SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of

Informatics, Systems and Communication

PhD program in Computer Science. Cycle XXXIV

Computational strategies to dissect the heterogeneity of multicellular systems via multiscale modelling and omics data analysis

Surname: Maspero Name: Davide

Registration number: 736996

Supervisor: Alex Graudenzi

Tutor: Prof. Raimondo Schettini

Coordinator: Prof. Leonardo Mariani

ACADEMIC YEAR 2021 - 2022

Contents

Abstract

Heterogeneity pervades biological systems and manifests itself in the structural and functional differences observed both among different individuals of the same group (e.g., organisms or disease systems) and among the constituent elements of a single individual (e.g., cells). The study of the heterogeneity of biological systems and, in particular, of multicellular systems is fundamental for the mechanistic understanding of complex physiological and pathological phenomena (e.g., cancer), as well as for the definition of effective prognostic, diagnostic, and therapeutic strategies.

This work focuses on developing and applying computational methods and mathematical models for characterising the heterogeneity of multicellular systems and, especially, cancer cell subpopulations underlying the evolution of neoplastic pathology. Similar methodologies have been developed to effectively characterise viral evolution and its heterogeneity. The research is divided into two complementary portions, the first aimed at defining methods for the analysis and integration of omics data generated by sequencing experiments, the second at modelling a multiscale simulation of multicellular systems. Regarding the first strand, next-generation sequencing technologies allow us to generate vast amounts of omics data, for example, related to the genome or transcriptome of a given individual, through bulk or single-cell sequencing experiments. One of the main challenges in computer science is to define computational methods to extract useful information from such data, taking into account the high levels of data-specific errors, mainly due to technological limitations. In particular, in the context of this work, we focused on developing methods for the analysis of gene expression and genomic mutation data. In detail, an exhaustive comparison of machine-learning methods for denoising and imputation of single-cell RNA-sequencing data has been performed. Moreover, methods for mapping expression profiles onto metabolic networks have been developed through an innovative framework that has allowed one to stratify cancer patients according to their metabolism. A subsequent extension of the method allowed us to analyse the distribution of metabolic fluxes within a population of cells via a flux balance analysis approach. Regarding the analysis of mutational profiles, the first method for reconstructing phylogenomic models from longitudinal data at single-cell resolution has been designed and implemented, exploiting a framework that combines a Markov Chain Monte Carlo with a novel weighted likelihood function. Similarly, a framework

that exploits low-frequency mutation profiles to reconstruct robust phylogenies and likely chains of infection has been developed by analysing sequencing data from viral samples. The same mutational profiles also allow us to deconvolve the signal in the signatures associated with specific molecular mechanisms that generate such mutations through an approach based on non-negative matrix factorisation.

The research conducted with regard to the computational simulation has led to the development of a multiscale model, in which the simulation of cell population dynamics, represented through a Cellular Potts Model, is coupled to the optimisation of a metabolic model associated with each synthetic cell. Using this model, it is possible to represent assumptions in mathematical terms and observe properties emerging from these assumptions.

Finally, we present a first attempt to combine the two methodological approaches which led to the integration of single-cell RNA-seq data within the multiscale model. This new modelling framework allows us to formulate data-driven hypotheses on the emerging properties of the system.

1

Introduction

Living beings can be considered as *complex systems*, because they exhibit the following key properties: (*i*) they are composed by a high number of interacting entities and sub-components, (*ii*) they show a hierarchical (multi-level) organisation, (*iii*) they present nonlinear interactions and feedbacks, (*iv*) they usually lack a central control and show patterns of self-organization, (*v*) their emerging dynamical behaviour is characterized by spontaneous order and metastable states, (*vi*) they are typically robust to perturbations [65].

For the same reason, all components of multicellular systems, i.e., *cells*, are complex systems themselves, and so are all the *ensembles* of biological entities that define their structure, functioning and dynamical organization, e.g., the genome, the transcriptome, the proteome, the metabolome, etc. Such ensembles are object of investigation of the so-called *omics* sciences (genomics, transcriptomics, epigenomic, proteomics, metabolomics, etc.) and, in broader terms, of *systems biology*. More recently, an attempt to combine concepts and methods from complex systems science, typically related to statistical physics, and those from systems biology, led to the definition of *complex systems biology* [31, 129]. Notice also that a significant portion of the so-called *network science* has been focused on the characterization of the static and dynamical features of the various omics layers, leading to the identification of many universal properties and regularities [60].

Within this lively scientific field, thanks to the continuous theoretical and methodological advancements of computer and data science, and to the ever-increasing computational power available in dry labs, it is now possible to: (*i*) design computational

methods to fully exploit the growing availability of omics data, and (*ii*) implement expressive modelling and simulation frameworks, so to deliver a fine characterization of complex biological systems and phenomena. In this work, I describe the attempts of designing computational methods to this end, with a specific focus on the dissection of the *heterogeneity* that naturally arises from the interaction of cells in multi-cellular systems and, especially, on cancer and viral evolution.

1.1 Biological background: the heterogeneity of (multi-cellular) biological systems.

Heterogeneity. Cells can be considered as the basic units of life [108]. In a nutshell, although all cells share key common features, they can be highly different in appearance and function, as outcome of billion years of evolution and adaptation to environmental changes. All cells are composed by the same classes of organic molecules, such as nucleic acids (i.e., DNA, RNA), proteins (e.g., enzymes, receptors, structural, or with other functions), and chemical species (e.g., carbohydrates, lipids), and they all rely on a number of strategies to survive, maintain their functioning and replicate. Together, cells constitute tissues that, in turn, form organs and eventually entire organisms, in a progressively more complex hierarchical organization.

The result of this multi-level and multi-scale interplay is the complex dynamical behaviour that defines the functioning of a living being. In other terms, the interaction of a large number of functionally distinct biological entities gives rise to an emerging behaviour that allows an organism to grow, survive, proliferate, maintain homeostasis and adapt to environmental changes.

One of the main phenomena related to this state of affairs, and which is observed in most biological systems, is the emergence of *heterogeneity* as, e.g., in tissues, in which ensembles of cells of highly-specialized types coexist and interact without any explicit central control. A notable example in this respect is that of cell differentiation. For instance, during embryogenesis, from a single egg cell, an entire organism is progressively generated [222], driven by fine-tuned gene regulation and morphogenetic processes. Notably, these two developmental processes are not independent, as changes in cell fate, determined by the transcriptional program, can modify the architecture of tissues, and vice versa [130]. Analogously, during the lifetime of an organism further cell differentiation processes are required to maintain tissue homeostasis, or to respond to disruptive events (e.g., wounds), via the generation of new specialized cells from stem cells [32, 157]. As a result, tissues and, in turn, organs are composed by a heterogeneous and ever-changing multitude of interacting cell types and modes.

Another important example is provided by diseases like cancer, a disorder arising over time and space from the complex genetic and environmental interactions of cells [66]. During the disease progression, tumors generally become more heterogeneous, resulting

in a diverse collection of cell subpopulations harbouring distinct molecular signatures and exhibiting different properties (e.g., proliferation rates, metastatic capacity, or sensitivity to therapies), typically known as cancer hallmarks [46]. Such heterogeneity is both *spatial*, i.e., it results in a non-uniform distribution of genetically distinct cancer subpopulations across and within disease sites, and *temporal*, i.e., the molecular composition of cancer cells changes over time [148]. Importantly, much of the attention of recent cancer research has been devoted to the investigation of intra-tumour heterogeneity (ITH), which is a major cause of drug resistance and relapse [70, 232].

Interestingly, even viruses can display complex population structures, where different viral subpopulations known as *quasispecies* coexist and are supposed to underlie most of the adaptive potential to the response of the immune system and to anti-viral therapeutic strategies [50, 214].

Overall, we note that heterogeneity is pervasive in any aspect of biology, and results both in the observed differences among organisms (e.g., inter-organism, inter-patient, inter-condition heterogeneity, etc.) and in those within the same system (e.g., intra-patient, intra-host, intra-tissue, intra-tumor heterogeneity, etc.). Accordingly, heterogeneity reflects into the different hierarchical layers of a biological system, i.e., the omics layers.

Omics layers. In brief, the *genome* can be considered the base-level of the organization of a biological system, as it is defined as the ensemble of all genes (i.e., roughly, the portions of DNA that encode for specific products) and regulatory elements of a given organism. A genome encodes the information needed to regulate, for example, the synthesis of every protein in a cell, at the right time and in a given environmental condition. Genome information is heritable and constitute the *genotype* of the organisms. For the sake of simplicity, genetic information is translated into the *proteome* through the *transcriptome*, i.e., the ensemble of RNA molecules obtained by transcribing specific genes. This hierarchy is effectively described by the so-called unified theory of gene expression [21]. One can have a general idea of the overall complexity by considering that a human cell includes more than 20000 genes, which produce more than 200000 transcripts [164], which undergo post-transcriptional and post-translational regulation processes to produce an even higher number of functional proteins, in highly complex gene regulatory networks [147]. Finally, the *metabolome* is the ensemble of chemical entities (i.e., metabolites) that are uptaken from the outside, transformed via biochemical reactions, and secreted. Metabolic biochemical reactions shape complex networks, in which most of the reactions are catalysed by enzymes, i.e., proteins synthesized through gene regulation [143].

Importantly, the Oxford dictionary defines the *phenotype* of an organism as *"the set of characteristics of a living thing, resulting from its combination of genes and the*

effect of its environment". Following this definition, all the omics layers downstream of the genome, i.e., transcriptome, proteome, metabolome, etc. represent the phenotype of a given organism. Metabolome, in particular, provides one of the best functional readouts of cellular phenotype of both healthy and pathological states [48]. However, phenotypic traits are not equally determined by genetics and environment. On the one hand, a higher contribution of the latter leads to the emergence of heterogeneous phenotypes under different conditions (i.e., phenotypic plasticity [43]). On the other hand, genomes can be robust to perturbations, as genetically distinct organisms can show similar phenotypic traits (i.e., genotype canalization [28]).

Therefore, a fine characterization of the genotype-phenotype map [14, 33] could be crucial in the investigation of complex biological systems and phenomena, such as cancer [172, 177, 226], even if it is hindered by the complex interaction of all omics layers, which is highly dependent on the environmental conditions and on the evolutionary history of the organism itself. This multi-level and multi-scale interplay shapes a dynamically evolving hierarchical organization, whose distinct levels are typically object of investigation of different scientific disciplines, including computer and data science, especially thanks to the increasing availability of omics datasets via high-throughput technologies.

In fact, different technologies and protocols now provide accurate and reliable measurements of the distinct omics layers including, e.g., genome sequences, and transcriptomic, proteomic and metabolomic states (i.e., abundance of transcripts, proteines, and metabolites). Furthermore, recently the resolution of these measurements has been extended to the single-cell level [105]. In this respect, excellent results are obtained for single-cell gene expression profiling, since it is currently possible to measure the whole transcriptome for many thousands single cells via single-cell RNA sequencing (scRNA-seq) experiments, hence allowing for the dissection of the intra-sample transcriptomic heterogeneity [162, 210].

Based on these many premises, two of the possible approaches to investigate the heterogeneity of multicellular systems are either by effectively exploiting the wealth of available omics data, or by simulating their behaviour via computational models.

1.2 Computational background: methods for the investigation of the heterogeneity of multicellular systems

In this work, I have focused on the definition of computational methods aimed at the investigation of the heterogeneity that arises from the interaction of cells in multicellular systems. As explained above, two major complementary methodological approaches exist, which can be summarized in: (*i*) methods for omics data analysis and integration (*bottom-up* strategies), and (*ii*) multiscale modelling and simulation (*top-down* strategies).

1.2.1 Methods for omics data analysis and integration

The first category of methods aims at investigating the heterogeneity of biological systems by extracting usable knowledge from the omics data generated via wet lab experiments (e.g., sequencing). As explained above, current technologies allow one to retrieve accurate measurements from all different biological layers. In this work, I have focused on methods aimed at exploiting either: (*i*) gene expression profiles or (*ii*) mutational profiles.

Gene expression. In the former category, a plethora of computational methods have been recently developed to analyse the gene expression profiles generated from either bulk or single-cell analyses (see Section 2.2).

Bulk data can be exploited to perform many different analyses, among which Differential Gene Expression, which is applied to detect significant changes in the gene expression in different conditions [190] or individuals [100]. Gene expression profiles can also be used to stratify patients to characterize their diseases [125], to retrieve specific gene expression signatures with prognostic power [224], to predict the response to therapies [126] or determine survival biomarkers [199], usually via machine learning approaches. Moreover, they allow one to estimate the relative proportion of the different cell types or cancer cells in a given sample by deconvolving the gene expression signal into cell type-specific signatures [116].

Single-cell data can be used with similar purposes, but instead of considering the difference among biological specimens, they are typically used to investigate the heterogeneity among the cell subpopulations present in a given sample, e.g., a biopsy or a patient-derived organoid or cell culture. For example, single-cell data can provide information about specific expression signatures to distinguish different cell types [171, 246], or can allow one to extract a temporal ordering of their differentiation state [196]. Other methods were recently introduced to predict the future state of individual cells by exploiting the ratio of abundance between spliced and unspliced mRNA in each cell [158, 207]. Clearly, this list is far from being conclusive, and a surge of analytical instruments to investigate gene expression from single-cell data is observed nowadays [217].

Given the high dimensionality and the data- and technology-dependent error rates, single-cell data are more challenging to handle than bulk data, and require ad-hoc methods for their analysis. In this regard, best practices in quality control and preprocessing were recently suggested [186]. These steps and other downstream analyses can be performed using toolkits available in different suites, such as Seurat [236], and SCANPY [168]. However, many challenges are still open and new computational strategies are needed to improve the reliability of existing analyses and pave the way for new ones [217].

Genomic mutations. The second category includes the computational frameworks designed to exploit mutation profiles for: (*i*) the reconstruction of the evolutionary history of a biological system or (*ii*) the characterization of the underlying mutational processes.

In general, the former analyses can be executed considering either an ensemble of independent biological systems, e.g., tumor samples from distinct oncological patients, – here we speak of *population-level* analyses –, or analysing a single biological system, e.g., a single tumour, – *individual-level* analyses.

Among the computational methods designed to perform population-level analyses, some exploit binarized mutation profiles, in which mutations are either present (1) or absent (0) in a given sample [34, 57, 103, 179, 228]. The aim is to retrieve the most likely trends of accumulation of genomic mutations (i.e., *drivers* [72]) in a given tumor type. Cross-sectional data can be exploited to this end, and the result is a probabilistic graphical model which explains the statistical trends existing in the data, which might be used to deliver prognostic insights.

The inference of individual-level evolution models starts, instead, from the dissection of ITH, which is often achieved reconstructing the underlying clonal structure of a tumor (i.e., the genetically distinct cancer cell subpopulations). This can be performed either from bulk or single-cell data. For example, bulk sequencing experiments performed on multi-region biopsies, return the Variant Allele Frequency (VAF) profiles, i.e., the frequency of each mutation detected in a given sample. VAFs are used to estimate the Cancer Cell Fraction (CCF) i.e., the relative abundance of cancer cells carrying the mutation, which in turn is exploited to determine the genotype and the abundance of each clone. Unfortunately, strong tumor-sampling bias, as well as errors in sequencing data (e.g., purity, ploidy, Copy Number Alterations, or mutation multiplicity) induce uncertainty in the inferred model of cancer evolution. Thus, several methods are available to reconstruct the clonal structure from this data type [86, 209]. Other techniques return also the evolutionary model by exploiting multiple-sample data [96], even in the longitudinal case [189].We finally highlight that a recent method was proposed to leverage the regularities detected across single tumors of the the same type, so to identify possible patterns of repeated evolution, via transfer learning [144].

Distinct computational strategies exist to exploit single-cell data to reconstruct individual-level models of cancer evolution. In principle, with these experiments, mutations are univocally assigned to each cell. However, technological limitations inflate data with false positives, false negatives and missing values. So, statistical inference approaches need to be specifically designed to retrieve consistent results from such noisy data type. For instance, the method SCITE [113] maximizes a likelihood function to find the evolution model that best fits the input single-cell mutational profiles, by assuming the so-called Infinite Sites Assumption (ISA) [9], similarly to

other methods such as TRAIT [192] , while other approaches relax such assumption, e.g., siCloneFit [200] or SASC [231] . Instead, ddClone [135] couples the information of both bulk and single-cell data, obtained from the same tumour samples, to improve the deconvolution steps and the inference.

In addition, mutational profiles can also provide information about the molecular mechanisms underlying their origination. In fact, many harmful mechanisms and the subsequent specific DNA repair processes generate specific point mutation rate spectra, called *signatures*. For example, it is proved that tobacco smoking damages the human DNA inducing mainly the C to A transversion [107]. Identifying which mutagenic processes are active in a tumour may indicate biological differences, or provide markers of therapeutic response. To this aim, signatures can be detected by decomposing the mutation spectra across multiple tumour samples, usually via Non-Negative Matrix Factorization (NMF), [59, 204, 240].

In this work I have been developing methods for the analyses of both gene expression (Section 3.1) and mutational profiles (Section 3.2). Interestingly, in one of my works I could integrate the information on both levels, characterizing the relation between genotype and phenotype at the single-cell level for the very first time (see paper P#5).

1.2.2 Multiscale modelling and simulation of multicellular systems

The second category involves methods aimed at *modelling* and *simulating* the emerging behaviour of the system under investigation. In fact, computational/mathematical models have repeatedly proven fruitful for understanding mechanisms and processes of complex biological systems and diseases, such as cancer [90]. Overall, simulations allow one to carry out a virtually unlimited number of experiments in a broad range of in-silico scenarios, with reduced costs, higher speed and increased feasibility than any real-world (e.g., wet-lab) experiment.

Two general and complementary approaches are possible in this regard: (i) highly-detailed models and (ii) simplified/abstract models.

Notice that the chosen classification is functional to the aim of the thesis. However, mathematical models could be more formally classified with a different approach. For instance, they can be grouped based on their mathematical formalization. Either detailed and abstract model categories can include models based on deterministic, hybrid or stochastic formulations.

- *Detailed models*
  Detailed models are rooted in reductionist theory [26] and are aimed at precisely characterize the functions and interactions of each component of the biological system under investigation. In this case, it is possible to define and tune the

parameters of a corresponding in-silico model to finely replicate its static and dynamical properties.

To mention a few examples in this sphere, some methods employ deterministic formulations (either via Ordinary Differential Equations or Partial Differential Equations) to deliver: a quantitative description of time-dependent biological processes like cell signaling pathway [61, 71], the dynamics of immune system activation [194], the metabolism [69], and the pharmacodynamics and pharmacokinetics of a drug [149].

In addition, it is possible to add a certain degree of uncertainty by employing stochastic formulations to the previous highly detailed models. This approach may be particularly effective when the system includes a few number of elements, e.g., chemical reactions far from thermodynamic equilibrium, which involve low copy numbers of chemical species. Another example is the binding and dissociation events of a signalling molecule and its receptor resulting from random encounters between them. Stochastic Differential Equations and Master Equations are one the most effective approaches to represent such processes [30, 136].

A clear limitation is the necessity of the parametrization of the model, e.g., setting all the kinetic constants of chemical reactions to properly characterize the biological systems. Constraint-based approaches are conceived to overcome this issue, by employing simplifying assumptions which hold in specific scenarios. For example, cellular metabolism can be simulated with such approach and flux balance analysis optimization, avoiding the parametrization of kinetic constants [76].

We also note that, interestingly, detailed models were repeatedly proven fruitful to propose diagnostic or therapeutic strategies [22, 80, 180].

- *Abstract models*
  This second category starts from the consideration that mathematical models can be effective also when the details of the system are mostly unknown. In this case, one starts by measuring high-level properties of the biological system to retrieve similar patterns in the behaviour of a highly simplified or abstract simulated model. If the same patterns emerge, the model can enhance our understanding of the general properties of the system. The approach can be helpful to reveal unknown biological principles or regularities, which can be missed by a qualitative approach, testing theories on quantitative grounds, evaluating assumptions, and providing predictions on emerging behaviours. For these reasons abstract models are widely used in *complex systems science* and *statistical physics*.

  Notable examples in this respect were provided by the many works of Stuart Kauffman and epigones, which modelled gene regulation via highly-simplified Boolean

network models, yet leading to the identification of important universal properties of real-world systems [7, 8].

A similar methodological approach was used to investigate cancer properties. Sui Huang started from the works on Boolean networks and embraced the concept of cancer attractor. In brief, healthy cells would wander in distinct attractors of the gene regulatory network state space representing normal cell types, whereas cancer cells would fall in aberrant attractors after genetic mutations or external perturbations [35, 64, 115].

In general, this approach has proven effective in cancer research. In fact, many aspects of cancer progression, such as accumulation of mutations during cell division, can be effectively modelled with highly simplified mathematical frameworks. For instance, branching [40, 78] and Moran processes [2] can be used to simulate the cancer initiation and its tissue hierarchy resulting from the accumulation of new somatic mutations, and set the basis, e.g., for the simulation of complex emerging structures and patterns of tumors [47, 104].

Multiscale modelling and simulation. Distinct biological layers and processes involving different spatial and temporal scales (e.g., the gene regulation and the cell population dynamics) can be coupled in a unique simulation framework, that of *multiscale modelling and simulation* [73]. A wide and increasing number of multiscale models currently exist, typically referred to specific (sub)systems, such as tissues or organs like the neuromuscolar system [193], complex processes like bone healing [117], or to specific diseases, such as neuropsychiatric disease [81], or infections [218]. Moreover, attempts to modelling an entire organism both unicellular like the *Mycoplasma genitalium* [52] or a plant [243] have been proposed. Note that, as shown, e.g., in [93], a sufficiently sophisticated model could predict complex, multi-network phenotypes such as the efficacy evaluation of a given drug, leading to a completely personalized therapy strategy for a disease.

In this work, we are mainly interested in investigating the behaviour of multicellular systems and, especially, of cancer emergence and evolution, which results from the complex interaction between cell subpopulations and the microenvironment [19, 178].

To this end, many multiscale models have been proposed, in most of which cancer cells are described as discrete agents that populate either in-lattice or off-lattice environments. In the former case, the space is split into discrete subportions, whereas in the latter, more complex geometrical representations are employed (see the reviews on the issue [62, 188]).

We finally highlight that, in the choice of any suitable modelling framework, there exists a harsh trade-off between the expressivity and the parsimoniousness of any mathematical/computational model, which translates into several theoretical issues in terms of model selection and parameter estimation [181]. When the number of parameters

of the models needed to represent real-world systems and phenomena is high, powerful statistical methods are required for proper parameter fitting. A notable and powerful example is given by the Approximate Bayesian Computation (ABC) approach, which tries to interpret experimental or observational data in light of mathematical models. ABC approaches allow one to tune the many parameters of a model, by computing summary statistics generated via the repeated simulation of the model, instead of computing the exact likelihood calculations [42, 88].

In this thesis, I defined a new multiscale modelling and simulation framework, to investigate the metabolic heterogeneity of cell subpopulations and their interactions. The approach combines an abstract model for the spatial population dynamics, simulated via Cellular Potts Model [124], with a more detailed one to simulate the cellular metabolism, via Flux Balance Analysis [76].

1.3 Computational challenges

Computer and data science provide an effective methodological background to solve open issues in biomedical sciences. Each biological question generates a number of computational problems that, in this work, were addressed according to the two research branches discussed in the Computational background section.

Multiscale modelling and simulation. Many computational challenges are related to the proper modelling and simulation of multicellular systems.

Challenge 1) Trade-off between model expressivity and over-parameterisation
In order to investigate the emergent properties and patterns of real-world systems, computational frameworks need to be sufficiently *expressive*, avoiding the concurrent explosion of the parameter space. The search for the optimal trade-off is one of the key challenges faced during the definition and implementation of the computational frameworks proposed in this work.

Challenge 2) Scalability
Multiscale models are typically computationally expensive, limiting the possibility of exploring in-silico experimental scenarios.

Challenge 3) New metrics to analyze of emergent/generic properties
An important issue is the definition of effective metrics to compare the emerging behaviour and properties of the simulated system with real-world observations.

Omics data analysis. Many computational challenges related to the current work are explicitly related to the methods developed and employed to analyse omics data.

- Gene expression

Challenge 4) Denoising and imputation
It is well known that in gene expression profile derived from single-cell RNA-seq experiments, a high rate of errors ad missing data is present, due to technical limitation (e.g., allele dropout) and biological variability (e.g., batch effect). One of the key challenges lies in the definition of computational strategies to denoise ad impute single-cell data, reducing sparsity in the matrix and improve the downstream analyses.

Challenge 5) Usability
Given the ever-increasing amount of computational methods for omics data analysis, another open challenge is the release of tools that are actually used in real-world research and also by non-experts in bioinformatics or computer science, by improving the overall usability.

Challenge 6) Feature selection in sample classification
Machine learning strategies are widely used to classify samples from gene expression data. Very often results can be improved in terms of performance and interpretability by reducing the features taken into account. Thus, an important computational challenge is the selection of the most relevant features for classification of (cancer) samples. To this end a promising approach is to focus on the topological features of metabolic nets.

- Genomic mutations

Challenge 7) Improving the robustness of phylogenetic models
Mutational profiles generated via variant calling from either single-cell or viral sequencing data can present high levels of false positives, false negatives and missing values. For this reason, a challenge is how to properly deal with such noisy data. For this purpose, the use of likelihood-based and Bayesian statistical frameworks, as well as of search and optimization methods (e.g., via Markov chain Monte Carlo) should be explored.

Challenge 8) Assessing model performance
Since no benchmark datasets for single-cell phylogenomics are currently available, the simulation of realistic datasets is essential, as proposed for instance in [140, 201]. Accordingly, ad-hoc synthetic datasets need to be created for any performance assessment.

Challenge 9) Lack of methods for longitudinal single-cell datasets
Longitudinal (i.e., time course) single-cell sequencing experiments are becoming popular, but no currently available phylogenetic method is specifically designed to handle this kind of data to reconstruct, e.g., the evolutionary history of tumours.

Challenge 10) Characterization of search space in MCMC

In phylogenomic inference, a challenging aspect is the vast search space, which exponentially increases with the number of variables (i.e., mutations) included in the model. Often, the optimization algorithms may get stuck in a local minimum. To overcome this problem, consensus methods [23] can be effective to collect information from all the solutions (i.e., trees) sampled during the Markov chain Monte Carlo search.

Challenge 11) Exploiting mutation types of viral samples
Deep sequencing experiments of viral samples provide high-resolution data allowing to distinguish high-frequency mutations (i.e., clonal or fixed), included in consensus sequences and most likely transmitted during infections, and low-frequency mutations (i.e., minor), which emerge due to host-related processes and might be used to improve the reconstruction of infection chains. New methods to exploit the distinct mutation types are needed.

Challenge 12) Identification of mutational signatures of viral samples.
Mutational signature detected in cancer samples highlight specific damaging mechanisms. Unfortunately, no method to decompose the mutation spectra of viral samples is currently available. Such analysis could be useful to retrieve insights on viral-host interactions.

- Omics data integration.
  Another current major challenge in bioinformatics and computional biology is the definition of effective strategies to integrate multiple omics information.

  Challenge 13) Multi-omics data mining
  The transcriptome is the biological layer between the genome (i.e., genotype) and the proteome (i.e., phenotype). Therefore, RNA-seq data might be exploited to retrieve: (*i*) the genomic mutation profiles, (*ii*) the transcriptome states, and (*iii*) an approximate estimate of protein abundances, of the same biological sample. Computational methods and pipeline to this end are needed and might allow one to perform multi-omics data mining.

  Challenge 14) Data integration into multiscale models
  Another challenging task is the integration of real-world data into multiscale model, in order to provide a finer parameter estimation. In a previous work [175], we proved that is possible to integrate single-cell data into metabolic models. Here, the challenge is to extend the approach to exploit RNA-seq data in the settings of the computational model.

- Data analysis (overall)
  Challenge 15) Reproducibility
  A key challenge in computer-science is the production of analyses that are repro-

ducible and robust. To this aim, computational pipelines should be coded using workflow managers, like Nextflow [122] or Snakemake [245], while dependency and required software should be provided using containers (e.g., Docker [83]). Such tools also allow one to improve the scalability and efficiency of the computational analyses.

1.4 Main achievements

The main contributions of this thesis can be divided into four main general topics, which are related to the challenges listed above: (*i*) omics data preprocessing pipelines, (*ii*) methods for omics data analysis and integration, (*iii*) multiscale modelling and simulation of multicellular systems, and (*iv*) data-driven multiscale modelling.

Overall, 8 computational methods were developed and released in public repositories such as GitHub, Bioconductor and Galaxy Project. In Appendix: Code repositories, one can find the references to the code repositories.

- **Omics data preprocessing pipelines** (Chapter 2)
  Mutational profiles of biological samples are typically obtained from (single-cell or bulk) DNA-seq experiments. However, since mRNAs are obtained by transcribing the genome, and the RNA editing in human cells is limited [53], I have implemented a novel pipeline to retrieve somatic mutation profiles from single-cell RNA-seq data. In paper P#2 we proved the reliability of the results delivered by the pipeline, also by detecting a likely experimental error in [165].

  Another achievement regarding omics data preprossing led to the most extensive review of denoising and imputation methods for single-cell RNA-seq data. In fact, it is reasonable to expect that distinct denoising methods could perform differently depending on the dataset being analyzed. Therefore, a quantitative benchmarking of such methods was needed to generate guidelines for researchers. In paper P#1 we compared 19 denoising methods, both on simulations and real data, to assess which approach provides better results under different conditions.

  A further work, included in the Appendix 1, concerns the definition of a self-contained pipeline for the discovery and assignment of mutational signatures from viral samples, which is related to another publication described in paper P#7. The pipeline is coded in Nextflow to ensure scalability and portability. Moreover, all the required bioinformatics tools and dependencies are pre-installed in a Docker image to improve the reproducibility of the results.

- **Methods for omics data analysis and integration** (Chapter 3)

Gene expression profiles. In [154] we proved that is possible to extract translational knowledge from RNA-seq datasets projecting gene expression profiles onto metabolic networks, and that such information is relevant for cancer progression. In paper P#4, we explored the feasibility of classifying cancer samples using the topological features of the related metabolic networks, as estimated from the corresponding expression profiles. As a result, just a few key topological features are sufficient to provide good classification results via state-of-the-art machine learning strategies, hence providing a new diagnostic automated tool for cancer research.

In addition, we translated the original method presented in [154] in a dedicated tool released on the Galaxy Project server (named MaREA4Galaxy), so to improve its overall usability. Galaxy Project is a web-based platform that was released to provide a user-friendly GUI interface and a repository with bioinformatic tools ready to be installed [142]. The tool is currently online on a public server that is part of the *ELIXIR* infrastructure (https://elixir-europe.org/).

Mutational profiles – cancer samples. Thanks to the continuous improvements of single-cell sequencing technologies, longitudinal (i.e., time-course) datasets are increasingly being generated, e.g., from patient-derived organoids, cell cultures or xenografts [163, 254, 258]. Thus, we developed the first computational method explicitly designed to handle single-cell mutational profiles generated from samples collected at distinct time points, to reconstruct the evolutionary history of a tumor, namely (LACE – Longitudinal Analysis of Cancer Evolution). Thanks to a robust statistical framework, which combines Boolean Matrix Factorization, a newly defined weighted likelihood function and Markov chain Monte Carlo search, LACE improves over state-of-the-art methods especially in condition of high noise and sampling limitations. Accordingly, LACE is effective with mutation profiles from single-cell RNA-seq data and this allows for the integration between genomic (mutations) and transcriptomic (gene expression) layers at the single-cell level (paper P#5).

Mutational profiles – viral samples. Most phylogenomic and phylodynamic studies on the evolution and diffusion of viruses, such as SARS-CoV-2, process consensus sequences generated from sequencing experiments of viral samples. Yet, most studies discard the information on low-frequency intra-host mutations, which can be retrieved from raw sequencing data, and which are supposed to play a key role in viral evolution and adaptation.

In paper P#6 we developed a new statistical framework (VERSO – Viral Evolution ReconStructiOn) with two distinct goals. The method first exploits the profiles

of the mutations included in consensus sequences to improve the reconstruction of the phylogeny, via a robust statistical approach borrowed from cancer evolution research, and which improves over the state-of-the-art especially in conditions of sampling inhomogeneity. VERSO then employs the profiles of low-frequency mutations to characterize intra-host heterogeneity and retrieve a fine-resolution map of infection events, and which would be impossible with consensus sequences only.

Furthermore, in an attempt of solving challenge 10, we developed a novel (unpublished) method to compute a consensus optimum branching tree, instead of the maximum likelihood one, by applying the Chu–Liu/Edmonds' optimal branching algorithm [11] to summarize the evolution models sampled during the MCMC search. We evaluate the improvement of our method for both local and global topological properties, over a large number of synthetic datasets representing different experimental scenarios (see section 3.2.1.1).

In addition, another method was developed to infer the mutational signatures of SARS-CoV-2, via Non-negative Matrix Factorization of mutational spectra. In paper P#7 we proved that this processes are host-specific and can be used to stratify samples in homogeneous clusters.

- **Multiscale modelling and simulation of multicellular systems** (Chapter 4) In section 4.2, we introduce a new multiscale multicellular framework, named Flux Balance Cellular Automata (FBCA), to investigate the metabolic heterogeneity of cancer subpopulations under different conditions. In particular, we propose to combine the spatial dynamics (i.e., cell growth, division, death, and migration) via Cellular Potts Model, with a realistic simulation of the cellular metabolism via Flux Balance Analysis.

  In particular, in paper P#8, we compared different experimental scenarios by computing population dynamics measures like, e.g., cell duplication time, average cell size, total accumulated biomass, clonal size (i.e., number of cells with the same ancestor), proving the effectiveness of the modelling approach.

  In paper P#9, we improved the representation of the microenvironment, by modelling nutrients diffusion and uptake/secretion. This further process allows for a better evaluation of the impact of the microenvironment, and of the emerging metabolic behaviour of cells.

  Note that key efforts were devoted to improve the overall scalability of the multiscale model. In fact, FBCA is coded in Matlab, an efficient programming language to handle matrix calculation. Specifically, we relied on a multicore approach to optimize the computation of cell metabolisms in parallel, and exploited Matlab utilities to refactor the code and optimize the overall computational time.

- **Data-driven multiscale modelling** (Chapter 5)
  FBCA is a flexible framework that is suitable for the integration with single-cell transcriptome data. Therefore, to fill the gap between theory (multiscale models) and data (omics data analysis and integration), in paper P#10 we propose a preliminary attempt to this end, and which might pave the way for further development in the direction of predictive modelling of complex biological phenomena.

Finally notice that in Appendix P#A2, a paper regarding the definition of a framework for the design of optimized therapies via control theory is also presented. While being only partially related to the other works presented in the thesis, this article provides a solid ground for future developments combining multiscale modelling and control theory.

1.4.1 Articles

In the following, the list of included papers is reported. My contribution to each of them are indicated considering the CRediT (Contribution Roles Taxonomy) author statement [170].

PAPER #1 *Omics data preprocessing pipelines*
Challenge 4

*Patruno, L., *Maspero, D., Craighero, F., Angaroni, F., Antoniotti, M., and Graudenzi, A. *A review of computational strategies for denoising and imputation of single-cell transcriptomic.* **Briefings in Bioinformatics** 22.4 (Oct. 2020). doi:10.1093/bib/bbaa222. *equal contribution.

CRediT: Conceptualization, Formal analysis, Investigation, Methodology, Software, Resources, Validation, Visualization, and Writing - Original Draft

PAPER #2 *Omics data preprocessing pipelines*
Challenge 13

Ramazzotti, D., Angaroni, F., Maspero, D., Ascolani, G., Castiglioni, I., Piazza, R., Antoniotti, M., and Graudenzi, A. *Variant calling from scRNA-seq data allows the assessment of cellular identity in patient-derived cell lines.* ***Under review.*** Preprint at: doi:10.1101/2021.04.13.439634.

CRediT: Formal analysis, Software, Resources, Validation, Visualization, and Writing - Original Draft

PAPER #A1 *Omics data preprocessing pipelines*
Challenge 15

Maspero, D., Angaroni, F., Porro, D., Piazza, R., Graudenzi, A., and Ramazzotti, D. *VirMutSig: Discovery and assignment of viral mutational signatures from sequencing data.* **STAR Protocols** 2.4 (2021), p. 100911. doi:10.1016/j.xpro.2021.100911.

CRediT: Software, Validation, Visualization, Writing - Original Draft, Review & Editing

PAPER #3 *Omics data analysis and integration*
Challenges 2,5,13

Damiani, C., Rovida, L., Maspero, D., Sala, I., Rosato, L., Di Filippo, M., Pescini, D., Graudenzi, A., Antoniotti, M., and Mauri, G. *MaREA4Galaxy: Metabolic reaction enrichment analysis and visualization of RNA-seq data within Galaxy.* **Computational and Structural Biotechnology Journal** 18 (2020), pp. 993–999. doi:10.1016/j.csbj.2020.04.008.

CRediT: Conceptualization, Resources

PAPER #4 *Omics data analysis and integration*
Challenges 6,13

Machicao, J., Craighero, F., Maspero, D., Angaroni, F., Damiani, C., Graudenzi, A., Antoniotti, M., and Bruno, O. M. *On the Use of Topological Features of Metabolic Networks for the Classification of Cancer Samples.* **Current Genomics** 22.2 (2021), pp. 88–97. doi:10.2174/1389202922666210301084151

CRediT: Conceptualization, Data Curation, Visualization

PAPER #5 *Omics data analysis and integration*
Challenges 7,8,9,13

*Ramazzotti, D., *Angaroni, F., *Maspero, D., Ascolani, G., Castiglioni, I., Piazza, R., Antoniotti, M., and Graudenzi, A. . *LACE: Inference of cancer evolution models from longitudinal single-cell sequencing data.* **Journal of Computational Science**, 58 (2022), 101523. doi:10.1016/j.jocs.2021.101523. *equal contribution.

CRediT: Conceptualization, Formal analysis, Investigation, Methodology, Software, Resources, Visualization, and Writing - Original Draft

PAPER #6 *Omics data analysis and integration*
Challenge 11

Ramazzotti, D., Angaroni, F., Maspero, D., Gambacorti-Passerini, C., Antoniotti, M., Graudenzi, A., and Piazza, R. *VERSO: A comprehensive framework for the inference of robust phylogenies and the quantification of intra-host genomic diversity of viral samples.* **Patterns** 2.3 (2021), p. 100212. doi:10.1016/j.patter.2021.100212

CRediT: Conceptualization, Formal analysis, Investigation, Methodology, Validation, Visualization

PAPER #7 *Omics data analysis and integration*
Challenge 12

*Graudenzi, A., *Maspero, D., *Angaroni, F., Piazza, R., and Ramazzotti, D. *Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity.* **iScience** 24.2 (2021). doi:10.1016/j.isci.2021.102116. *equal contribution.

CRediT: Conceptualization, Formal analysis, Investigation, Methodology, Validation, Visualization, and Writing - Original Draft

PAPER #8 *Multiscale modelling and simulation*
Challenges 1,2,3

*Graudenzi, A., *Maspero, D., and *Damiani, C. *FBCA, A Multiscale Modeling Framework Combining Cellular Automata and Flux Balance Analysis.* **Journal of Cellular Automata** 15 (2020), pp. 75–95. url: jca-volume-15-number-1-2-2020/jca-15-1-2-p-75-95. *equal contribution.

CRediT: Conceptualization, Formal analysis, Investigation, Methodology, Software, and Visualization

PAPER #9 *Multiscale modelling and simulation*
Challenges. 1,2,3

Maspero, D., Damiani, C., Antoniotti, M., Graudenzi, A., Di Filippo, M., Vanoni, M., Caravagna, G., Colombo, R., Ramazzotti, D., and Pescini, D. *The Influence of Nutrients Diffusion on a Metabolism-driven Model of a Multi-cellular System.* **Fundamenta Informaticae** 171 (2020). 1-4, pp. 279–295. doi:10.3233/FI-2020-1883.

CRediT: Conceptualization, Formal analysis, Investigation, Methodology, Software, Visualization, and Writing - Original Draft

PAPER #10 *Data-driven multiscale modelling*
Challenges 1,2,3,13,14

Maspero, D., Di Filippo, M., Angaroni, F., Pescini, D., Mauri, G., Vanoni, M., Graudenzi, A., and Damiani, C. *Integration of single-cell RNA-sequencing data into Flux Balance Cellular Automata.* **Lecture Notes in Computer Science**. Ed. by Springer. Proceedings of Computational Intelligence methods for Bioinformatics and Biostatistics. In press. Computational Intelligence Methods for Bioinformatics and Biostatistics, CIBB 2019. Bergamo, Italy, Sept. 2019.

CRediT: Conceptualization, Formal analysis, Investigation, Methodology, Software, Resources, Visualization, and Writing - Original Draft

PAPER #A2
Challenge 1

Angaroni, F., Graudenzi, A., Rossignolo, M., Maspero, D., Calarco, T., Piazza, R., Montangero, S., and Antoniotti, M. *An Optimal Control Framework for the Automated Design of Personalized Cancer Treatments.* **Frontiers in Bioengineering and Biotechnology** 8 (2020), p. 523. doi:10.3389/fbioe.2020.00523.

CRediT: Investigation, Software

1.5 Structure of the thesis

The thesis is divided into 6 chapters and 2 appendices.

Chapters $2-5$ cover the specific contribution areas presented above. In particular, chapter 2 introduces technologies to produce omics data and explains our preprocessing pipelines. Chapter 3 covers methods for omics data analysis and integration which exploit either transcriptomic and genomic data. Chapter 4 focus on the use of multiscale modelling and simulation of multicellular system to investigate the heterogeneity and other systematic emerging properties. Finally, chapter 5 presents a data-driven multiscale modelling combining chapter 3 and 4 approaches (see Figure 1.1).

The articles related to the distinct topics are included in their entirety (for the Supplementary Material, please refer to the online version). For the sake of readability, only green-marked papers are present in the main body, while blue-marked ones are reported in the appendix A. We also highlight that papers are reported without any modification, including the specific reference numbering, which is different from that of the thesis. All chapters include a background section, as well as an introduction that explains the motivations and assumptions of the distinct methods.

The discussion section presents our final comments on the achievements presented in this thesis. We also propose possible future developments and research directions.

The following section 1.6 includes the list of abbreviations. Finally, appendix B includes the code repositories.

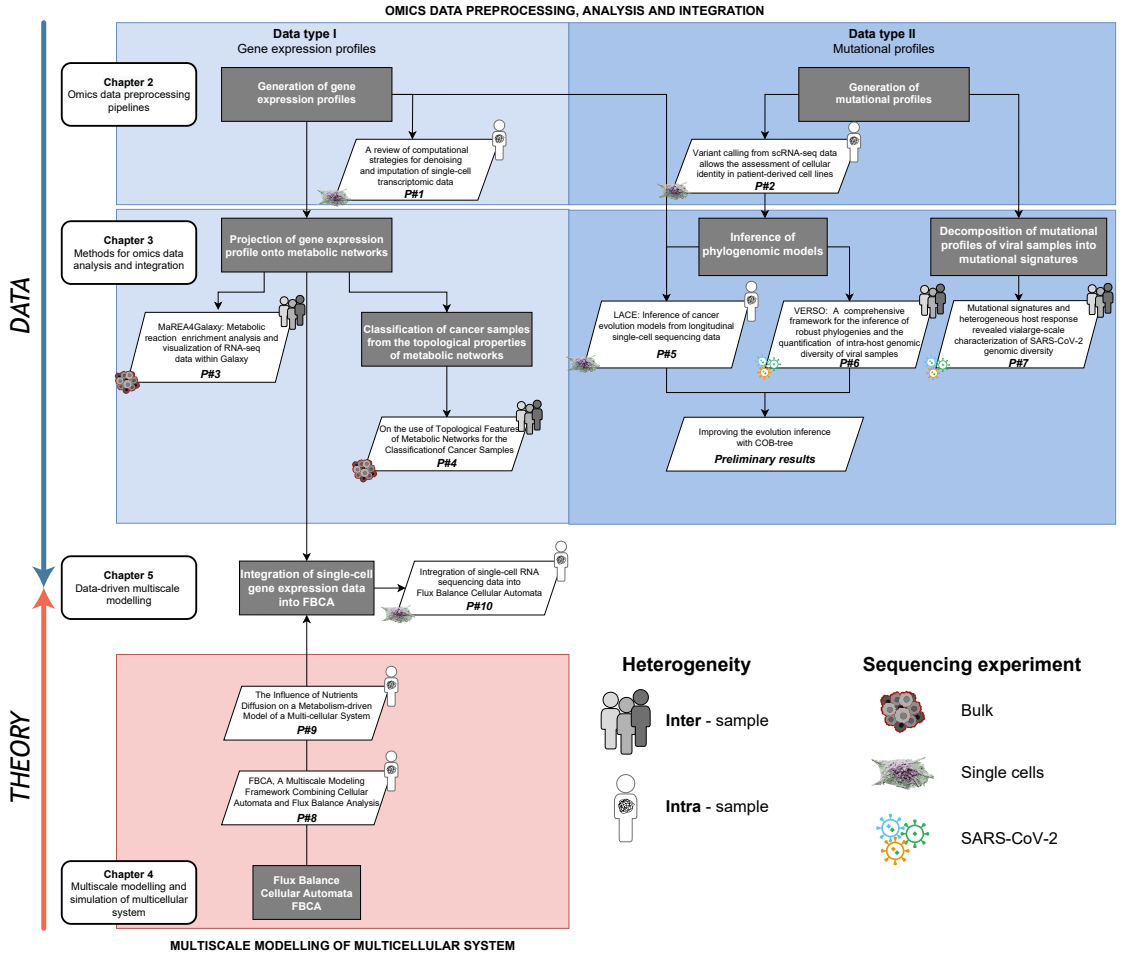


Figure 1.1: Schematic structure of the thesis.

1.6 Abbreviation list

Abbreviation	Description
CCF	Cancer Cell Fraction
CML	Chronic Myeloid Leukemia
COB	Consensus Optimum Branching
CPM	Cellular Potts Model
DEG	Differential Expressed Gene
FACS	Fluorescence-Activated Cell Sorted
FBA	Flux Balance Analysis
FBCA	Flux Balance Cellular Automata
FN	False-Negative
FP	False-Positive
ISA	Infinite Site Assumption
ITH	Intra-Tumour Heterogeneity
LACE	Longitudinal Analisys of Cancer Evolution
ML	Machine Learning
ML tree	Maximum Likelihood tree
NGS	Next-Generation Sequencing
NMF	Non-negative Matrix Factorization
RAS	Reaction Activity Score
RNA-seq	RNA sequencing
scFBA	single-cell Flux Balance Analysis
scRNA-seq	single-cell RNA sequecing
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SV	Structural Variant
TCGA	The Cancer Genome Atlas
VAF	Variant Allele Frequency
VERSO	Viral Evolution ReconStructiOn
VF	Variant Frequency

2

Omics data preprocessing pipelines

In this chapter, the new omics data preprocessing pipelines developed during the PhD project are presented, starting from a general overview of data generation via bulk, single-cell and viral sequencing experiments (section 2.1).

The results related to preprocessing of gene expression data are discussed in Section 2.2, whereas those regarding mutational profile generation in Section 2.3. Notice that an additional pipeline for the inference of viral mutational signatures is included in the paper included in Appendix, (paper P#A1).

2.1 Background: omics data generation via Next Generation Sequencing

From biological samples to BAM files

High-throughput sequencing technologies have underlied most of the recent advancements in biomedical sciences, with dramatic improvements over experimental workflows and best practices [127, 182]. For example, while the first human genome required years to be sequenced, at stellar costs, nowadays the sequencing of a human genome takes only a few hours, and at extremely limited costs. Massively parallel sequencing is at the root of the improvement observed with respect to the classical Sanger sequencing approach [89]. In general, the new approach and the concurrent advancements in workflow steps

led to colossal cost reductions in data generation [74].

In addition to genome sequencing, sequencing technology can be applied to collect data from many further biological layers, such as epigenomic (ChIP-seq or ATAC-seq) [38, 92]) and transcriptomic (RNA-seq [39]). Data produced via such technologies may hold the promise of personalized medicine, leading to routinely available sequencing tests that can guide patient treatment decisions. Readers interested in the history of the evolution of sequencing technologies and of their applications may refer to [166, 184], and to [212] for a specific focus on RNA sequencing technologies evolution.

Specifically, RNA-seq experiments are de-facto the new standard tool for the analysis of gene expression, and replaced Microarray assays by improving the range of detection, reducing the technical noise and allowing the detection of novel transcripts [198]. Similarly to any other omics analysis, transcriptomic measurement via RNA-seq can be executed at the resolution of either (*i*) bulk or (*ii*) single-cell experiments [239]. In the following, some details of the two experimental approaches are reported.

2.1.1 Bulk sequencing experiments

Bulk RNA-seq experiments allow one to measure the average expression level for every gene across the many cells included in a biological sample, e.g., a tissue biopsy. The data generated in this way are widely used, for instance, to perform differential gene expression analyses, by comparing different conditions or ensembles of patients, allowing one to quantify expression signatures, or to stratify them into distinct clusters [219]. The principal limitation is the lack of direct insights into subpopulation heterogeneity, which is relevant, in particular, when complex diseases (e.g., cancer) are the object of investigation. Nonetheless, the bulk approach offers relevant advantages over single-cell analyses, especially in terms of quality of the coverage (i.e., the number of unique reads aligned on each position of the reference genome). Higher coverage indicates a higher quality of data [87]).

Sequencing experiment protocol. The protocol to perform bulk RNA-seq experiments is here briefly sketched.

1. **Library preparation.** The procedure starts from the extraction and purification the ensemble of messenger RNAs (mRNAs) (i.e., single-stranded RNA molecules that are complementary to one of the DNA strands) extracted from the specimen. mRNAs are then broken into fragments and converted into cDNA, a more stable molecule.

2. **Sequencing.** The sequencing step is done via high-throughput platforms (e.g., Illumina HiSeq or NovaSeq) [85, 244]. The process generates millions of short reads from one end of each cDNA fragment (i.e., single-end sequencing) or of both ends

(i.e., paired-end sequencing). The result is one or more FASTQ files reporting the reads (i.e., strings of nucleotides detected) and the corresponding quality scores. For more technical details on the file format, please refer to [41].

3. **Alignment.** Raw reads are processed to remove tags and low-quality nucleotides (i.e., trimming). Then, they are mapped on a proper reference genome [75] to retrieve the genomic coordinate to each mRNA fragment using a class of tools called *aligners* (e.g., STAR [63], or Bowtie 2 [55]). This information is stored in a Sequence Alignment/Map (SAM) file or its compressed format known as BAM (i.e., Binary Alignment Map file) [37]

The starting biological material is usually abundant enough to require few amplification cycles. So, many issues present in single-cell data (e.g., amplification bias or allele dropout) are avoided. However, some biases are present also in bulk RNA-seq experiments, so proper preprocessing steps are required to make experiments comparable [94].

2.1.2 Single-cell sequencing experiments

Over the last years, further improvement of the RNA-seq sensitivity reduced the minimal input amount of biological material required. At the same time, the development of new microfluidic-based technologies enabled the automatic isolation of cells. Such advancements allowed to increase the resolution up to the single-cell level (scRNA-seq) [68]. scRNA-seq experiments produce a gene expression profile for each cell, scaling from hundreds to thousands or even millions. Accordingly, the data so generated allow one to investigate the subpopulation heterogeneity at the highest possible resolution.

Sequencing experiment protocols. The broad range of single-cell RNA sequencing protocols can be divided into two major groups: (*i*) full-length (e.g., Smarter and Smartseq2) [84], and (*ii*) UMI-based (e.g., Drop-seq and CEL-seq) protocols [51, 173].

The principal difference with respect to bulk sequencing lies in the library preparation steps, whereas, sequencing and alignment are conceptually identical. The main steps required to collect data from single cells are briefly illustrated below, in order to facilitate the comprehension of the properties of the distinct datasets, which primarily determine the computational preprocessing steps required prior to their analysis.

1. **Cell dissociation.** Single cells from biological samples, like blood samples, cancer biopsies, patient-derived organoids, or cell lines, are collected and dissociated to create a suspension of single cells. During this step cells could suffer of stress or even damages, which can reflect on their expression profiles.

2. **Cell isolation.** Cells isolation is performed via different techniques depending on the choice of the protocol. They can be grouped in *plate-based* or *droplet-based*. Typically, plate-based techniques are associated to full-length protocols, and droplet-based to UMI-based protocols (even though exceptions are possible, see [110]). With plate-based methods, each cell is usually sorted into a unique small plate using, for instance, a microfluidic sorter (e.g., Fluidigm C1 [152]) or via flow cytometry approaches. The number of cells collected ranges between 96 up to 800, which remain isolated until sequencing. Droplet-based techniques use a specific platform (e.g., 10x Chromium) that creates droplets containing a single cell and all the necessary enzymes for downstream chemical processes. The cell number increases from a thousand to many thousand. For a detailed overview of cells isolation method and platforms, please refer to [167].

3. **Library preparation.** During library preparation cells were lysed to extract mRNAs that are fragmented and amplified before sequencing. This is the main difference between full-length and UMI-based protocols. In the latter, the mRNA fragments from all the cells are mixed before the sequencing steps. So, the procedure requires adding specifically designed nucleotide sequences (i.e., tags) to each mRNA copy, in order to identify the cell of belonging and the original molecular transcripts. Such tags are attached via chemical reactions only to one transcript end (usually 3'-prime). Thus, only fragments that include the transcript end can be sequenced. On the other hand, plate-based methods keep mRNA fragments separated into their original plates. Adding a cell identifier is unnecessary, so all of them can be sequenced.

As anticipated, sequencing and alignment steps are similar to those described for bulk RNA-seq experiments. The few differences are reported below, which depend on the protocols. Usually, full-length sequencing protocols return one FASTQ file for every single cell. Instead, the UMI-base ones return a unique FASTQ file that must be demultiplexed into specific single cells, considering their identifiers. For more details on the technologies and applied protocols, please refer to [141, 215]. Please notice that Smartseq 2 and 10x Chromium protocols are the most widely used for full-length and UMI-based sequencing. The principal differences are the number of cells analyzed and the average coverage. Thus, researchers should select appropriate methods for their sample type and research aims. In particular, even if only a few cells can be analyzed simultaneously with a fluidigm C1 platform, it is possible to obtain in-depth transcriptome information for each cell. The genomic coverage is higher and more uniformly distributed along the coding regions. Instead, Chromium allows one to analyze thousand of cells, yet at the cost of less precise information on individual cells. Moreover, we note that, in generale, it is unfeasible to call somatic mutations from UMI-based protocol, because the coverage

involves only a tiny portion of each gene loci.

The BAM file obtained from full-length protocol have to be further processed to reduce amplification bias. In particular, the *Picard* tool, from the Broad Istitute can be used to detect and remove duplicated reads, which are supposed to be generated from the very same mRNA fragment. UMI-based protocol instead avoids amplification bias using a Unique Molecular Identifier to unequivocally identify each mRNA fragment.

2.1.3 Sequencing of viral samples

The sequencing of the viral genome obtained from biological samples, such as nasal swabs, present certain differences with respect to the methods described above. Most of the experiments are executed applying the Sanger sequencing technology [252]. This method returns the *consensus* sequence of the virus, which includes the base detected with highest relative frequency in every single genome position. Conversely, lower-frequency mutations are discarded, losing possible relevant information [262].

However, also viral samples can be analyzed with NGS technologies, for example, to obtain the full mutation spectrum and improve the downstream analyses as explained in [230]. Such deep sequencing data can be obtained following different protocols and sequencing technologies. One of the most commonly used approaches was proposed by ARTIC network (https://artic.network/). In particular, an Amplicon sequencing method for sequencing viral samples was optimized [133], and successively adjusted for the application to the SARS-CoV-2 genome [220]. In this approach, viral RNA is amplified by applying specific primers, avoiding human genetic material contamination, and by increasing the specificity of the RNA selected. The library prepared is then sequenced via Illumina or other technologies, with single-end or paired-end layouts. For additional details and other available sequencing methods, please refer to [230].

The workflow to obtain BAM files is similar to that applied to bulk or single-cell experiments. In the following, the main differences are reported.

- **Sample collection.** The viral genome is usually collected from infected hosts (e.g., via oral swab) and isolated. Notice that SARS-CoV-2 is a single-strand RNA virus. [229].
- **Reference genome.** One of the main differences between the analysis of either human or viral samples is the choice of the reference genome. In fact, a standard and curated reference genome of the SARS-CoV-2 currently does not exist. Many works employ as reference one of the two early sequenced genomes of SARS-CoV-2 [205, 206], that are identical except for five nucleotide positions. To this end, in paper P#6 we solved the discrepancy between the two genomes by considering Bat-CoV-RaTG13 genome and Pangolin-CoV genome, which were identified as closely related genomes to SARS-CoV-2 [208]

- **Alignment.** Instead of using a splice-aware aligner like STAR or Bowtie2, the Burrows-Wheeler Alignment tool (BWA) [36] is often used, since the viral genome does not include any untranslated regions. BWA returns BAM files, also in this case duplicated reads should be removed, especially if the library preparation was made via a RNA-seq based protocol.

In [238], the authors highlighted the importance of NGS in the investigation of the complexity of intra-host RNA viral populations. Such technologies could enable the development of more effective prevention strategies or antiviral therapeutics such as, e.g., structures of transmission networks, infection stage, and drug resistance.

2.2 Data type I: generation of gene expression profiles

From BAM files to read count matrix

The reads mapped in each BAM file are collected into genome features/loci (e.g., exons, transcripts, or genes), depending on the aim of the experiment and the average coverage. The results are positive matrices, which include either samples (from bulk experiments) or cells (single-cell experiments) as columns, and genome features as rows. Each entry represents the number of reads that are mapped in a given genome feature/locus for a given sample.

Data preprocessing. Both bulk and single-cell raw count matrices must be normalized before the downstream analyses. On the one hand, the preprocessing of bulk matrices is often applied to make sample expression profiles comparable (i.e., between sample normalization). In particular, the aim is to reduce biases due to different library sizes (i.e., different numbers of total reads obtained for each sample). Please refer to [150] for additional details on the available normalization methods.

On the other hand, single-cell expression matrices, particularly those produced with droplet-based protocols, require more work before using their gene count values. The preprocessing steps performed depend on the chosen protocol (i.e., UMI-based or full-length) and the experimental design. If interested, an exhaustive guide about the best practices in single-cell data analysis can be found in [186].

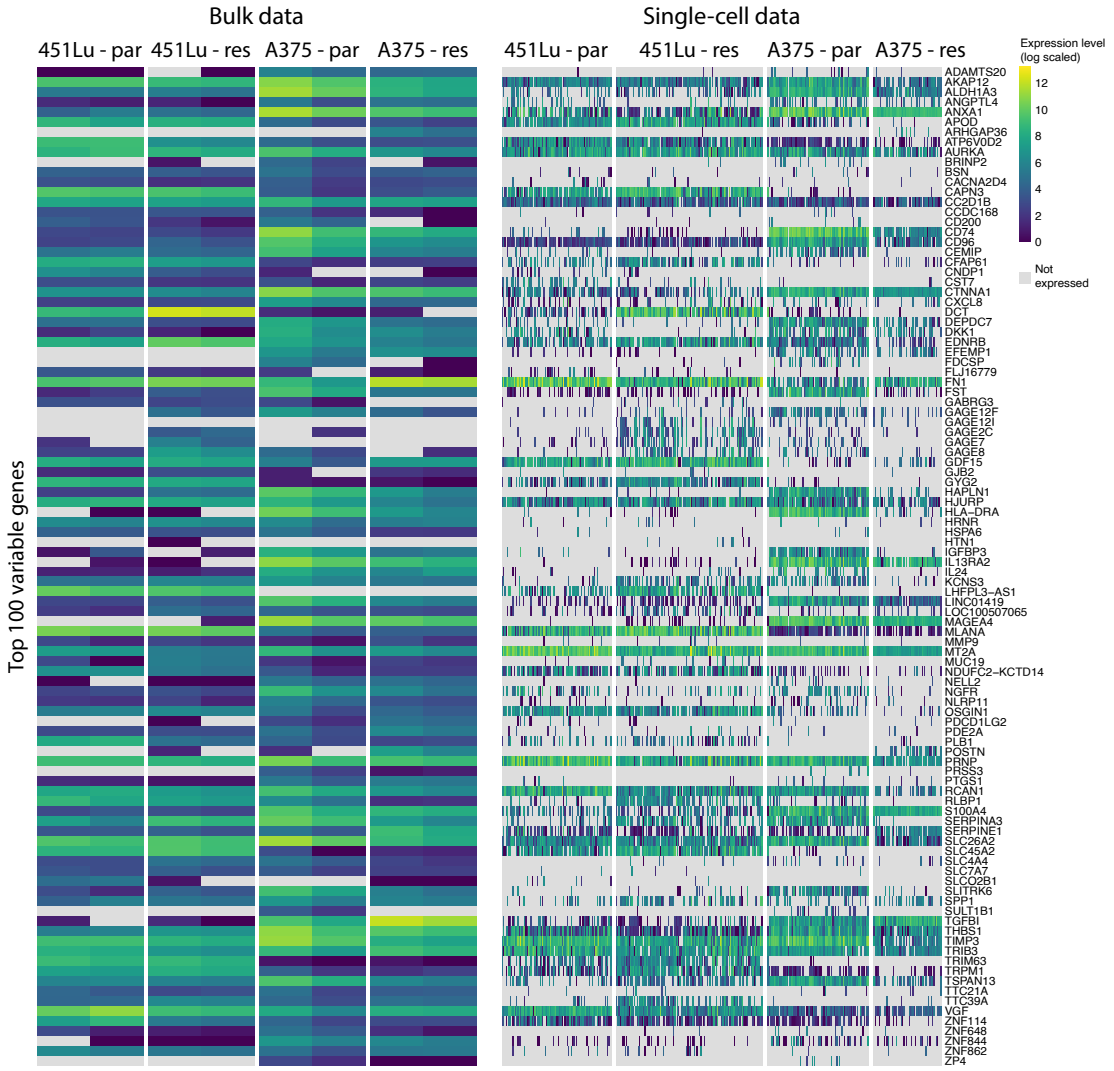


Figure 2.1: Comparison of gene expression profiles obtained via Bulk and single-cell sequencing protocol from [156]. Raw data are downloaded from GEO database with the following accession number: GSE108382 (bulk), and GSE108383 (single-cell). Top 100 variable genes were selected with Seurat toolkit considering only the single-cell dataset. Gene with less than 1 count are considered as not expressed. Expression values are log-scaled.

As an example, in figure 2.1 two typical read count matrices obtained via bulk and single-cell full-length sequencing protocol are reported. They are obtained from the dataset generated in [156] and downloaded from Gene Expression Omnibus database using the following accession number: GSE108382 (bulk), and GSE108383 (single-cell). It is possible to notice some major differences. First, the bulk matrix includes fewer non-

expressed genes (11.8%) than the single-cell matrix (60.3%). However, it is also evident that the high resolution of single-cell data enables the investigation of the subpopulation heterogeneity, despite the increased uncertainty in the data.

The data generated via both bulk and single-cell sequencing experiments can be used for downstream analyses, starting from the analysis of gene expression.

2.2.1 Comparative assessment of denoising and imputation methods for scRNA-seq data

Since the number of non-expressed genes in single-cell data is most likely inflated due to technical errors (i.e., capture efficiency, amplification bias, sequencing depth, and batch effect), many computational methods aimed at recovering the corrupted information in single-cell RNA-seq datasets have been developed. In general terms, most methods take as input noisy single-cell gene expression matrices (often unnormalized) and return a new matrix with the exact dimension of the original, but with recovered expression profiles. The denoised matrix should have lower 0-entries and more consistent data.

In all related papers, the authors claim to improve over the state-of-the-art, yet a fair comparative assessment executed by an independent research group might be a valuable resource, also to evaluate how protocols, technologies and data types might affect the overall performance. For these reasons, in paper P#1, we reviewed 19 methods for denoising and imputation of single-cell gene expression. In the review, both synthetic and real-world datasets are used. The former allowed us to compute quantitative metrics considering the *ground-true* expression profiles. Real-world datasets are also used to test the performances and to highlight the possible differences.

In particular, four different computational tasks were considered in the comparison: (*i*) imputation of dropout events, (*ii*) recovery of the gene expression profiles, (*iii*) characterization of cell similarity, and (*iv*) identification of Differentially Expressed Genes (DEG). The final aim is to propose a guideline for researchers to use the best method based on the specific dataset under investigation. Notice that, the supplementary information of the article is not included in this thesis. The complete documents is available on the online version of the manuscript [250].

Briefings in Bioinformatics, 22(4), 2021, 1–19

https://doi.org/10.1093/bib/bbaa222
Method Review

A review of computational strategies for denoising and imputation of single-cell transcriptomic data

Lucrezia Patruno, Davide Maspero, Francesco Craighero, Fabrizio Angaroni, Marco Antoniotti and Alex Graudenzi

Corresponding authors: Marco Antoniotti, Department of Informatics, Systems and Communication, University of Milan-Bicocca, Milan, Italy. Tel: +39 0264487901; E-mail: marco.antoniotti@unimib.it; Alex Graudenzi, Institute of Molecular Bioimaging and Physiology, Consiglio Nazionale delle Ricerche (IBFM-CNR), Segrate, Milan, Italy. Tel: +39 0221717551; E-mail: alex.graudenzi@ibfm.cnr.it
Lucrezia Patruno and Davide Maspero are equal contributors. Marco Antoniotti and Alex Graudenzi are co-senior authors.

Abstract

Motivation. The advancements of single-cell sequencing methods have paved the way for the characterization of cellular states at unprecedented resolution, revolutionizing the investigation on complex biological systems. Yet, single-cell sequencing experiments are hindered by several technical issues, which cause output data to be noisy, impacting the reliability of downstream analyses. Therefore, a growing number of data science methods has been proposed to recover lost or corrupted information from single-cell sequencing data. To date, however, no quantitative benchmarks have been proposed to evaluate such methods. **Results.** We present a comprehensive analysis of the state-of-the-art computational approaches for denoising and imputation of single-cell transcriptomic data, comparing their performance in different experimental scenarios. In detail, we compared 19 denoising and imputation methods, on both simulated and real-world datasets, with respect to several performance metrics related to imputation of dropout events, recovery of true expression profiles, characterization of cell similarity, identification of differentially expressed genes and computation time. The effectiveness and scalability of all methods were assessed with regard to distinct sequencing protocols, sample size and different levels of biological variability and technical noise. As a result, we identify a subset of versatile approaches exhibiting solid performances on most tests and show that certain algorithmic families prove effective on specific tasks but inefficient on others. Finally, most methods appear to benefit from the introduction of appropriate assumptions on noise distribution of biological processes.

Key words: denoising; imputation; single-cell RNA-sequencing; machine learning

Lucrezia Patruno is a PhD student in computer science at the Department of Informatics, Systems and Communication of the University of Milan-Bicocca. Her studies focus on data analysis and machine learning methods for the study of complex biological phenomena.
Davide Maspero is a PhD student in computer science at the Department of Informatics, Systems and Communication of the University of Milan-Bicocca. His studies focus on data integration methods for complex biological systems.
Francesco Craighero is a PhD student at the Department of Informatics, Systems and Communication of the University of Milan-Bicocca. His research is devoted to deep learning and explainable AI, with the aim of explaining deep networks inner sparse representations.
Fabrizio Angaroni is a postdoc researcher at the Department of Informatics, Systems and Communication of the University of Milan-Bicocca. His research is focused on mathematical methods for modeling and data analysis of complex biological system.
Marco Antoniotti is an associate professor at the Department of Informatics, Systems and Communication of the University of Milan-Bicocca. His main research topics are bioinformatics, computational systems biology, simulation, verification and cancer data analysis.
Alex Graudenzi is a research fellow at the IBFM-CNR. His research integrates (bio)informatics, complex systems, statistics and systems biology to deliver computational methods for the investigation of complex biological phenomena.
Submitted: 26 May 2020; **Received (in revised form):** 7 August 2020

© The Author(s) 2020. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Introduction

In recent years, an increasing number of studies has involved data generated from single-cell RNA sequencing (scRNA-seq) experiments [1, 2], which quantify gene expression levels at single-cell resolution, thus providing insights into cell population heterogeneity [3]. scRNA-seq methods can be used to perform accurate transcriptome quantification with a relatively small number of sequencing reads, isolating a typically large number of single cells. In optimal conditions, scRNA-seq data can recapitulate the results of standard sequencing experiments from bulk samples, yet with a much higher resolution [4].

This is a great advantage, as many works report that even cells in a homogeneous population may have heterogeneous expression profiles [5–8]. For instance, scRNA-seq data can be used to characterize rare cell subpopulations that had been hidden in the output of bulk RNA sequencing experiments [9], as well as in the analysis of cancer evolution, where they can be exploited to study the heterogeneity of tumor cell subpopulations [10] and the processes that lead to drug resistance or metastasis [11]. The wide use of scRNA-seq technologies has also allowed the creation of cell atlases for simple organisms such as, for example, the *Caenorhabditis elegans* [12]; most importantly, there is an ongoing effort to create such map for the human organism, i.e. the Human Cell Atlas [13]. However, the analysis of single-cell sequencing data is affected by the complex combination of biological variation and technical noise, which typically result in sparse and noisy single-cell expression profiles.

On the one hand, stochasticity of gene expression is inherent in most biological systems, with respect to both the biochemical processes related to gene regulation and the fluctuations of other cellular components and phenomena [14]. For this reason, even cells of the same type within the same tissue may display different gene expression distributions, complicating the identification and characterization of cellular states and transitions [15].

On the other hand, currently available sequencing technologies are still hindered by various technical issues [2, 16, 17]. In particular, the most common approaches for scRNA-seq are based on either droplet platforms (e.g. Drop-seq [18], InDrop [19] and Chromium 10x [20]) or plate-based platforms (e.g. Smart-Seq2 [21], MATQseq [22], MARS-seq [23], CEL-seq [24] and SPLIT-seq [25]), while some further approaches rely on microfluidics (e.g. C1 SMARTer [26]) or nanowell arrays (e.g. SEQ-well [27]). Typically, droplet platforms allow to isolate a large number of single cells (from a few to many thousands), by sequencing the 3′-end and by employing unique molecular identifiers (UMIs) [28], which allow the tagging of each transcript before amplification, thus distinguishing original transcripts from amplification duplicates [29]. Conversely, plate-based platforms usually employ full-length sequencing protocols and, accordingly, allow to sequence a much lower number of single cells (∼hundreds), yet with a considerably higher coverage. Overall, all sequencing protocols are affected by a number of technological and experimental issues, which typically result in noisy measurements.

- Capture efficiency: due to (i) the low quantity of RNA in a given single cell, and (ii) the stochastic nature of gene expression patterns at the single-cell level, certain gene can display null expression level, since none of its transcripts may be captured, thus resulting in zero expression levels. These are the so-called dropout events [30] and might be particularly relevant for scarcely expressed genes. This issue causes both noise and a high sparsity in the data [9].
- Amplification bias: the amplification phase may be subject to potential PCR biases in the quantification of the abundance of each gene, such as preferential amplification of certain templates. UMI-based approaches are able to mitigate this issue, yet in any case, amplification biases can be a potential source of noise in the data.
- Sequencing depth: the number of sequenced reads per cell varies between different experimental settings and platforms, and this can result in noisy and sparse outputs, especially when the depth is relatively low [29].
- Batch effects: technical sources of systematic variation may add a confounding factor in downstream analysis. Batch effects can be generated by analyzing samples with different technologies, in different laboratories or in different runs [31, 32]. When multiple experiments are considered, it is appropriate to remove such bias. In recent years, many methods were proposed to reach this goal. However, the comparison of the performance of methods for batch removal requires an in-depth investigation that is beyond the scope of this work (see [33] for a recent review).

As a consequence, it is safe to suppose that (i) nonzero expression values may not coincide with the true transcript abundance in the cell and (ii) zero values observed in the gene expression profiles may be either due to truly non-expressed genes—in this case, we refer to structural zeros, as proposed in [34]—or to technical limitations of the sequencing technology, i.e. dropout events.

For this reason, many computational approaches have been developed to retrieve lost and corrupted information from scRNA-seq data, with the goal of returning an estimation of the correct expression levels in each single cell. Such methods are typically grouped in two major categories: (*i*) imputation methods, with the general goal of recovering the missing values in the data and (*ii*) denoising methods, aimed at adjusting the data by removing biological and technical noise. Very often, the two categories are mentioned indistinctly (see e.g. [35]), even though they comprise substantially different computational tasks.

To better distinguish the two categories, here, we propose a rigorous categorization of imputation and denoising methods for scRNA-seq data, in order to reduce the possible ambiguity in the definition of the underlying computational tasks (an analogous distinction was recently proposed in [36]).

- Imputation methods for scRNA-seq data include two major steps. The first step is aimed at distinguishing structural zeros (associated to non-expressing genes) from dropout events (i.e. genes whose transcripts were not captured during the sequencing process due to technical issues). Accordingly, in the second step, such methods strive to impute the values of dropout entries only. Nonzero entries and structural zeros are left unchanged.
- Denoising methods for scRNA-seq data ideally include both an imputation step (see above) and an additional computational step, which is aimed at modifying the entries which include falsely increased or decreased gene expression levels due to, e.g. biological variation or technical noise. According to this definition, all denoising methods are also imputation methods while the opposite is typically not true (a rigorous definition of the two categories is provided in section 1 of the Supplementary Material).

Methods in both categories rely on different assumptions and employ different algorithmic strategies to perform their tasks.

Thus, as reported in [37], a comprehensive comparison of all available approaches might be useful and timely to clarify which methods are more suitable for different circumstances and distinct data types. In particular, in [37], the different approaches are grouped in the following typologies.

- Data smoothing: the methods in this category aggregate the expression profiles of similar cells in order to perform denoising and imputation. In this category, we find DrImpute [38], DEWÄKSS [39], scHinter [40], kNN-smoothing [41], LSImpute [42], MAGIC [43], netSmooth [44], PRIME [45] and RESCUE [46]. Finally, other methods that use data smoothing to impute missing values are G2S3 [47] and scTSSR [48]. However, the former aggregates the information across similar genes to perform imputation, while the latter considers both similar cells and similar genes.
- External knowledge integrators: these methods exploit external knowledge to impute or denoise gene expression profiles. In this category, we find ADImpute [49], netSmooth [44], netNMF-sc [50], SAVER-X [51], SCRABBLE [52], scNPF [53] TRANSLATE [54] and URSM [55].
- Machine learning (ML): these methods employ ML techniques to correct for technical noise. We can find very recent methods that employ Artificial Neural Networks (ANNs) to infer the denoised or imputed version of the dataset, which are AutoImpute [56], DeepImpute [57], DCA [58], EnImpute [59], GraphSCI [60], LATE [54], scIGANs [61], SAUCIE [62], scScope [63], scVI [64] and SISUA [65]. Next, we have methods that use regression to correct for noise in the dataset, which are 2DImpute [66] and RIA [67].
- Matrix theory: these methods decompose the observed gene expression matrix in a low-dimensional space to remove noise. In this category, we find ALRA [68], ENHANCE [69], scRMD [70], CMF-Impute [71], deepMc [72], McImpute [73], PBLR [74], WEDGE [75], ZIFA [76] and Randomly [77].
- Model-based: these methods make assumption on the statistical model of the distribution of technical and biological variability and noise and perform denoising and imputation by estimating the parameters of the distributions. In this category, we find bayNorm [78], BISCUIT [79], BUSseq [80], CIDR [81], MISC [82], SAVER [83], scImpute [84], scRecover [85], SCRIBE [86], SIMPLEs [87] and VIPER [88].

We here present a comparative assessment of denoising and imputation methods for scRNA-seq data, with the goal of providing a general overview of their features, strengths and limitations, in order to understand in which data analysis task they are most computationally and statistically efficient. In particular, we selected a subset of 19 different methods out of the list mentioned above, by including some of the most widely used approaches and which fall in the following categories.

- Data smoothing methods: DrImpute [38], kNN-smoothing [41] and MAGIC [43].
- ML methods: AutoImpute [56], DCA [58], DeepImpute [57], SAUCIE [62], SAVER-X [51], SCScope[63] and scVI [64].
- Matrix factorization/theory methods: ALRA [68], ENHANCE [69], McImpute [73], Randomly [77] and scRMD [70].
- Model-based methods: bayNorm [78], SAVER [83], scImpute [84] and VIPER [88].

The comparative assessment was carried out both on simulated data, generated via the widely used tool SymSim [89], and four real-world scRNA-seq datasets from [90–93]. All computational methods were tested with respect to a number of metrics, in order to assess the effectiveness in imputing dropout events, recovering the true expression profiles, characterizing the similarity among cells and improving the identification of differentially expressed genes (DEGs), in addition to quantify their scalability. In the Results section, we present the results of the extensive comparative assessment, also by releasing a summary for a quick evaluation of the distinct techniques in different scenarios and experimental settings.

We note that previous works reviewing imputation methods have been proposed. In particular, in [94], the authors focus on understanding whether six different imputation strategies introduce false positives in the results of differential expression analysis. In [95], eight different methods are analyzed to understand whether they improve the result of clustering and differential expression analysis. Both works, however, do not include in the analysis the most recent methods and assess the performance of a relatively limited number of computational strategies. In addition, both works mainly focus on the imputation task, without assessing how denoising techniques may recover corrupted information. Finally, a recent preprint on a similar subject [35] exploits real-world data to assess the performance of imputation methods on downstream analyses. While this work includes a more extensive assessment of recent methods, it does not employ simulated data, which are necessary to evaluate a number of ground truth (GT)-based performance metrics. Further comments in this respect are provided in the Discussion section.

In the Methods section, we provide a brief description of each denoising and imputation method included in the study, discuss the performance assessment describing both the synthetic data generation and the real-world datasets and present the different metrics used in the analysis. In the Results section, we present the results of the comparative assessment on both simulated and real data, also by releasing a summary for a quick evaluation of the distinct techniques in different scenarios and experimental settings. Finally, in the Discussion section, we draw conclusions about the comparison and discuss possible future developments.

Methods

In this section, we describe in detail the 19 methods included in the comparative assessment; we discuss the synthetic data generation and present the 4 real-world scRNA-seq datasets from [90–93] employed in the analysis, as well as the performance metrics.

Description of denoising and imputation methods

The 19 methods that have been analyzed and tested can be partitioned into the following four families, according to their assumptions and modeling techniques: smoothing, model-based, matrix factorization/theory and ML. In the following sections, we provide a brief description of each method. For additional details, we refer the reader to the original papers.

Data smoothing methods

The first category includes methods that aggregate the expression profiles of similar cells, e.g. by averaging the expression values, in order to impute (DrImpute) or denoise (MAGIC and kNN-smoothing) their expression values.

DrImpute [38] imputes dropout events with the following three steps: first, it computes a distance matrix between cells, then it runs the k-means algorithm and, lastly, it defines the expected value of a dropout event as the average value of that

gene over the cells belonging to the same cluster. To make the estimations more robust, the similarity matrix is computed with both Pearson and Spearman correlations and a range of number of clusters is tested. The averaged estimation over all combinations is taken as the final imputation value, reducing the risk of over-imputation.

kNN-smoothing [41] improves the signal-to-noise ratio of single-cell expression profiles with a two-phase algorithm: first, the k-nearest neighbors (kNNs) of each cell are identified, then the gene expression profile of each cell is smoothed by considering its neighbor profiles. The initial step of the algorithm is performed by normalizing the expression profiles and stabilizing their variance. Then, to overcome the problem of finding the best assignment for k, smoothing is applied in a progressive fashion, by starting from $k = 1$ and increasing k step-by-step until the desired level of smoothness is reached.

MAGIC [43] extracts the true similarity between cells by amplifying biological trends, while simultaneously filtering out spurious correspondences due to noise in the data. First, to overcome the problem of data sparsity, a nearest neighbor graph based on cell–cell expression distance is built. Then, an affinity matrix is defined by applying a Gaussian kernel on the principal components of the graph. Lastly, a diffusion process [96] is applied on the similarity matrix to obtain a smoothed, more faithful affinity matrix. The final imputation involves computing the new expression of each gene as a linear combination of the same expression in similar cells, weighted by the similarity strength obtained in the previous steps.

ML methods

This group includes methods that apply ANNs to solve the denoising problem (further details on ANN types are reported in section 2 of the Supplementary Material). As reported in [37], an increasing number of methods fall in this category (see above). In particular, we selected DeepImpute, DCA, SAVER-X, SAUCIE, scScope, AutoImpute and scVI.

AutoImpute [56] employs a sparse autoencoder, to learn the distribution of the input gene expression matrix and perform imputation. With regard to the implemented loss function, this method takes advantage of standard reconstruction errors such as (root) mean squared error, applied only on the nonzero expressed genes. After training the autoencoder (AE), the reconstructed matrix is taken as the imputed output.

DCA [58] employs AEs to perform denoising. Instead of the classical AE decoder output, it defines a parametric decoder that models each gene count as a negative binomial (NB) or a zero-inflated negative binomial (ZINB) distribution; consequently, the reconstruction error is defined as a likelihood. The predicted distribution is then used to generate the denoised output.

DeepImpute [57] employs a deep feedforward network (DFN) to perform imputation. After the initial preprocessing, where only relevant genes are kept, N random groups of genes G_i are defined. Then, for each gene in each G_i, a set I_i with the top five Pearson correlated genes not in G_i is built. Lastly, each I_i will be an input for a different DFN, trained to output G_i. The output of each DFN is then used for imputing dropout events.

SAUCIE [62] is an AE-based denoising method that also supports batch correction and enhanced clustering and visualization capabilites. More in detail, the AE embedding layer is used for both low-dimensional visualization and batch correction, by minimizing the difference between the probability distribution of layer's activations belonging to different batches. Moreover, the activations of the decoding part are binarized to define an encoding of each cell, which is then used for clustering. Lastly, denoising is performed by minimizing the reconstruction error, i.e. the mean squared error, that deals both with noise and dropout events.

SAVER-X [51] is an extension of SAVER [83] that pairs the Bayesian model with an AE. A NB distribution is used to model technical and biological noise, while the AE is used to estimate the portion of gene expression that is predictable by the other genes. Lastly, Bayesian shrinkage is used to compute a weighted average of the predicted expression values and the observed data, to get the final denoised value. Additionally, SAVER-X allows transfer learning [97] across species, thanks to the flexibility of AEs, allowing to extract information from data belonging to different species and experimental conditions.

scScope [63] exploits a deep learning approach for imputation, combining an AE with a recurrent layer. The architecture of the neural network is composed by a first layer that performs batch correction. Successively, the encoding and decoding layers of the AE perform compression and reconstruction, respectively, of the batch corrected input. Lastly, the imputation layer corrects the missing values and sends back the imputed output to the encoding-decoding layers, to re-learn a compressed representation. The loss function is defined as a standard reconstruction error, on the nonzero entries.

scVI [64] employs a variational AE to specify a ZINB distribution, which models the true gene expression. More in detail, the neural network takes as input each batch-annotated cell expression and successively learns a variational distribution accounting for, separately, the cell-specific scaling factor and the remaining gene variation; furthermore, the defined latent space allows to perform both clustering and visualization. Lastly, the ZINB distribution is specified based on the learned latent representation and the cell scaling factor.

Matrix factorization and matrix theory methods

The third category comprises four methods that denoise (ENHANCE) or impute (ALRA, McImpute and scRMD) the observed gene expression data by solving a matrix factorization problem [98]. For the sake of simplicity, we added to this category also a method that performs imputation by exploiting random matrix theory (RMT): Randomly.

ALRA [68] performs imputation by low-rank matrix completion [99] of the observed gene expression matrix. The algorithm is composed by two phases: firstly, a low-rank approximation with Singular Value Decomposition [100] is computed. Then, to distinguish dropouts from true zeros, the authors observed that biological zeros in the computed low-rank matrix are assigned to small values around 0, due to the approximation error. Consequently, by taking the magnitude of the smallest negative value of each gene as an approximation of the error, it is possible to define a gene-wise threshold to distinguish dropouts and extract the imputed values.

ENHANCE [69] is a method that combines PCA and cell aggregation using kNNs to denoise the observed count matrix. The algorithm can be divided into two main steps. The first one accounts for reducing the bias toward highly expressed genes, by aggregating the expression of similar cells based on the distance between their principal component scores. The second phase projects the aggregate matrix on the first k principal components, where k is selected to represent only true biological differences. Lastly, the selected components are used to derive the final denoised matrix.

McImpute [73] is a low-rank matrix completion approach to impute missing values in a gene expression matrix. This method aims at finding a lower-dimensional decomposition of the input matrix. They formulated a low-dimensional nonnegative matrix factorization problem as an optimization problem, solved using the majorization-maximization technique [101]. To ensure the convexity of the problem, McImpute solves a relaxed version of the original objective: nuclear norm minimization. Lastly, the resulting decomposition is used to impute missing values.

Randomly [77] is a recent denoising method that extracts the true biological signal from the gene expression data by analyzing the eigenvector statistics predicted by RMT [102]. The algorithm is composed by three steps. In the pre-processing step, expression counts are normalized and genes contributing to a sparsity-induced nonbiological signal are removed; then, the random matrix accounting for the noise is estimated. Lastly, the eigenvalues carrying the true biological signal are extracted following RMT, providing a low-rank representation of the input data; additionally, the genes that are mostly responsible for the signal directions can be separated from the less relevant ones.

ScRMD [70] is a method that approaches the imputation task by means of a robust matrix decomposition (RMD) approach [103]. The authors assumed that we can decompose each gene expression in the following components: the mean expression of cells belonging to the same cluster, the specific cell variability, the measurement error and the dropouts events. The method defines each component as a matrix decomposition problem, solved with an alternating direction method of multiplier, by also applying a regularizer to account for the low-rank of the biological signal and the sparseness of the observed counts.

Model-based methods

This category is composed by methods that model the observed expression value of each gene in each cell as a random variable and perform imputation (scImpute and VIPER) and denoising (bayNorm and SAVER) by estimating the parameters of their distributions.

bayNorm [78] employs a Bayesian approach to perform denoising. The posterior distribution of the original counts is composed by (i) the likelihood of obtaining the observed transcripts, modeled as a Binomial distribution, and (ii) a prior on each gene expression value. In order to model biological variability, bayNorm employs a prior on the underlying true gene expression levels, by modeling them as variables following an NB distribution. Parameters can then be estimated locally or globally, depending on one's interest in amplifying or not, respectively, the intergroup differences between cells.

SAVER [83] estimates the true gene expression levels by modeling observed counts as a NB distribution. More in detail, the technical noise in the gene expression signal is approximated by the Poisson distribution, while the gamma prior accounts for the uncertainty in the true expression. The final recovered expression is a weighted average of the normalized observed counts and the predicted true counts.

scImpute [84] is a method that performs imputation, in a three-step algorithm. Initially, it identifies subpopulations of cells by first applying PCA and, successively, spectral clustering [104] on the remaining dimensions. To infer which genes are affected by dropout, it models genes in each subpopulation with a gamma-normal mixture model. Lastly, only highly probable dropout events are considered, to reduce over-imputation, and the final imputation value is computed as a linear combination of the expression of the other cells in the same subpopulation, weighted by the pairwise similarity.

VIPER [88] is an imputation method composed of four phases. The first step performs a pre-selection of candidate similar cells, to reduce overfitting. Then, a least-squares method is used to choose a local neighborhood for each cell. To prevent imputing missing values, VIPER estimates the dropout probability and the expected expression for each zero-valued neighbor. Furthermore, to adjust dropout events, the gene expressions in each neighborhood are assumed to follow a zero-inflated Poisson mixed model, estimated using expectation maximization. Lastly, imputation is performed by computing the weighted sum of the expression of each neighbor, by also taking into account the computed dropout adjustments.

Performance assessment

In the original articles, the imputation and denoising methods introduced above are often compared with competing approaches. However, such comparisons typically involve a limited number of denoising methods and a small number of selected experimental settings. In order to provide a comprehensive evaluation of performances, in this work, we tested all methods on a large number of both simulated and real-world datasets, with respect to several metrics.

In particular, we generated an extensive array of simulated data, for which the GT is available and which allow to quantify the ability of each method to actually recover the lost information (see Supplementary Material section 3 for details about the generation of such data). Moreover, we tested all methods on four real-world scRNA-seq datasets generated via distinct experimental protocols and settings.

Simulations

We employed the tool SymSim [89] to generate a large number of synthetic scRNA-seq datasets (for a total of 90 distinct synthetic datasets). SymSim takes as input the number of single cells, the number of genes, the number of cell subpopulations (characterized by distinct gene expression patterns) and a number of parameters that tune the amount of biological variability and technical noise.

The tool returns as output (*i*) a GT expression matrix, which includes biological variability but no noise; (ii) a theoretical expression profile (TEP) for each cell subpopulation, which is obtained by removing the biological variability from the GT; and (iii) a noisy (and sparse) expression matrix (NEM), which is finally derived by simulating the steps of a sequencing experiment.

In this work, we generated datasets simulating two main experimental scenarios and, in particular,

(i) non-UMI full-length datasets (i.e. high-coverage, high-amplification bias), including 100 single cells and modeling a typical plate-based full-length sequencing experiment (e.g. Smart-Seq2). Thirty datasets were generated with distinct parameter settings;
(ii) UMI datasets (i.e. low-coverage, low-amplification bias), including 3000 single cells (30 datasets) and 10000 single cells (30 datasets) and modeling a typical droplet sequencing experiment (e.g. Chromium 10x).

The different datasets in each scenario are characterized by distinct parameter settings, in terms of number of cell subpopulations ($\{3, 5\}$), noise level (5 levels), number of selected (most variable) genes ($\{500, 2000, 10000\}$) (Table 1). A detailed

Table 1. Summary of the simulated datasets. We simulated a total of 90 datasets, with the following combinations of parameters: 3 values of sample size (number of single cells) ×2 different numbers of subpopulations ×5 levels of noise ×3 numbers of selected most variable genes

Protocol	UMI	Non-UMI full-length
No datasets	60	30
No cells	{3000; 10000}	100
GT sub-populations	{3; 5}	{3; 5}
Capture efficiency	Low	High
Amplification Bias	No	Yes
Coverage	Low	High
Noise level[a]	{1; 2; 3; 4; 5}	{1; 2; 3; 4; 5}
No genes	{500; 2000; 10000}	{500; 2000; 10000}

[a] The levels of noise present in the simulated datasets are defined in section 3 of the Supplementary Material, which we refer for further details on synthetic data generation.

description of synthetic data generation can be found in the Supplementary Material section 3.

Real-world datasets

All methods were tested also on four distinct real-world scRNA-seq datasets, generated with distinct protocols and experimental specifications. In detail, we have the following.

- **RW-D** #1 (PBMCs – 10x) [90]: this widely employed scRNA-seq dataset is generated via 10x Genomics platform [20] and includes 68579 peripheral blood mononuclear cells (PBMCs), which are annotated with 11 cell types of the immune system, via correlation with benchmark gene expression profiles. This dataset was used in our analysis to assess the performance of imputation and denoising methods in characterizing cell similarities (for further details on the dataset, please refer to [90]; instructions for download are provided in the Supplementary Material).
- **RW-D** #2 (lung cell lines – 10x) [91]: this scRNA-seq dataset is generated via the 10x Genomics platform and includes 3918 cells from 5 distinct cell lines, which were assigned to its corresponding identity by exploiting known genetic differences (i.e. SNPs) between cell lines [91]; this allows not to rely on gene expression profiles for cell labeling. We employed this dataset to assess the robustness of the characterization of cell similarity.
- **RW-D** #3 (pancreatic islets – Smart-Seq2) [92]: this scRNA-seq dataset is generated via the full-length Smart-Seq2 protocol and includes 3514 cells from human pancreatic islets of four diabetic patients and five healthy samples. We employed this dataset to assess the performance of imputation and denoising methods with respect to cell similarity characterization when processing data from non-UMI full-length protocols.
- **RW-D** #4 (melanoma cell lines – 10x, Fluidigm/Smart-Seq, bulk) [93]: this dataset includes three different measurements from the same biological samples, namely (i) bulk RNA-seq experiments, (ii) 10x Genomics scRNA-seq experiments with 737280 barcodes, (iii) Fluidigm/Smart-Seq scRNA-seq experiments with approximately 100 single cells. Since no cell type labels are provided in this dataset, we here used the data to compare the performance of imputation and denoising methods with respect to the correct identification of DEGs, by setting the results obtained on bulk data as baseline.

All real-world datasets were preprocessed to consider only high-quality single cells, and downsampled, to ensure a uniform assessment scheme for all methods. In Table 2, one can find the main features of all datasets employed in the analyses (see Supplementary Material section 4 for further details on preprocessing and downsampling).

Performance metrics

To evaluate the performance of the 19 selected methods, we employed a number of metrics, which were assessed with respect to either simulated or real-world data, according to the specific cases. All metrics are further detailed in section 5 of the Supplementary Material.

Imputation of dropout events (simulations) The effectiveness of the methods in identifying and correcting dropouts events can be evaluated by employing the GT expression matrix obtained from simulations (see Supplementary Material section 5 for additional details). In order to quantify the correct imputation of the dropout entries present in the GT, we employed three distinct metrics.

In particular, we computed (i and ii) precision and recall on dropout entries only (i.e. entries that are > 0 in the GT and are $= 0$ in the NEM), (iii) the Spearman correlation delta between the imputed/denoised expression matrix (for the sake of readability, we will refer to as denoised expression matrix, from now on) and the GT with respect to all the zero entries in the NEM, which allows to evaluate how imputed entries are correlated with GT values (this metric is shown in the Supplementary Material section 5).

Notice that the false discovery rate (FDR) can be easily determined from precision (FDR $= 1-$precision) and, in this case, allows to evaluate the effectiveness of the methods in not imputing structural zeros (i.e. entries that are 0 both in the GT and in the NEM).

Recovery of true gene expression profiles (simulations) To estimate the ability of each method in recovering the true single-cell gene expression profiles, we relied on both the GT and the NEM obtained from simulations.

In particular, we computed the difference between the Spearman correlation coefficient ρ computed after imputation or denoising (i.e. ρ between denoised expression matrix and GT) and that computed before imputation or denoising (i.e. ρ between NEM and GT). This measure is denoted as delta correlation in the following, $\Delta\rho$.

Characterization of cell similarity (simulations and real-world data) In order to evaluate the effectiveness of each method in capturing the similarity among cells, we computed the average silhouette coefficient (or width) [105] by grouping single cells according to the GT labels, i.e. cell subpopulations labels for both simulated data, and cell type/line labels for real data. Higher values of the average silhouette coefficient indicate that cells are grouped consistently with GT labels. Therefore, we here measured the difference between the average silhouette coefficient obtained from denoised data and that computed from the NEM (i.e. silhouette delta). Further detail about the evaluation of such metric is given in the Supplementary Material section 5.

We finally remark that, with regard to simulations, we here employed the TEP of all cell subpopulations as performance benchmark.

Table 2. Features of real-world datasets. Main features of the four real-world datasets used in the assessment of imputation and denoising methods: **RW-D**#1 [90], **RW-D**#2 [91], **RW-D**#3 [92] and **RW-D**#4 [93]

Dataset			Number of cells		
RW-D	Name	Protocol	Original	Employed	Task
#1	PBMC [90]	UMI	68579	3000	Cell sim.
		UMI	68579	10000	Cell sim.
#2	Lung cell lines [91]	UMI	3918	3918	Cell sim.
#3	Pancreatic islets [92]	Non-UMI	352	245	Cell sim.
		Non-UMI	383	243	Cell sim.
		Non-UMI	383	197	Cell sim.
		Non-UMI	383	224	Cell sim.
		Non-UMI	383	196	Cell sim.
		Non-UMI	383	263	Cell sim.
		Non-UMI	383	93	Cell sim.
		Non-UMI	384	275	Cell sim.
		Non-UMI	384	293	Cell sim.
#4	Sake (Parent.) [93]	UMI	737280	3178	DEGs
		Non-UMI	113	113	DEGs
	Sake (Resist.) [93]	UMI	737280	3085	DEGs
		Non-UMI	84	84	DEGs

Identification of DEGs (real-world data)
To assess the improvement on the identification of DEGs due to the application of imputation/denoising methods, we employed real-world dataset **RW-D**#4 which includes two independent cell populations, namely parental and resistant, for which single-cell 10x, single-cell Fluidigm/Smart-Seq and bulk sequencing experiments were executed.

We proceeded as follows: for each single-cell dataset (10x and Fluidigm/Smart-Seq), we performed a standard Wilcoxon test to select the DEGs ($p < 0.05$) between parental and resistant populations, with respect to both the NEM and the denoised expression matrix, and which results in two distinct lists of DEGs.

The expression profiles of the DEGs are then used to calculate the Spearman correlation coefficient between each single cell and the corresponding bulk profile. The distribution of the difference of the Spearman correlation coefficient as computed on denoised data and that on the NEM is used to evaluate the performance for this task.

Computation time (simulations) We finally analyzed the computational time of each tested method to impute or denoise datasets with distinct numbers of observations (i.e. single cells) and of variables (i.e. genes), with respect to a selected number of simulated datasets. All computations were performed on a HP® Z8 G4 Workstation equipped with two Intel® Xeon® Gold 6240 processors at 2.60 GHz, 1 TB DDR4 RAM at 2933 MHz and Linux Mint 19.2 Tina.

We note that, in the original papers, the authors do not declare any theoretical worst-case performance in terms of $O(\cdot)$ notation; although for many of them, it would be derivable from literature. We therefore present an empirical study of the relative performances of the methods.

Parameter settings of computational methods

Most methods were run on both simulated and real-world datasets using default settings and following guidelines provided from the authors, if any. For additional details on parameter settings of all methods, please refer to section 6 of the Supplementary Material and to Supplementary Table 4.

Note that we report the results SAVER-X without pre-training, as its performance seems to be only slightly affected by pre-training on real-world datasets, as shown in Supplementary Figure 9. Besides, for analyses involving synthetic datasets, we did not run AutoImpute, McImpute, scImpute and VIPER on datasets with 10 000 cells and 10 000 genes, and we did not execute VIPER on **RW-D**#1 (downsampled to 10 000 cells and 10 000 genes), due to the high computational time required by such methods. Furthermore, for 10 out of 30 non-UMI full-length simulated datasets, SAUCIE collapsed all cells into one unique profile. Thus, such datasets were not included in the analysis. Finally, please note that for Fluidigm/Smart-Seq datasets in **RW-D**#4, the computation of bayNorm and ENHANCE raised errors and, therefore, their results are not reported.

Results

We start by providing a qualitative example of the effect of the tested imputation and denoising methods: Figures 1 and 2 show the tSNE low-dimensional representation [106] of a synthetic dataset (3000 cells, 5 subpopulations and 2000 genes) and of one real dataset (**RW-D**#1, downsampled to 3000 cells and 2000 genes; see the Methods section for further details). For the synthetic dataset, we show the GT expression matrix, the NEM and the denoised datasets returned by each method, whereas for **RW-D**#1 we show its original expression matrix and the corresponding denoised versions.

From this qualitative analysis, one can appreciate the substantial different data transformations which are determined by the distinct methods.

While it is difficult to draw conclusion from single experiments, certain methods apparently tend to reduce the variability of gene expression profiles, resulting in more compact representations on the tSNE space (e.g. kNN-smoothing, SAUCIE, MAGIC), some others appear to enhance the inter-cluster distance (scImpute, SAVER and ENHANCE), while most methods seem to preserve the original disposition in the transcriptomic space, with some exceptions (note that in this and subsequent analyses, AutoImpute seems not to have reached convergence, with default parameters).

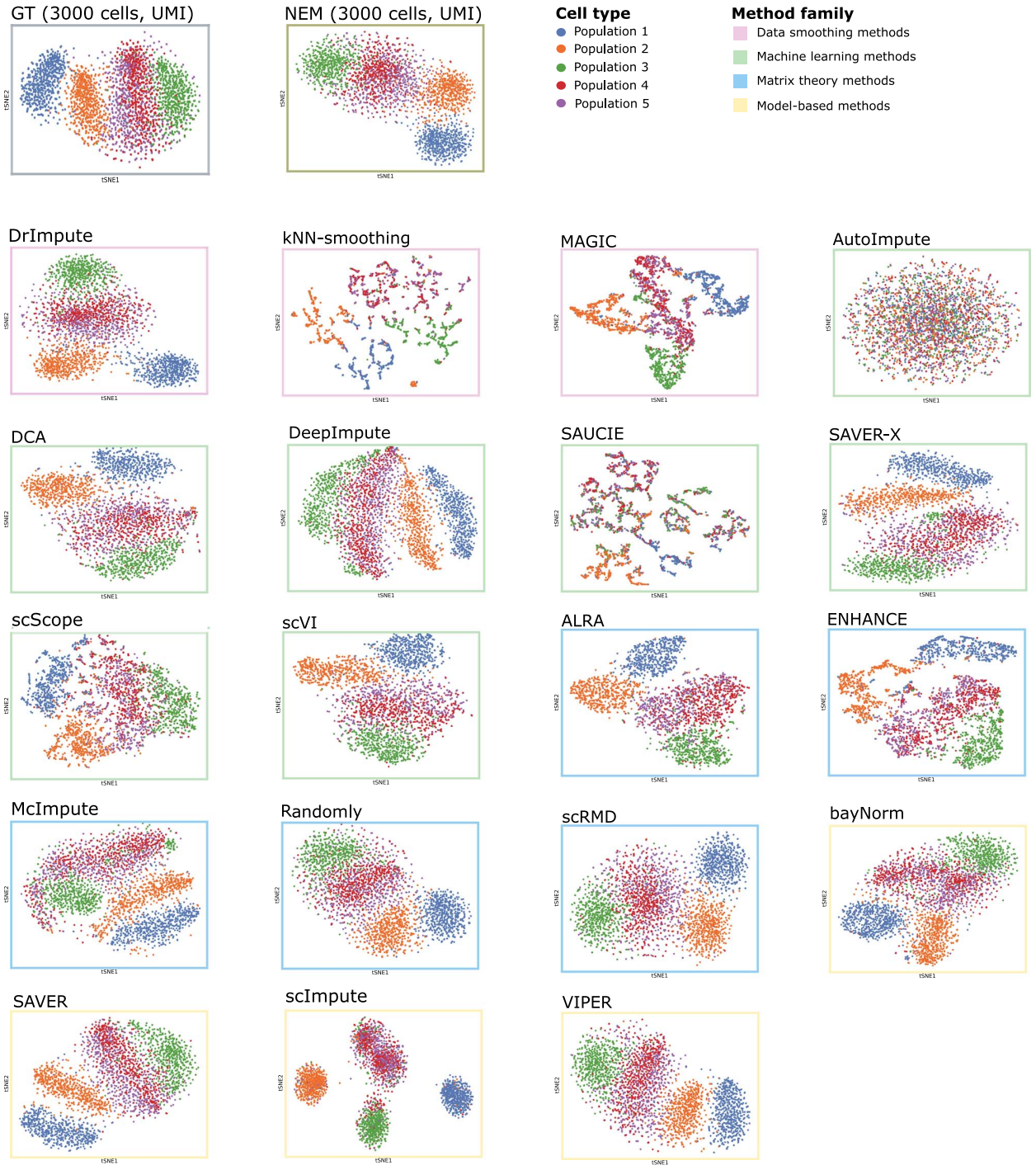


Figure 1. Effect of 19 imputation and denoising methods on a selected simulated scenario via tSNE low-dimensional representation. tSNE low-dimensional representation [106] of the gene expression profile of 3000 single cells of a selected synthetic UMI dataset with 5 subpopulations and 2000 genes. For this dataset, we present the tSNE plot of the GT expression matrix generated via SymSim and the NEM obtained after simulating the sequencing experiment. The remaining tSNE plots represent the gene expression of the cells after the application of all tested denoising and imputation methods to the NEM.

The visualization of three further synthetic datasets and of real-world datasets **RW-D**#2 and **RW-D**#3 are shown in Supplementary Figures 1–5. The results of the quantitative assessment with respect to the metrics described in the Methods section are presented in the following.

Imputation of dropout events (simulations)

We first assessed the performance of all methods in imputing dropout events (i.e. entries $= 0$ in the NEM but > 0 in the GT expression matrix), leaving structural zeros unchanged (i.e. entries $= 0$ both in the NEM and the GT). The parameters of all

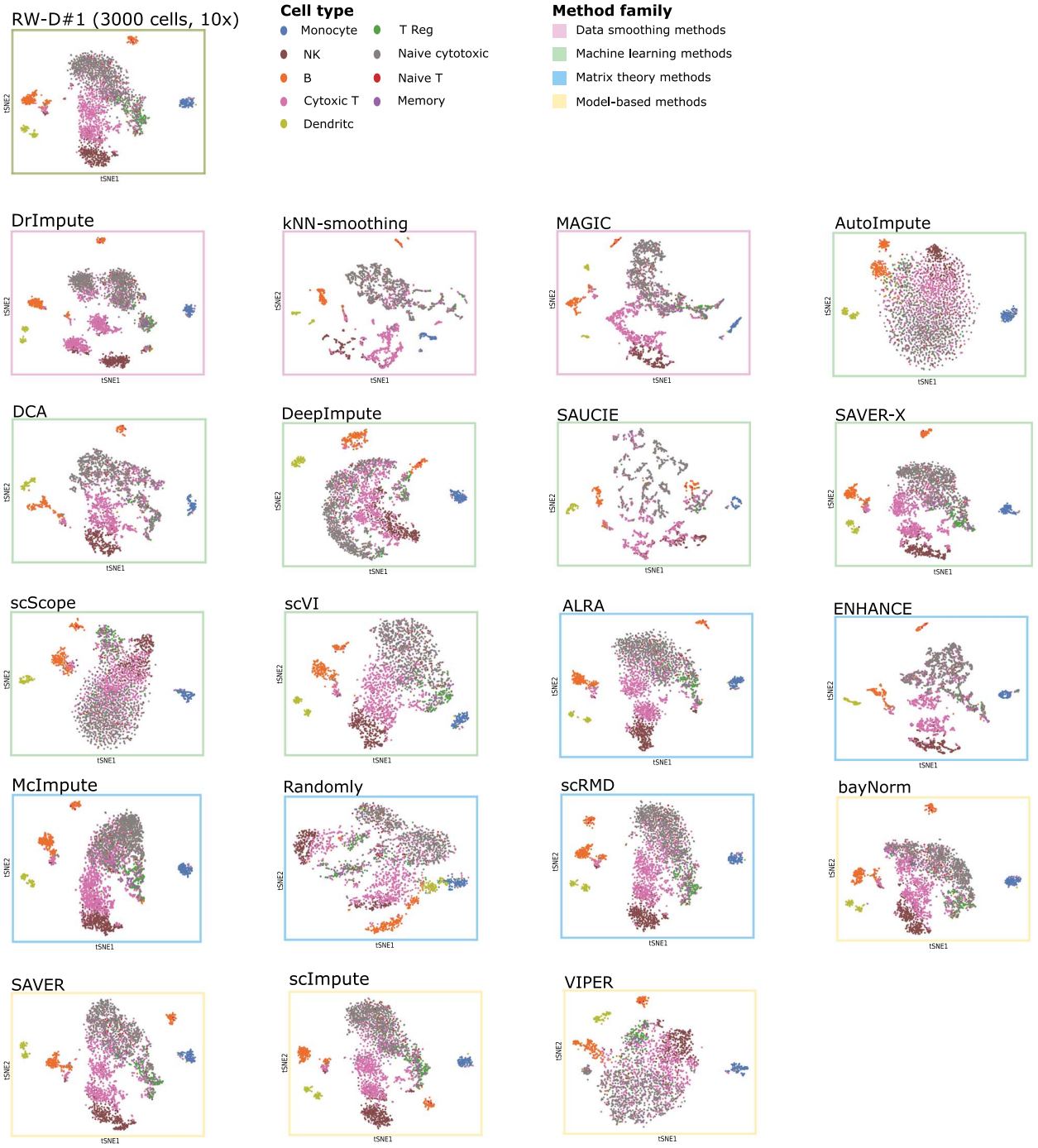


Figure 2. Effect of 19 imputation and denoising methods on real-world dataset RW-D #1 via tSNE low-dimensional representation. tSNE low-dimensional representation [106] of the gene expression profile of 3000 selected cells from RW-D #1 (PBMCs – 10x) [90] as computed on the 2000 most variable genes. For this dataset, we present the tSNE projection of the original dataset, which includes nine cell types and the tSNE plots of the single-cell expression profiles after the application of all methods under analysis.

simulations are recapitulated in Table 1 and in Supplementary Tables 1 and 2. Please refer to the Methods section and to Supplementary Material sections 3 and 5 for details on synthetic data generation and performance metrics. Note that Randomly was not included in this test, since it provides an already scaled expression matrix as output.

In Figure 3, one can find, for each method, the median precision and recall on correctly imputed dropouts (in this case, a true positive is an entry > 0 both in the GT and in the denoised expression matrix but $= 0$ in the NEM), grouped according to the number of (most variable) selected genes ($\{500, 2000, 10\,000\}$) and the number of single cells (100 for non-UMI full-length

Figure 3. Performance assessment on imputation of dropout events (simulations). Assessment of imputation of dropouts, as evaluated on non-UMI full-length simulated datasets (100 single cells) and UMI simulated datasets ({3000, 10000} single cells), with {500, 2000, 10000} genes. In each panel, we display a scatter-plot returning, for each imputation and denoising method, the median precision (y-axis) and recall (x-axis) as computed on correctly imputed dropouts (computed on 10 simulations per setting). In this case, a true positive, is an entry that is > 0 in the denoised expression matrix and in the GT but is $= 0$ in the NEM (see the Methods section for further details and Supplementary Table 3 for the confusion matrix). The squared shade indicates methods with precision and recall > 0.80. In Supplementary Figure 6, the distribution of precision and recall is displayed.

and {3000, 10000} for UMI datasets). In order to identify the methods showing high precision (i.e. how many imputed entries are dropouts) and high recall (i.e. how many dropouts are imputed) scatter-plot areas corresponding to high values for both measures ($>$ 0.80) were highlighted (in Supplementary Figure 6 the distributions of precision and recall on settings are displayed).

As a first result, most methods struggle when dealing with non-UMI full-length datasets (with 100 cells), as proven by the relatively lower value of average precision. This aspect is likely due to the low number of observations (single cells) as compared with the number of variables (genes) and consistently affects the performance of all methods on most tasks (see below).

Conversely, we observe a subset of methods that achieve extremely positive performances (both precision and recall $>$ 0.80) for UMI datasets with 3000 and 10 000 cells. In detail, VIPER provides the best performance with datasets with 500 genes, while for datasets with 2000 and 10 000 genes, ALRA, bayNorm, DrImpute, ENHANCE, kNN-smoothing, MAGIC, SAVER, SAVER-X and scVI consistently provide optimal and analogous performances. In particular, such methods show values of recall very close to 1 in all experimental settings (with the exception of kNN-smoothing). While this effect might be due to over-imputation, such methods also display significantly high precision in most settings. Notice also that higher values of precision implicate a lower fraction of wrongly imputed structural zeros (entries $= 0$ both in the GT and the NEM), as measured by the false discovery rate (FDR $= 1-$ precision).

Finally, we note that scRMD and scImpute display the highest values of precision in most settings, which, however, are most likely due to the conservative nature of the approaches, which tend to limit the number of imputed values. This observation is strengthened by considering the low values of recall for both methods: indeed, as recall corresponds to the fraction of imputed dropouts, a value close to 0 indicates that the method did not impute most of the events.

To further extend the analysis on imputation of dropouts, in Supplementary Material section 7 (Supplementary Figure 7), we return the analysis of the Spearman correlation coefficient computed considering zero entries of the NEM and which allows to quantify the correlation between imputed entries and the corresponding GT expression values. On the one hand, bayNorm, DrImpute, ENHANCE, MAGIC, SAVER and SAVER-X provide the most accurate and robust results in most scenarios, proving effective in correctly recovering the true expression values of imputed entries. On the other hand, ALRA, kNN-smoothing and scVI and VIPER, which exhibit good values of precision and recall on imputed dropouts (see above), display a relatively lower performance in terms of correlation of the imputed entries with respect to the GT expression values.

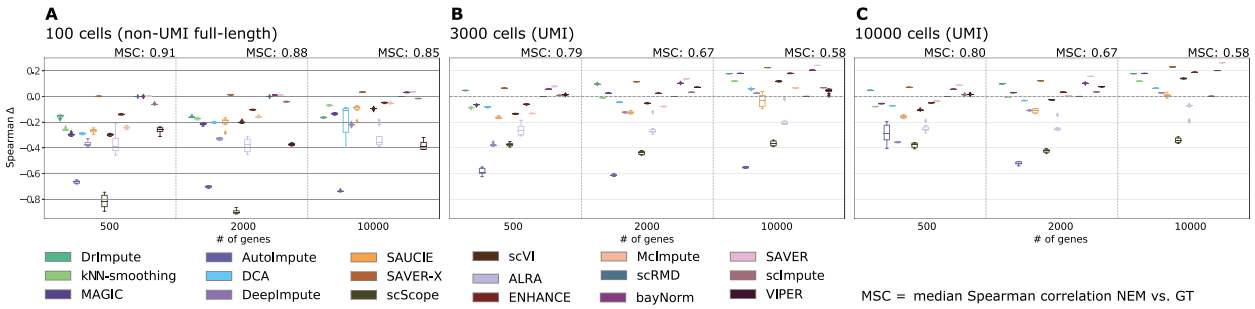


Figure 4. Performance assessment on recovery of true gene expression profiles (simulations). Assessment of recovery of true expression profiles, as evaluated on non-UMI full-length simulated datasets (100 single cells, panel A) and UMI simulated datasets ({3000, 10000} single cells, panels B and C), with {500, 2000, 10000} genes. The boxplots return the distribution of correlation delta, $\Delta\rho$, i.e. the difference between the Spearman correlation coefficient computed between the denoised expression matrix and the GT and that computed between the NEM and the GT, for all methods in each experimental setting. The baseline median Spearman correlation coefficient (MSC) between the NEM and GT is reported on top of the panels, for each setting, while in Supplementary Figure 8, the relative distributions are returned.

Recovery of true gene expression profiles (simulations)

We next tested the capability of each method in recovering the GT gene expression profiles, by using simulated data. In Figure 4, one can find the difference of the Spearman correlation coefficient as computed between the GT and the denoised expression matrix after the application of all 19 methods and that computed between the GT and the NEM. Such difference is denoted as correlation delta, $\Delta\rho$, from now on (see the Methods section and Supplementary Material section 5 for further details).

In particular, the results are displayed according to the number of genes, {500, 2000, 10000} and number of cells, 100 for non-UMI full-length and {3000, 10 000} for UMI experiments, as this allows to analyze the performance under different experimental settings. Note that, as for the analysis on imputed entries, we here do not include the output of Randomly, which provides a scaled output matrix.

As expected, sample size and protocol-type highly influence the capability of any method to recover corrupted information, as the performance of all methods generally improves with datasets with a larger number of single cells and generated via UMI-based protocols. More specifically, most methods appear to struggle when processing non-UMI full-length datasets characterized by a low number of cells (i.e. = 100), delivering unreliable and often erroneous denoised expression profiles, as proven by the negative Spearman correlation delta observed in most cases (up to −0.45 for some methods).

Conversely, correlation deltas progressively improve with UMI datasets including larger numbers of cells and/or genes, and, in particular, all methods with the exception of ALRA and scScope, achieve a positive median delta with datasets with 10 000 genes and 10 000 cells.

Examining the methods in greater detail, we observe that bayNorm, SAVER and SAVER-X are the methods with the best overall performance, as they always provide a positive correlation delta and achieve the best results with both non-UMI full-length and UMI datasets. Furthermore, we note that such approaches show an extremely low variance, suggesting that the results are robust. Among the other approaches, we note that DrImpute displays a high correlation delta with UMI datasets, whereas both ENHANCE and MAGIC exhibit remarkable performances with datasets with more than 3000 cells and more than 2000 genes.

All in all, the results of this and the previous analyses suggest that bayNorm, SAVER and SAVER-X might be an adequate choice for both imputing dropouts and recovering corrupted information, as they show the most accurate and stable performances with both UMI and non-UMI full-length datasets, whereas DrImpute, ENHANCE and MAGIC are similarly effective when processing UMI datasets.

Characterization of cell similarity (simulations and real-world data)

When analyzing scRNA-seq data, one might be interested in characterizing the possible heterogeneous populations included in the dataset, typically performing unsupervised clustering. For example, the Scanpy [107] and Seurat [108] packages for single-cell analyses incorporate the Louvain and Leiden algorithms for community detection [109], which identify clusters based on a nearest neighbors graph constructed from the profiles of each single cell. Therefore, it is clear that improving the identification of cell similarities might result in better clustering performances. To this end, we assessed the effectiveness of all tested methods in enhancing cell similarity with respect to both simulated and real data.

In Figure 5, we show the difference between the average silhouette coefficient computed on denoised expression matrix and that obtained from the NEM, by grouping single cells according to the GT labels. Higher values of the average Silhouette coefficient indicate that cells are close to other cells of the same subpopulation and separated from those belonging to other subpopulations. In particular, GT labels are provided by cell subpopulation labels for simulated data and by cell type/line labels for real-world datasets (see the Methods section and the Supplementary Material for further details). We remark that the silhouette coefficient allows one not to rely on arbitrarily chosen clustering approaches, to evaluate the correct grouping of single cells. In fact, currently available clustering methods for scRNA-seq data are characterized by different properties, goals and specifications and produce results that are extremely sensitive to parameter choices and variations, and which might, in turn, undermine the comparison of denoising and imputation methods on this specific task.

Results are shown for simulated datasets with {500, 2000, 10000} genes and 100 (non-UMI full-length) or {3000, 10000} single cells (UMI), as well as for real-world datasets **RW-D**#1, **RW-D**#2 and **RW-D**#3. Note that we employed the TEP of all cell subpopulations as benchmark for the assessment on simulated datasets: in particular, the silhouette coefficient delta between the TEP and the NEM represents the largest theoretical improvement in each setting.

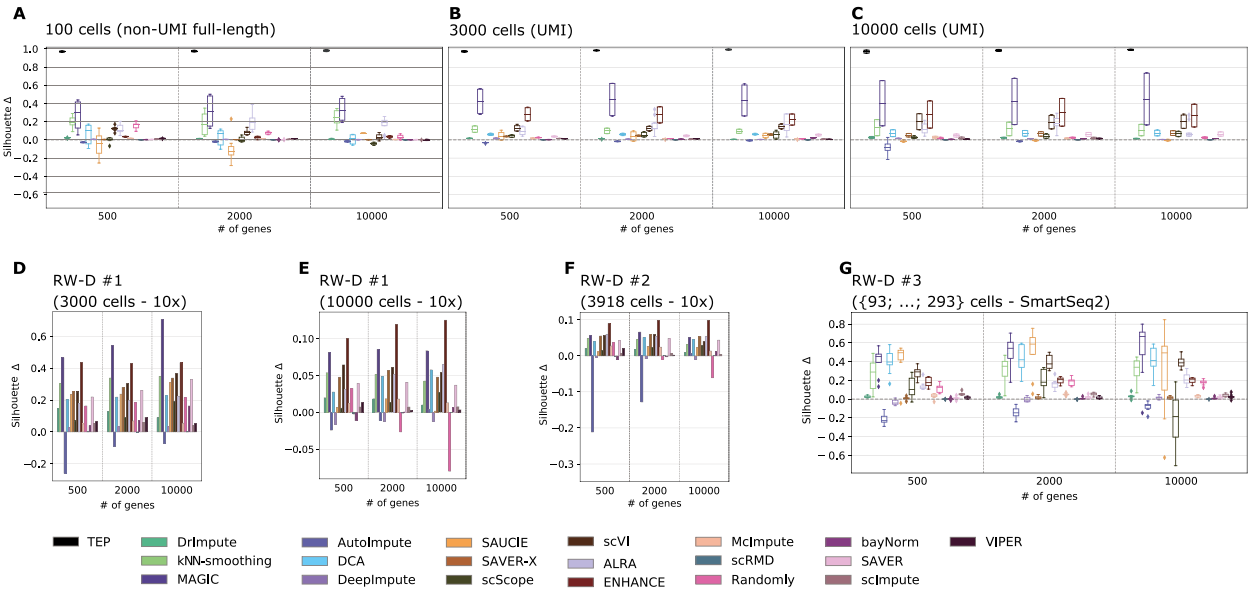


Figure 5. Performance assessment on cell similarity characterization (simulations and real-world data). Assessment of enhancement of cell similarity characterization after denoising, as evaluated on (i) simulated datasets (non-UMI full-length with 100 single cells and UMI-simulated datasets with {3000, 10000} single cells, panels A–C) and (ii) real-world datasets RW-D#1 (downsampled to {3000, 10000} cells, 10x platform, panels D and E), RW-D#2 (3918 cells, 10x platform, panel F) and RW-D#3 ({93, 196, 197, 224, 243, 245, 263, 275, 293} cells, Smart-Seq2, panel G). The boxplots (respectively, barplots) in all panels, depict the distribution (respectively, values) of the Silhouette delta, i.e. the difference between the average silhouette coefficient computed on the denoised expression matrix and that computed on the NEM, for all methods. The difference between the average silhouette coefficient evaluated on the TEP and that computed on the NEM is also shown for all simulated datasets.

Overall, most methods cause an increase of the average silhouette coefficient in most settings, suggesting that imputation and denoising approaches are indeed effective in enhancing the similarity of the expression profiles of cells belonging to the same sub-populations.

This effect is significantly intensified with datasets with larger sample size and generated (or simulated) with UMI protocols, as proven by the overall increase in delta magnitude. In particular, MAGIC and ENHANCE appear to produce the best results, with respect to both simulated and real-world datasets, yet with noteworthy variance in some scenarios, and with the latter method improving its performance with UMI datasets. We further notice that ALRA, kNN-smoothing and scVI deliver notable performances in most scenarios, closely followed by DCA. Surprisingly, SAUCIE exhibits a negative delta with simulated non-UMI full-length datasets but produces good results with real-world Smart-Seq2 dataset RW-D#3.

We recall that, among the best performing methods for the imputation and expression recovery tasks (see above), in addition to the aforementioned MAGIC and ENHANCE, SAVER-X and SAVER consistently produce improvements of the average silhouette delta in most simulated and real-world scenarios, whereas bayNorm and DrImpute appear to be less effective with respect to this specific task.

We finally specify that the results on simulated and real-world datasets are mostly coherent across experimental scenarios, further proving the suitability of simulations in assessing the performance of imputation and denoising methods.

Identification of DEGs (real-world data)

In order to quantify the effect of denoising and imputation methods on the identification of DEGs, we leveraged on bulk RNA-sequencing data included in real-world dataset RW-D#4 [93]. In detail, we first computed the DEGs between the parental and resistant samples included in the dataset, with respect to both the original expression matrix and the denoised matrix (via Wilcoxon test, $P < 0.05$), and which resulted in two distinct lists of DEGs. The analysis was repeated for both the Fluidigm/Smart-Seq dataset (84 and 113 single cells for resistant and parental cell lines, respectively) and the 10x datasets (3085 and 3178; see the Methods section and the Supplementary Material section 4 for further details).

In Figure 6, we display the difference of the Spearman correlation coefficient between the expression profile of the DEGs obtained from the denoised expression matrix and the bulk expression profile (computed for each single cell), and the one computed on the profiles of DEGs determined from the original expression matrix.

Noteworthy, most approaches produce an increase of the correlation with respect to the bulk expression profile. In particular, kNN-smoothing, MAGIC and SAUCIE deliver a median Spearman delta > 0.10 for both the Fluidigm/Smart-Seq and the 10x datasets, while bayNorm, ENHANCE, SAVER, SAVER-X and scVI show a median Spearman delta > 0.10 for the latter protocol only.

Overall, this result indicates that, in many cases, imputation and denoising methods might be effective in improving downstream analyses, such as the identification of DEGs.

Computation time (simulations)

Figure 7 reports the results of the computational time assessment on three simulated datasets: (i) non-UMI full-length (100 cells) (ii) UMI (3000 cells), and (iii) UMI (10 000 cells), with respect to {500, 2000, 5000} genes, plotted in logarithmic scale.

We can observe that all methods suffer an approximately exponential increase of computational time with respect to the number of cells and the number of genes, with extremely significant difference in magnitude. Overall, the most scalable

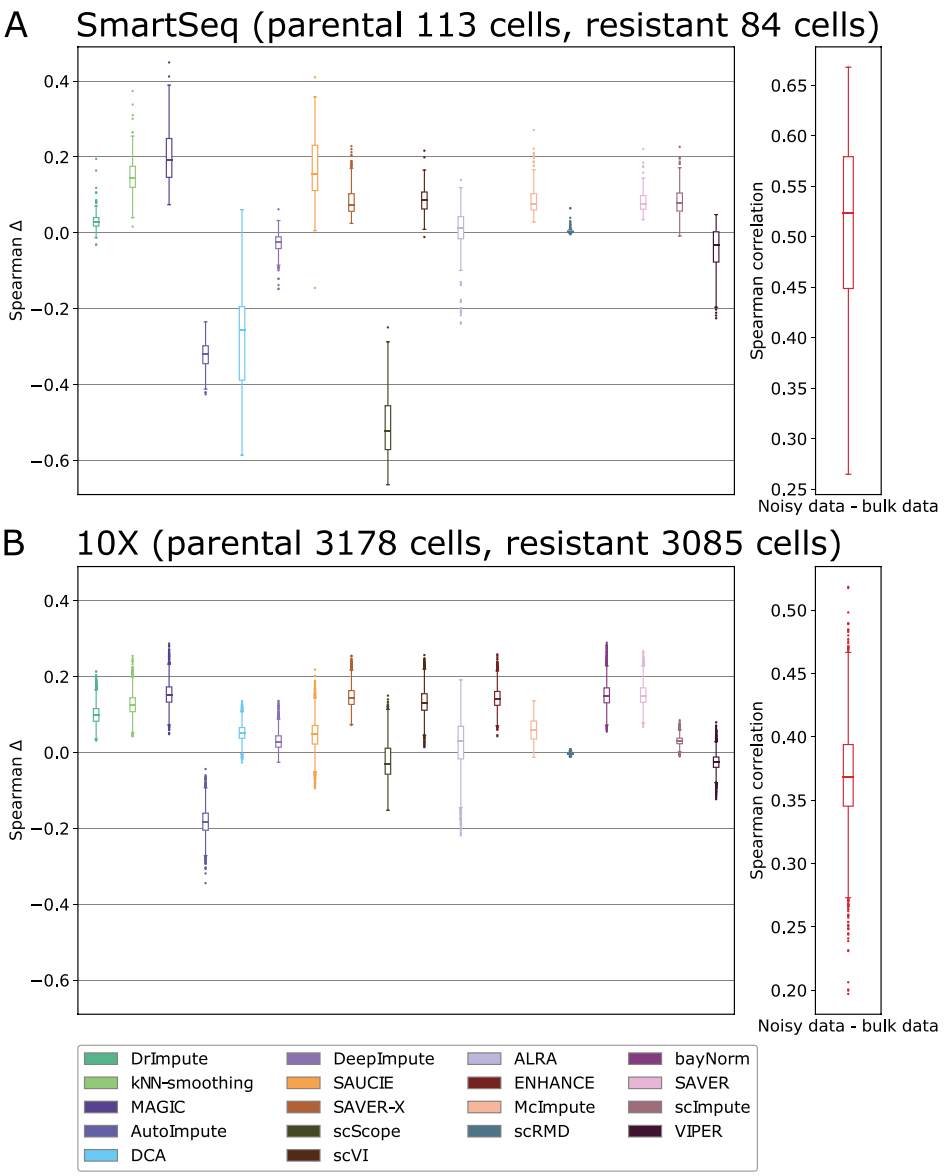


Figure 6. Performance assessment on identification of DEGs (real-world data). Assessment of identification of DEGs, as computed on **RW-D**#4 [93]. DEGs between parental and resistant cell lines of RW-D#4 are identified via Wilcoxon test ($P < 0.05$), both starting from the original scRNA-seq dataset and from the corresponding denoised matrices, for both Fluidigm/Smart-Seq and 10x datasets (panel A and B). The Spearman correlation coefficient between the expression profile of all single cells and the corresponding bulk expression profile is computed with respect to all the DEGs included in the distinct lists. The distribution (on all single cells) of the difference between the Spearman correlation coefficient computed with original data matrix and that computed with the denoised version is then shown as boxplots for both 10x and Fluidigm/Smart-Seq datasets. In the rightmost panels, the baseline distribution of the Spearman correlation coefficient between the NEM and bulk data (with respect to the corresponding list of DEGs) is shown, for both scenarios.

algorithms appear to be ALRA, kNN-smoothing and scRMD while, in general, matrix theory appears to be the most computationally efficient category.

Summary of the performance assessment on denoising and imputation methods

In Figure 8, we present a recapitulation of the performance assessment. The schema includes seven panels, structured as follows:

- imputation of dropout events,
- recovery of gene expression profiles,
- characterization of cell similarity,
- identification of DEGs,
- computation time,
- task,
- release code quality.

In particular, we selected a subset of simulated datasets, characterized by selected parameter settings in terms of single-cell number ($\{100, 3000, 10000\}$), sequencing protocol {non-UMI full-length, UMI} and number of genes (2000 for all settings)—and all four real-world datasets (see the Methods section), which we employed to compute a schematic ranking of all methods with respect to the distinct tasks.

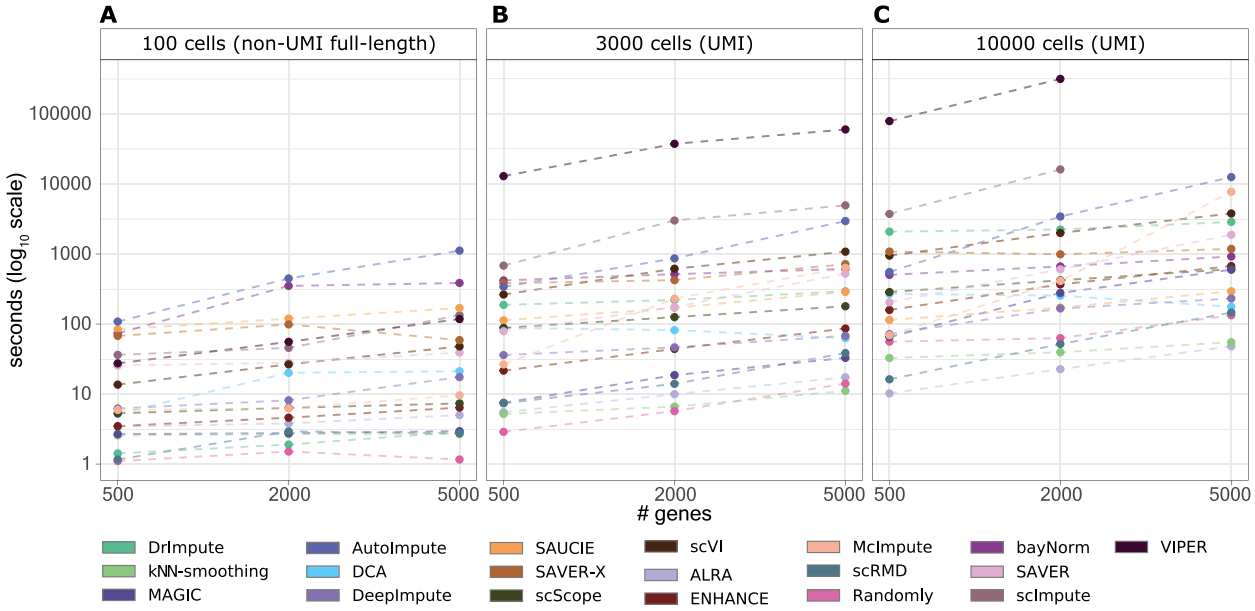


Figure 7. Computational time assessment. Running time of each method in denoising/imputing datasets with increasing number of cells and genes. In **(A)** results with 100 cells, in **(B)** results with 3000 cells and in **(C)** results with 10 000 cells. Values are plotted in logarithmic scale.

More in detail, for each selected parameter setting of the simulated dataset and for each real-world dataset, we ordered all 19 methods with respect to the average values of the following metrics:

(i) average Spearman correlation delta for zero entries of the NEM (for imputation of dropout events),
(ii) average Spearman correlation delta on the whole expression matrix (for recovery of gene expression profiles),
(iii) average silhouette delta (for characterization of cell similarity),
(iv) average Spearman correlation delta (for identification of DEGs),
(v) computation time.

The ranking is visually represented with dots with respect to each experimental setting, where the largest dot corresponds to the best performing method (green) and the smallest dot to the worst performing method (red).

The task panel indicates whether each method performs either denoising or imputation (see the Introduction section and Supplementary Material section 1 for a rigorous classification of the two tasks). Finally, the last panel reports a summary of selected quality code metrics, which were used to evaluate the different tools. In particular, usability and documentation range from 1, i.e. the worst result, to 4, corresponding to the best score. Usability is calculated by considering a set of characteristics that contribute in worsening the overall usability of the tool: (i) either input preprocessing, preliminary operation, e.g. clustering, or output post-processing, e.g. re-normalization, are required to the user; (ii) at least one parameter depends on the input, i.e. a grid-search is required; (iii) parameters meaning is not intuitive, e.g. it has no biological meaning; (iv) the tool is not available on a package distribution platform, e.g. `Bioconductor` or `pip`/`conda`. If a tool has none of the previously introduced features is assigned to the maximum score of 4; otherwise, the scoring is reduced to a minimum of 1. Documentation score is assigned as follows: 1 indicates that the authors did provide neither a documentation nor a detailed tutorial, 2 indicates that the authors provided a tutorial but did not write a detailed explanation for the parameters, 3 indicates that a detailed tutorial is available and 4 indicates that the authors provided both a detailed tutorial and a full explanation of all parameters. Finally, we indicate both whether the program is maintained, i.e. updated in the past 2 years, and the programming language on which the tool was implemented.

Discussion

We presented a review of the current state-of-the art of computational approaches for denoising and imputation of scRNA-seq data. Extensive tests on both real and synthetic datasets allowed to evaluate the performances and the robustness of each method under different experimental scenarios.

In light of the presented results, distinct methods appear to be more suitable for different tasks. In particular, ENHANCE, MAGIC, SAVER, and SAVER-X provide the best overall compromise and show robust performances with respect to all considered tasks. In addition to such methods, bayNorm and DrImpute are especially effective in recovering the true expression profiles and imputing dropout entries, while kNN-smoothing and scVI in improving the characterization of cell similarity and the grouping of single cells in coherent subpopulations, as well as the identification of DEGs.

We also note that, as expected, most methods appear to struggle with non-UMI full-length datasets, likely due to the low number of observations (cells) as compared with the high number of variables (genes). Furthermore, as already mentioned and as reported in [110], denoised expression values returned by any method should be considered with caution, due to the presence of possible artifacts, as proven by the low correlation with GT expression profiles from simulations recorded in many cases and, particularly, with non-UMI full-length datasets.

By focusing on machine learning frameworks, we notice that methods that employ assumptions on biological variability and technical noise (i.e. DCA, SAVER-X, scVI) typically exhibit

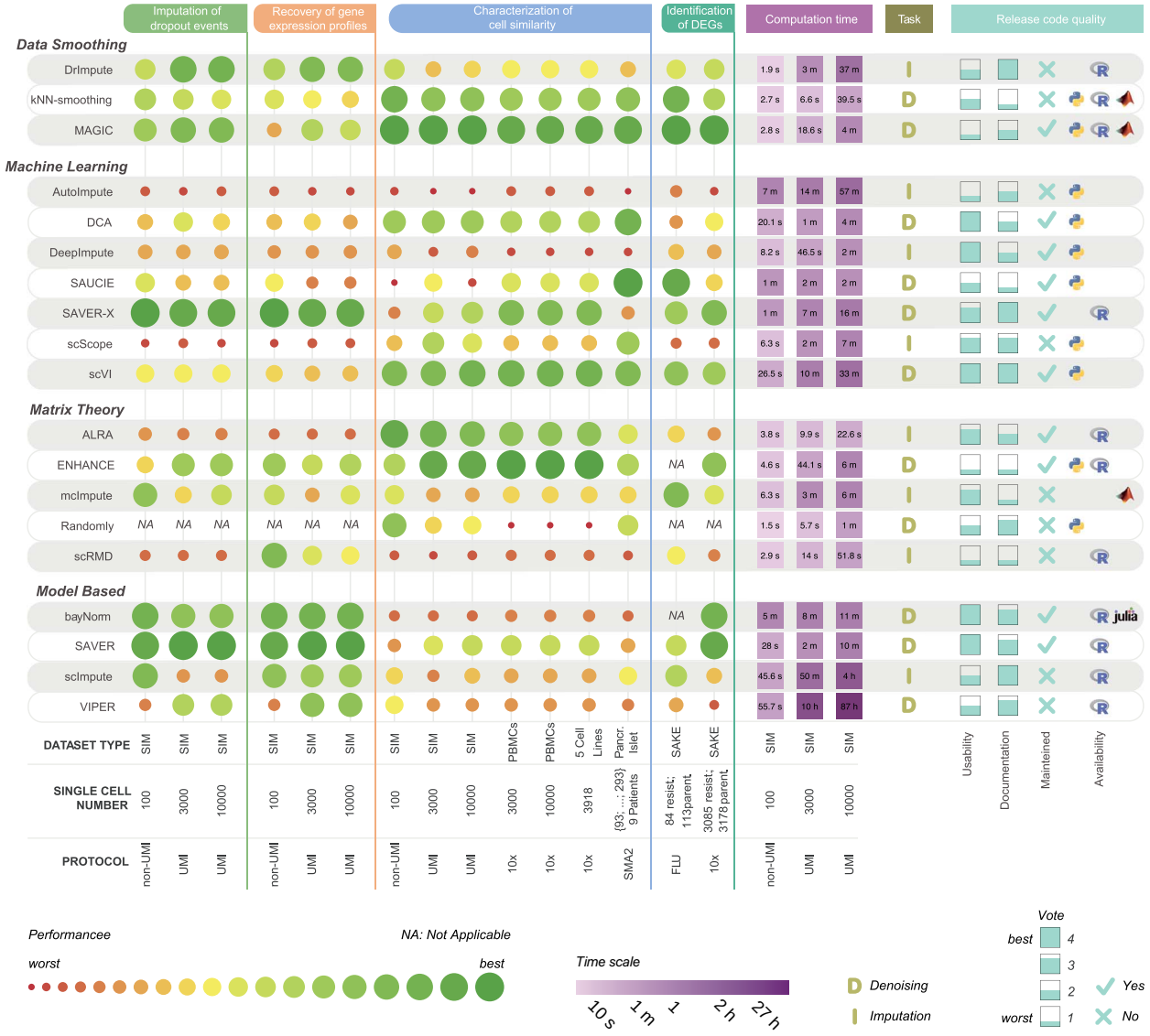


Figure 8. Summary of the performance assessment on denoising and imputation methods. The five leftmost panels report a schematic ranking of all 19 tested denoising and imputation methods, as computed on a selected panel of synthetic and real-world datasets, in terms of average Spearman correlation delta for zero entries of the NEM (for imputation of dropout events), average Spearman correlation delta on the whole expression matrix (for recovery of gene expression profiles), average silhouette delta (characterization of cell similarity), average Spearman correlation delta (identification of DEGs) and computation time. The size of each round marker is proportional to the ranking, with the largest (green) dots corresponding to the best performing tool and the smallest (red) dots to worst performing tool, with respect to the considering metric. The task panel indicates whether the method can perform either denoising or imputation tasks. Finally, the rightmost panel reports a summary of a quality code metrics that were used to evaluate the different tools in terms of usability, documentation, maintenance and availability (please refer to the Methods section (Summary of the performance assessment on denoising and imputation Methods) for further details).

better performances, hinting at the importance of including prior knowledge to inform the learning algorithms. Model-based methods present a typically good performance in both imputation and expression recovery, yet at a usually high computational cost, and generally showing suboptimal performance in cell similarity enhancement. Matrix theory-based techniques show good performance in terms of characterization of cell similarity, in addition to noteworthy scalability, even with large datasets. Finally, data smoothing approaches present typically good performances, yet with significant differences according to the specific task.

All in all, the performance of all methods appear to be highly dependent on the specific features of the dataset, as very distinct results are observed for the same method in different experimental scenarios, as recapped in Figure 8. This summary should guide potential users in selecting an optimal method according to the research needs and the available data types.

We further note that a review on a similar subject can be found as a preprint in [35]. Despite achieving similar conclusions on several methods included in our review, such work does not include comparisons on simulated data, which allow to evaluate a number of metrics with respect to the GT. For instance, certain methods that were identified as highly performing in [35], appear to struggle in dealing with true expression profiles recovery, an effect that can be evaluated only via simulations. The virtually unlimited number of in silico scenarios that can

be generated via methods such SymSim [89] suggests that simulations should be increasingly used to quantitatively assess the performance of data science methods and especially to test the robustness of their results.

Possible limitations of our assessment might be related to the application of most methods with default parameters, while one can expect improvements when fine tuning the parameters. In this respect, setting guidelines provided by the authors were followed when present and appear to be extremely beneficial to increase the overall usability and performance of the methods.

We also recall that for some methods, such as those based on AEs, it would be possible to use the latent variable space to perform single-cell clustering, while in our analysis we chose to use the denoised expression profiles, to provide a fair comparison for all methods.

We finally remark that scalable methods for denoising of single-cell transcriptomic data might pave the way for refined downstream analyses, for instance, by improving the reliability and accuracy of variant calling pipelines from scRNA-seq data to provide an accurate mapping of genotype and phenotype of single cells [111, 112], as well as by allowing a better estimation of metabolic fluxes from scRNA-seq data in the investigation of cancer metabolism [113, 114].

Key Points

- Extensive tests on synthetic and real datasets provide a quantitative assessment of the performance of denoising and imputation methods in distinct scenarios.
- Some methods are effective in improving the characterization of cell similarity, some others in recovering the true gene expression profiles and imputing dropouts.
- Appropriate assumptions on the noise model are beneficial to recover lost information.
- Overall, ENHANCE, MAGIC, SAVER and SAVER-X constitute a good compromise on all tasks.
- Corrected expression values returned by any method should be considered with caution in downstream analyses.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Data availability

The source code used to replicate all our analyses, including synthetic and real datasets, is available at this link: https://github.com/BIMIB-DISCo/review-scRNA-seq-DENOISING.

Acknowledgments

We thank Chiara Damiani, Daniele Ramazzotti and Giulio Caravagna for helpful discussions.

Funding

This work was supported by the Cancer Research UK and Associazione Italiana per la Ricerca sul Cancro (CRUK/AIRC) "Accelerator Award" (award #22790) 'Single-cell Cancer Evolution in the Clinic'. Partial support was also provided by the Italian node of the Elixir network (https://elixir-europe.org/about-us/who-we-are/nodes/italy) and the SysBioNet project, a Ministero dell'Istruzione, dell'Universitá e della Ricerca initiative for the Italian Roadmap of European Strategy Forum on Research Infrastructures.

References

1. Dalerba P, Kalisky T, Sahoo D, *et al*. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat Biotechnol* 2011;**29**(12):1120.
2. Vieth B, Parekh S, Ziegenhain C, *et al*. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat Commun* 2019;**10**(1):1–11.
3. Angela RW, Norma F, Neff TK, *et al*. Quantitative assessment of single-cell RNA-sequencing methods. *Nat Methods* 2014;**11**(1):41.
4. Kalisky T, Blainey P, Quake SR. Genomic analysis at the single-cell level. *Annu Rev Genet* 2011;**45**:431–45.
5. Huang S. Non-genetic heterogeneity of cells in development: more than just noise. *Development* 2009;**136**(23):3853–62.
6. Li L, Clevers H. Coexistence of quiescent and active adult stem cells in mammals. *Science* 2010;**327**(5965):542–5.
7. Shalek AK, Satija R, Shuga J, *et al*. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 2014;**510**(7505):363–9.
8. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* 2018;**50**(8):1–14.
9. AlJanahi AA, Danielsen M, Dunbar CE. An introduction to the analysis of single-cell RNA-sequencing data. *Mol Ther Methods Clin Dev* 2018;**10**:189–96.
10. Lawson DA, Kessenbrock K, Davis RT, *et al*. Tumour heterogeneity and metastasis at single-cell resolution. *Nat Cell Biol* 2018;**20**(12):1349–60.
11. Shaffer SM, Dunagin MC, Torborg SR, *et al*. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* 2017;**546**(7658):431.
12. Cao J, Packer JS, Ramani V, *et al*. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 2017;**357**(6352):661–7.
13. Regev A, Teichmann SA, Lander ES, *et al*. Science forum: the human cell atlas. *elife* 2017;**6**:e27041.
14. Elowitz MB, Levine AJ, Siggia ED, *et al*. Stochastic gene expression in a single cell. *Science* 2002;**297**(5584):1183–6.
15. Marinov GK, Williams BA, McCue K, *et al*. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res* 2014;**24**(3):496–510.
16. Haque A, Engel J, Teichmann SA, *et al*. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med* 2017;**9**(1):75.
17. Ziegenhain C, Vieth B, Parekh S, *et al*. Comparative analysis of single-cell RNA sequencing methods. *Mol Cell* 2017;**65**(4):631–43.
18. Macosko EZ, Basu A, Satija R, *et al*. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 2015;**161**(5):1202–14.
19. Klein AM, Mazutis L, Akartuna I, *et al*. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 2015;**161**(5):1187–201.

20. Fraction of mRNA transcripts captured per cell. https://kb.10xgenomics.com/hc/en-us/articles/360001539051-What-fraction-of-mRNA-transcripts-are-captured-per-cell.
21. Ramsköld D, Luo S, Wang Y-C, *et al*. Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* 2012;**30**(8):777.
22. Sheng K, Cao W, Niu Y, *et al*. Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat Methods* 2017;**14**(3):267–70.
23. Jaitin DA, Kenigsberg E, Keren-Shaul H, *et al*. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 2014;**343**(6172):776–9.
24. Hashimshony T, Wagner F, Sher N, *et al*. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* 2012;**2**(3):666–73.
25. Rosenberg AB, Roco CM, Muscat RA, *et al*. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* 2018;**360**(6385):176–82.
26. Pollen AA, Nowakowski TJ, Shuga J, *et al*. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat Biotechnol* 2014;**32**(10):1053.
27. Gierahn TM, Wadsworth MH, II, Hughes TK, *et al*. Seq-well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat Methods* 2017;**14**(4):395–8.
28. Islam S, Zeisel A, Joost S, *et al*. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* 2014;**11**(2):163.
29. Ziegenhain C, Vieth B, Parekh S, *et al*. Comparative analysis of single-cell RNA sequencing methods. *Mol Cell* 2017;**65**(4):631–43.
30. Haque A, Engel J, Teichmann SA, *et al*. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med* 2017;**9**(1):75.
31. Goh WWB, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol* 2017;**35**(6):498–507.
32. Tung P-Y, Blischak JD, Hsiao CJ, *et al*. Batch effects and the effective design of single-cell gene expression studies. *Sci Rep* 2017;**7**:39921.
33. Tran HTN, Ang KS, Chevrier M, *et al*. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol* 2020;**21**(1):12.
34. Zhu L, Lei J, Devlin B, *et al*. A unified statistical framework for single cell and bulk RNA sequencing data. *Ann Appl Stat* 2018;**12**(1):609.
35. Hou W, Ji Z, Ji H, *et al*. A systematic evaluation of single-cell RNA-sequencing imputation methods. *bioRxiv* 2020. doi: 10.1101/2020.01.29.925974.
36. Agarwal D, Wang J, Zhang NR, *et al*. Data denoising and post-denoising corrections in single cell RNA sequencing. *Stat Sci* 2020;**35**(1):112–28.
37. Lähnemann D, Köster J, Szczurek E, *et al*. Eleven grand challenges in single-cell data science. *Genome Biol* 2020;**21**(1):1–35.
38. Gong W, Kwak I-Y, Pota P, *et al*. DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC Bioinformatics* 2018;**19**(1):220.
39. Tjaernberg A, Mahmood O, Jackson CA, *et al*. Optimal tuning of weighted kNN- and diffusion-based methods for denoising single cell genomics data. *bioRxiv* 2020. doi: 1101/2020.02.28.970202.
40. Ye P, Ye W, Ye C, *et al*. scHinter: imputing dropout events for single-cell RNA-seq data with limited sample size. *Bioinformatics* 2020;**36**(3):789–97.
41. Wagner F, Yan Y, Yanai I. K-nearest neighbor smoothing for high-throughput single-cell RNA-seq data. *bioRxiv* 2018. doi: 10.1101/217737.
42. Moussa M, Măndoiu II. Locality sensitive imputation for single cell RNA-seq data. *J Comput Biol* 2019;**26**(8):822–35.
43. Van Dijk D, Sharma R, Nainys J, *et al*. Recovering gene interactions from single-cell data using data diffusion. *Cell* 2018;**174**(3):716–29.
44. Ronen J, Akalin A. netSmooth: network-smoothing based imputation for single cell RNA-seq. *F1000Res* 2018; 7:8.
45. Jeong H, Liu Z. PRIME: a probabilistic imputation method to reduce dropout effects in single cell RNA sequencing. *Bioinformatics* 2020;**36**(13):4021–9.
46. Tracy S, Yuan G-C, Dries R. RESCUE: imputing dropout events in single-cell RNA-sequencing data. *BMC Bioinformatics* 2019;**20**(1):388.
47. Wu W, Dai Q, Liu Y, *et al*. G2S3: a gene graph-based imputation method for single-cell RNA sequencing data. *bioRxiv* 2020. doi: 10.1101/2020.04.01.020586.
48. Jin K, Ou-Yang L, Zhao X-M, *et al*. scTSSR: gene expression recovery for single-cell RNA sequencing using two-side sparse self-representation. *Bioinformatics* 2020;**36**(10):3131–8.
49. Leote AC, Wu X, Beyer A. Network-based imputation of dropouts in single-cell RNA sequencing data. *bioRxiv* 2019. doi: 10.1101/611517.
50. Elyanow R, Dumitrascu B, Engelhardt BE, *et al*. netNMF-sc: leveraging gene–gene interactions for imputation and dimensionality reduction in single-cell expression analysis. *Genome Res* 2020;**30**(2):195–204.
51. Wang J, Agarwal D, Huang M, *et al*. Data denoising with transfer learning in single-cell transcriptomics. *Nat Methods* 2019;**16**(9):875–8.
52. Peng T, Zhu Q, Yin P, *et al*. SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data. *Genome Biol* 2019;**20**(1):88.
53. Ye W, Ji G, Ye P, *et al*. scNPF: an integrative framework assisted by network propagation and network fusion for preprocessing of single-cell RNA-seq data. *BMC Genomics* 2019;**20**(1):347.
54. Badsha MB, Li R, Liu B, *et al*. Imputation of single-cell gene expression with an autoencoder neural network. *Quant Biol* 2020;1–17.
55. Zhu L, Lei J, Devlin B, *et al*. A unified statistical framework for single cell and bulk RNA sequencing data. *Ann Appl Stat* 2018;**12**(1):609.
56. Talwar D, Mongia A, Sengupta D, *et al*. AutoImpute: autoencoder based imputation of single-cell RNA-seq data. *Sci Rep* 2018;**8**(1):1–11.
57. Arisdakessian C, Poirion O, Yunits B, *et al*. DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. *Genome Biol* 2019;**20**(1):1–14.
58. Eraslan G, Simon LM, Mircea M, *et al*. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun* 2019;**10**(1):390.
59. Zhang X-F, Ou-Yang L, Yang S, *et al*. EnImpute: imputing dropout events in single-cell RNA-sequencing data via ensemble learning. *Bioinformatics* 2019;**35**(22):4827–9.

60. Rao J, Zhou X, Lu Y, *et al.* Imputing single-cell RNA-seq data by combining graph convolution and autoencoder neural networks. *bioRxiv* 2020. doi: 10.1101/2020.02.05.935296.
61. Xu Y, Zhang Z, You L, *et al.* scIGANs: single-cell RNA-seq imputation using generative adversarial networks. *bioRxiv* 2020. doi: 10.1101/2020.01.20.913384.
62. Amodio M, Van Dijk D, Srinivasan K, *et al.* Exploring single-cell data with deep multitasking neural networks. *Nat Methods* 2019;**62**:1139–45.
63. Deng Y, Bao F, Dai Q, *et al.* Scalable analysis of cell-type composition from single-cell transcriptomics using deep recurrent learning. *Nat Methods* 2019;**16**(4):311–4.
64. Lopez R, Regier J, Cole MB, *et al.* Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;**15**(12):1053–8.
65. Mehtonen J, González G, Kramer R, *et al.* Semisupervised generative autoencoder for single-cell data. *J Comput Biol* 2019;**27**(8):1190–203.
66. Zhu K, Anastassiou D. 2DImpute: imputation in single-cell RNA-seq data from correlations in two dimensions. *Bioinformatics* 2020;**36**(11):3588–9.
67. Tran B, Tran D, Nguyen H, *et al.* Ria: a novel regression-based imputation approach for single-cell RNA sequencing. In: *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, 2019. pp. 1–9. New York City, NY, USA: IEEE.
68. Linderman GC, Zhao J, Kluger Y. Zero-preserving imputation of scRNA-seq data using low-rank approximation. *bioRxiv* 2018. doi: 10.1101/397588.
69. Wagner F, Barkley D, Yanai I. Accurate denoising of single-cell RNA-seq data using unbiased principal component analysis. *bioRxiv* 2019;**655365**. doi: 10.1101/655365.
70. Chen C, Wu C, Wu L, *et al.* scRMD: imputation for single cell RNA-seq data via robust matrix decomposition. *Bioinformatics* 2020;**36**(10):3156–61. 03.
71. Xu J, Cai L, Liao B, *et al.* CMF-Impute: an accurate imputation tool for single-cell RNA-seq data. *Bioinformatics* 2020;**36**(10):3139–47.
72. Mongia A, Sengupta D, Majumdar A. deepMc: deep matrix completion for imputation of single-cell RNA-seq data. *J Comput Biol* 2019;**27**(7):1011–9.
73. Mongia A, Sengupta D, Majumdar A. McImpute: matrix completion based imputation for single cell RNA-seq data. *Front Genet* 2019;**10**:9.
74. Zhang L, Zhang S. PBLR: an accurate single cell RNA-seq data imputation tool considering cell heterogeneity and prior expression level of dropouts. *bioRxiv* 2018. doi: 10.1101/379883.
75. Hu Y, Li B, Liu N, *et al.* WEDGE: recovery of gene expression values for sparse single-cell RNA-seq datasets using matrix decomposition. *bioRxiv* 2019;864488. https://doi.org/10.1101/864488.
76. Pierson E, Yau C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol* 2015;**16**(1):241.
77. Aparicio L, Bordyuh M, Blumberg AJ, *et al.* A random matrix theory approach to denoise single-cell data. *Patterns* 2020;;**1**(3):100035.
78. Tang W, Bertaux F, Thomas P, *et al.* bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data. *Bioinformatics* 2020;**36**(4):1174–81.
79. Azizi E, Prabhakaran S, Carr A, *et al.* Bayesian inference for single-cell clustering and imputing. *Genomics Comput Biol* 2017;**3**(1):e46–6.
80. Song F, Chan GM, Wei Y. Flexible experimental designs for valid single-cell RNA-sequencing experiments allowing batch effects correction. *Nat Commun* 2020;**11**:3274.
81. Lin P, Troup M, Ho JWK. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol* 2017;**18**(1):59.
82. Yang MQ, Weissman SM, Yang W, *et al.* MISC: missing imputation for single-cell RNA sequencing data. *BMC Syst Biol* 2018;**12**(7):114.
83. Huang M, Wang J, Torre E, *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods* 2018;**15**(7):539.
84. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat Commun* 2018;**9**(1):997.
85. Miao Z, Li J, Zhang X. scRecover: discriminating true and false zeros in single-cell RNA-seq data for imputation. *bioRxiv* 2019;665323. https://doi.org/10.1101/665323.
86. Zhang Y, Liang K, Liu M, *et al.* SCRIBE: a new approach to dropout imputation and batch effects correction for single-cell RNA-seq data. *bioRxiv* 2019;793463. https://doi.org/10.1101/793463.
87. Hu Z, Songpeng Z, Liu JS. SIMPLEs: a single-cell RNA sequencing imputation strategy preserving gene modules and cell clusters variation. *bioRxiv* 2020. doi: 10.1101/2020.01.13.904649.
88. Chen M, Zhou X. VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. *Genome Biol* 2018;**19**(1):1–15.
89. Zhang X, Xu C, Yosef N. Simulating multiple faceted variability in single cell RNA sequencing. *Nat Commun* 2019;**10**(1):2611.
90. Zheng GXY, Terry JM, Belgrader P, *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017;**8**(1):1–12.
91. Tian L, Dong X, Freytag S, *et al.* Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat Methods* 2019;**16**(6):479–87.
92. Segerstolpe Å, Palasantza A, Eliasson P, *et al.* Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab* 2016;**24**(4):593–607.
93. Ho Y-J, Anaparthy N, Molik D, *et al.* Single-cell RNA-seq analysis identifies markers of resistance to targeted BRAF inhibitors in melanoma cell populations. *Genome Res* 2018;**28**(9):1353–63.
94. Andrews TS, Hemberg M. False signals induced by single-cell imputation. *F1000Res* 2018;**7**:1740.
95. Zhang L, Zhang S. Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Trans Comput Biol Bioinform* 2018;**17**(2):376–89.
96. Coifman RR, Lafon S. Diffusion maps. *Appl Comput Harmon Anal* 2006;**21**(1):5–30.
97. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, MA, USA: The MIT Press, 2016.
98. Wang Y-X, Zhang Y-J. Nonnegative matrix factorization: a comprehensive review. *IEEE Trans Knowl Data Eng* 2012;**25**(6):1336–53.
99. Candès EJ, Recht B. Exact matrix completion via convex optimization. *Found Comput Math* 2009;**9**(6):717.
100. Eckart C, Young GM. The approximation of one matrix by another of lower rank. *Psychometrika* 1936;**1**:211–8.
101. Sun Y, Babu P, Palomar DP. Majorization-minimization algorithms in signal processing, communications, and

machine learning. *IEEE Trans Signal Process* 2016;**65**(3):816–794.
102. Livan G, Novaes M, Vivo P. *Introduction to Random Matrices: Theory and Practice*, Vol. 26. London, UK: Springer, 2018.
103. Hsu D. Kakade SM, Zhang T. Robust matrix decomposition with sparse corruptions. *IEEE Trans Inf Theory* 2011;**57**(11):7221–34.
104. Ng AY, Jordan MI, Weiss Y. On spectral clustering: analysis and an algorithm. In: *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, 2002. pp. 849–56. Vancouver, BC, Canada: MIT press.
105. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987;**20**:53–65.
106. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.
107. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018; **19**(1):15.
108. Butler A, Hoffman P, Smibert P, *et al*. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 2018;**36**(5):411.
109. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 2019;**9**(1):1–12.
110. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* 2019;**15**(6):e8746.
111. Ramazzotti D, Angaroni F, Maspero D, *et al*. Longitudinal cancer evolution from single cells. *bioRxiv* 2020. doi: 10.1101/2020.01.14.906453.
112. Zhou Z, Xu B, Minn A, *et al*. DENDRO: genetic heterogeneity profiling and subclone detection by single-cell RNA sequencing. *Genome Biol* 2020;**21**(1):1–15.
113. Damiani C, Maspero D, Di Filippo M, *et al*. Integration of single-cell RNA-seq data into population models to characterize cancer metabolism. *PLoS Comput Biol* 2019;**15**(2):e1006733.
114. Graudenzi A, Maspero D, Damiani C. FBCA, a multiscale modeling framework combining cellular automata and flux balance analysis. *J Cell Autom* 2020;**15**:75–95.

2.3 Data type II: generation of mutational profiles

From BAM files to VCF files

In general, any experiment producing nucleotide sequences can be used to retrieve genetic mutations.

DNA-sequencing methods, such as whole-genome (WGS), whole-exome (WES) or targeted sequencing [67, 202, 211], allow the reconstruction of either the whole DNA sequence, the protein-coding regions of the genome, or specific genomic regions, respectively, thus permitting to naturally detect the mutations with respect to a reference genome. However, even the sequencing technologies designed for different purposes, such as RNA-seq, ATAC-seq or ChIP-seq, generate reads which can be exploited to this end [56, 242].

In fact, BAM files generated from the alignment of sequences can be used to assess the presence of genomic alterations, e.g., single-nucleotide (SNV), structural variants (SV), etc. Notice that SVs involve large genome regions (i.e., > 50 base pairs *bp*), often include complex genetic rearrangements, and are more difficult to identify with respect to the single-nucleotide variants. For instance, in contrast to SNVs, SVs can cover a relevant portion of a sequenced read or even be larger than the read length, thus complicating mapping and requiring specific variant calling tools [187]. We highlight that, in our analyses, we focused on single-nucleotide variants (SNVs), discarding other kinds of mutations such as insertion-deletion (indels), focal copy number variations (CNVs), or other genomic aberrations [121]. For this reason, *SNV*, *variant*, and *mutation* are used in our works as synonyms.

After obtaining BAM files, SNVs can be called comparing the aligned reads string with the corresponding substring on a proper reference genome. Reference genomes are available for a broad range of species. In section 2.1.3, we introduced the SARS-CoV-2 reference genome, while human samples are usually aligned using the genome *GRCh38* produced and curated by the *Genome Reference Consortium* (ncbi.nlm.nih.gov/grc/human). Variant caller tools [185] produce a Variant Calling Format (VCF) file [45] from each BAM file. The VCF file includes a list of genetic variations and rich annotations.

We can distinguish two classes of single nucleotide variants: (*i*) germline mutations and (*ii*) somatic mutations. The former are inherited from the parents and are present in all the cells of an individual. The latter emerge after conception, due to random errors or mutational processes. Notice that, while germline mutations are called Single-Nucleotide Polymorphisms (SNPs) and are distributed in a human population with a known frequency defined as Minor Allele Frequency (MAF), their somatic mutation counterparts are called Single-Nucleotide Variants (SNVs).

Most somatic mutations are neutral, i.e., they do not impair the cellular functions,

but sometimes they can induce a transformation in a cell, leading to cancer or other diseases (i.e., driver events). Thus, it is necessary to distinguish somatic from germline mutations in cancer research. Typically, to achieve this goal, germline mutations are detected and filtered out by genotyping a healthy tissue from the same patient or exploiting public databases, such as dbSNP [20].

In the next paragraphs, we briefly report the approaches designed to retrieve mutations from bulk, single-cell and viral sequencing experiments. Although NGS technologies are similar for such experiments (see section 2.1), the noise model, the assumptions, and the resulting mutational profiles are different.

Bulk sequencing experiments. The Broad Institute collected bioinformatic tools in a Genome Analysis Toolkit (GATK) to perform Variant Discovery in High-Throughput Sequencing Data, also maintaining the pipeline to call point somatic mutation from bulk FASTQ files, obtained from bulk RNA sequencing experiments (gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels-). Note that, in paper P#2, we proved that the same workflow can be successfully applied to single-cell data as well.

VCF files resulting from the pipeline (one for each samples) can be combined to produce a continuous-value matrix, which includes samples on the rows and mutations on the columns. Each patient's tissue, biopsy or sample sequenced with bulk approach returns one mutation profile.

Single-cell sequencing experiments Reliable mutation profiles can be obtained only from BAM files generated with a full-length protocol, but, in general, it is still not possible to use BAM files from UMI-based ones (despite a notable exception proposed in [191]). The principal reason is that the latter produces a lower number of reads that cover only a tiny portion of one transcript end. Instead, full-length sequencing produce a more uniform coverage on the transcriptome. Different variant calling methods exist and have been applied to retrieve mutation profile from single-cell RNA-seq data [195].

With single-cell experiments, mutations are unequivocally assigned to any single cell. The number of reads with the mutation should correlate with the number of DNA copies carrying the mutations, which are two in a diploid organism. So, it is also expected that a somatic mutation accumulated during the tumour progression should be observed only in about half of the reads, because it hits only one of the DNA molecules. Unfortunately, different noise sources can affect this value. Primarily, if RNA-seq data are used, genes are differentially expressed, and sometimes, only one chromosome is actively transcribed (i.e., allele-specific expression [24]). Moreover, genetic aberrations (e.g., loss of heterozygosity) or sequencing errors (e.g., allelic dropout) may affect the fraction of reads displaying the mutations (also with DNA sequencing experiment). Please, see figure 2.2

for an example of alignment of scRNA-seq reads.

Finally, mutations observed in a small number of reads (e.g., less than %5 of the total coverage) are usually discarded, since they may be related to sequencing artefacts. In our workflow, mutations selected after quality control and filtering are binarized, and considered as either present or absent in a given single cell. Thus, the single-cell mutational profile is represented as a binary matrix with cells as rows and mutation as columns.

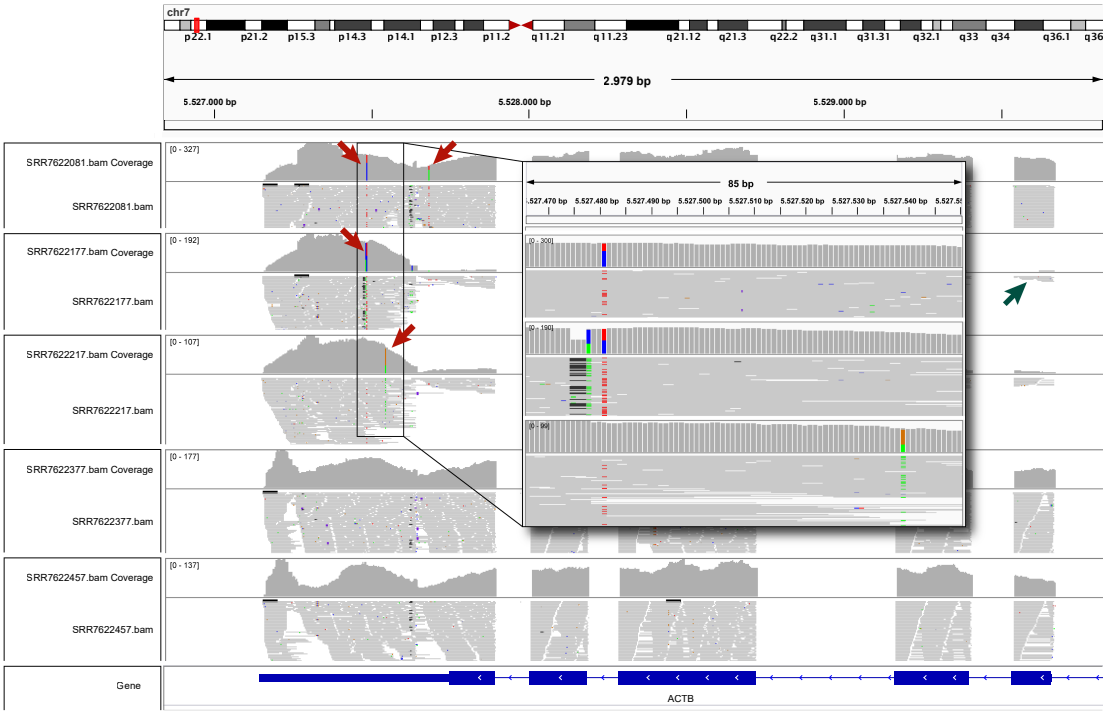


Figure 2.2: The figure reports the scRNA-seq reads alignment for five single cells collected by the authors of [165] and downloaded from Sequence Read Archive (SRA) with the corresponding identifier. Red arrows highlight the presence of cell-specific mutations. In the magnified selection, it is possible to observe their heterozygosity, suggesting their somatic nature. Only a portion of the ACTB gene locus in chromosome 7 is shown. A green arrow indicates low coverage regions. Finally, it is possible to observe probably noisy mutations present only in a few reads.

During my work, I first applied the tools and pipelines described above to call somatic mutations from longitudinal cancer single-cell RNA-seq experiments. We evaluated the feasibility of calling SNVs from such data in paper P#2, where we were able to asses the genetic identify of cells. Then, we used the same approach to obtain the input data for our new evolution inference framework presented in paper P#5.

2.3.1 Pipeline for variant calling from scRNA-seq

The genotyping analysis of full-length scRNA-seq data is rarely applied, with some exceptions [225], because of the high level of uncertainty usually present in the data. Systematic reviews, such as [185], which test the performance of different variant caller methods dealing with this particular kind of data, are necessary to prove the reliability of the mutational profiles retrieved. However, assessing the effectiveness of detecting the cellular identity in real-world scenario is a non-trivial task.

In this regard, involuntary help comes from Sharma and colleagues, authors of a study about the divergent modes of chemoresistance in patient-derived oral squamous cell carcinomas cell lines [165]. In brief, they produced multiple full-length scRNA-seq datasets by sampling single cells in different time points: before therapy administration (*Pri*), after the emergence of chemoresistance mechanisms (*PCR*), and after leaving the resistant cell in a drug-free environment (*PCRDH*). They performed the analyses for cells obtained from two oncological patients (HN120 and HN137).

We planned to use such RNA-seq dataset to perform variant calling and use the resulting mutational profile to infer the tumour evolution via our computational framework (presented in paper P#5). Unfortunately, during the analyses, we noted that the mutational profiles of the two replicates displayed inconsistent behaviours. In particular, HN120P cells were more similar to HN137PCR compared to HN120PCR ones, and vice-versa. Alarmed by this unexpected observation, we performed further analyses. We concluded that it is highly probable that the authors switched the identifier of the cell cultures before sequencing. This mistake may have led to erroneous conclusions, so we produced a *Matter Arising* paper to report our analyses and findings (currently under review) (paper P#2).

All in all, we proved the efficacy of genotyping scRNA-seq to assess both the genetic and the phenotypic heterogeneity of single cells.

Notice that we applied the same pipeline to obtain the somatic mutation profiles from a longitudinal scRNA-seq dataset of patient-derived xenografts (PDXs) of $\text{BRAF}^{\text{V600E/K}}$ mutant melanoma cells [163]. The resulting mutational profiles were used as input in our inference framework to investigate the tumour evolution and the impact of the therapy on the clonal structure. The study is presented in paper P#5. Details on the pipeline, the employed tools and the parameter settings are reported in the corresponding GitHub repository (see Appendix: Code repositories).

VARIANT CALLING FROM SCRNA-SEQ DATA ALLOWS THE ASSESSMENT OF CELLULAR IDENTITY IN PATIENT-DERIVED CELL LINES

Daniele Ramazzotti[1], **Fabrizio Angaroni**[2], **Davide Maspero**[2,3], **Gianluca Ascolani**[2], **Isabella Castiglioni**[4], **Rocco Piazza**[1], **Marco Antoniotti**[2,5], **Alex Graudenzi**[3,5,*]

[1] Dept. of Medicine and Surgery, Univ. of Milan-Bicocca, Monza, Italy
[2] Dept. of Informatics, Systems and Communication, Univ. of Milan-Bicocca, Milan, Italy
[3] Inst. of Molecular Bioimaging and Physiology,
Consiglio Nazionale delle Ricerche (IBFM-CNR), Segrate, Milan, Italy
[4] Department of Physics "Giuseppe Occhialini", Univ. of Milan-Bicocca, Milan, Italy
[5] Bicocca Bioinformatics, Biostatistics and Bioimaging Centre – B4, Milan, Italy
[*] Corresponding author: alex.graudenzi@ibfm.cnr.it

ABSTRACT

Matters Arising from: Sharma, A., Cao, E.Y., Kumar, V. et al. Longitudinal single-cell RNA sequencing of patient-derived primary cells reveals drug-induced infidelity in stem cell hierarchy. *Nat Commun* **9**, 4931 (2018). https://doi.org/10.1038/s41467-018-07261-3.

In Sharma, A. et al. *Nat Commun* **9**, 4931 (2018) the authors employ longitudinal single-cell transcriptomic data from patient-derived primary and metastatic oral squamous cell carcinomas cell lines, to investigate possible divergent modes of chemo-resistance in tumor cell subpopulations. We integrated the analyses presented in the manuscript, by performing variant calling from scRNA-seq data via GATK Best Practices. As a main result, we show that an extremely high number of single-nucleotide variants representative of the identity of a specific patient is unexpectedly found in the scRNA-seq data of the cell line derived from a second patient, and vice versa. This finding likely suggests the existence of a sample swap, thus jeopardizing some of the translational conclusions of the article. Our results prove the efficacy of a joint analysis of the genotypic and transcriptomic identity of single-cells.

The integration of omics data from single-cell sequencing experiments enables the analysis of cell-to-cell heterogeneity at unprecedented resolution and on multiple levels [3]. This is especially relevant in the study of cancer evolution and will be essential to shed light on the key mechanisms underlying intra-tumor heterogeneity, metastasis, drug resistance and relapse [4]. In particular, scRNA-seq experiments are increasingly employed, typically in the characterization of the gene expression patterns of single-cells in a variety of experimental settings [6]. However, an increasing number of studies is proving that scRNA-seq data can be used to efficiently call genomic variants, thus providing an available and cost-effective alternative to whole-genome/exome and targeted sequencing [5, 16]. Despite known pitfalls, such as the impossibility of calling variants from non-transcribed regions and the typically high rates of noise and dropouts [9], the mutational profiles so obtained can be promptly used to determine the identity of single-cells. This aspect is important, for instance, in the analysis of the clonal evolution of tumors and in the detection of rare clones [11].

In [13], the authors employ single-cell transcriptomic data from patient-derived primary and metastatic oral squamous cell carcinomas (OSCC) cell lines (from a previously characterized panel [1]), to investigate possible divergent modes of chemo-resistance in tumor cell subpopulations. We integrated the analyses presented in the manuscript, by performing variant calling from scRNA-seq data via GATK Best Practices [2]. On the one hand, this analysis may allow one to reconstruct the longitudinal evolution of the tumor in presence of the treatment [10]. On the other hand, this allows one to deliver an explicit mapping between genotype and phenotype of single cells, thus providing important hints on the relation between clonal evolution and phenotypic plasticity. This might have a significant translational relevance, given

the current shortage of accurate and affordable technologies for DNA and RNA sequencing of the same cells, despite the recent introduction of new protocols [7, 8].

In particular, we selected the scRNA-seq datasets of two cell lines derived from distinct OSCC patients – HN120 and HN137 – which include different data points, marked with the following suffixes: -P (primary line), -M (metastastic line), -CR (after cisplatin treatment), -CRDH (after drug-holiday). Since for the HN137P cell line two library layouts are provided (*single-end* and *paired-end*), which we here consider separately, and no HN137MCRDH is provided, we have a total of 12 datasets (all datasets are included in the GEO online repository, accession code `GSE117872`; please refer to the original article [13] for further details on the experimental setup).

In detail, we selected single cells labeled as "good data" on the GEO repository e performed variant calling (the whole procedure is described in detail in the Supplementary Material). $4,924,559$ unique variants were detected on a total of $1,116$ single cells included in all datasets. Given the known limitations and the high levels of experimental noise of scRNA-seq data, we then applied a number of quality-control filters on variants, to ensure high confidence to the calls and to reduce the number of both false alleles and miscalls. In particular, we *removed*: (i) indels and other structural variants – to limit the impact of possible sequencing and alignment artifacts, (ii) variants mapped on mitochondrial genes, (iii) variants on positions with total read counts < 5 in $> 50\%$ of the cells in each time point – to focus the analysis on well-covered positions, (iv) variants detected in less than 20% of both HN120P and HN137P (*single-end*) cells – to focus on recurrent variants, (v) variants detected (≥ 3 alternative reads) in *both* HN120P and HN137P (*single-end*) – to define a list of variants that clearly characterize the identity of the two primary cell lines. We finally selected the variants observed in at least 1 cell (≥ 3 alternative allele reads, ≥ 5 total reads) of HN120P and in exactly 0 cells of HN137P (*single-end*), *and* the variants observed in at least 1 cell (≥ 3 alternative allele reads, ≥ 5 total reads) of a HN137P (*single-end*) and in exactly 0 cells of HN120P.

As a result, we identified 67 single-nucleotide variants (SNVs) that are representative of HN120P cell identity, and that are present in 0 cells of HN137P (*single-end*). Such variants are observed in high frequency in HN120P *and* in HN137P (*paired-end*), HN137PCR, HN137PCRDH, HN137M, HN137MCR, whereas are not observed ($< 1\%$ of the cells) in HN120PCR, HN120PCRDH, HN120M, HN120MCR, HN120MCRDH and HN137P (*single-end*). In Figure 1A we display the mutational profiles of all single-cells in all datasets. The total allele reads matrix and the alternative allele reads matrix for such variants are provided as Supplementary Table 1.

Analogously, we identified 112 unique SNVs that are strongly informative for HN137P (*single-end*) identity, and that are present in 0 cells of HN120P (see Figure 1A). Such variants are observed in high frequency in HN137P (*single-end*) *and* in HN120PCR, HN120PCRDH, HN120M, HN120MCR, HN120MCRDH, whereas are not observed ($< 1\%$ of the cells) in HN137P (*paired-end*), HN137PCR, HN137PCRDH, HN137M, HN137MCR and HN120P. Supplementary Table 2 includes a summary of the analysis, in which for each SNV we report: genome position, reference and alternative alleles, rsID (if available), minor allele frequency (if available), the count and the ratio of single cells displaying the variant (total read count $>= 5$, alternative read count $>= 3$) in each dataset, the average total read count and the average alternative read count relative to the variant in each dataset.

From this analysis it is evident that the genotypic identity of HN120P cell line is inconsistent with that of the other HN120 datasets and with that of HN137P (*single-end*), whereas it is consistent with that of the remaining HN137 datasets. Conversely, the genotypic identity of HN137P (*single-end*) cell line is inconsistent with that of the other HN137 datasets and with that of HN120P, whereas it is consistent with that of all the other HN120 datasets. This consideration holds whether such SNVs are either germline or somatic, as genotypes are unquestionable footprints of cell identity (notice also that 177 on 179 variants have a rsID). These surprising results can be hardly explained by cancer-related selection phenomena, by random effects, or by sampling limitations. Instead, these observations suggest the presence of a methodological issue, which might be explained by a label swap of samples HN120P and HN137P (*single-end*).

This hypothesis is further supported by the single-cell transcriptomic analysis, which we performed via Seurat [15] (details are provided in the Supplementary Material). In Figure 1B one can find the t-SNE plot as computed on the 1000 most variable genes and in which single-cells are colored according to dataset label. Consistently with the genotype analysis, the transcriptomic analysis of single-cells highlights the presence of two distinct clusters, the first one including HN120P cells and all cells from HN137 datasets, excluded HN137P (*single-end*), the second one including HN137P (*single-end*) cells and all cells from HN120 datasets, excluded HN120P.

Unfortunately, we believe that this methodological error might have led to erroneous conclusions in [13, 14, 12]. In [13], for instance, the authors state that HN137 cell line is comprised of a mix of epithelial (ECAD+) and mesenchymal (VIM+) cells, whereas the HN120 cell line would include phenotypically homogeneous population of ECAD+ cells. However, by looking at the expression level of VIM (Figure 1C), one can clearly notice that this gene is up-regulated

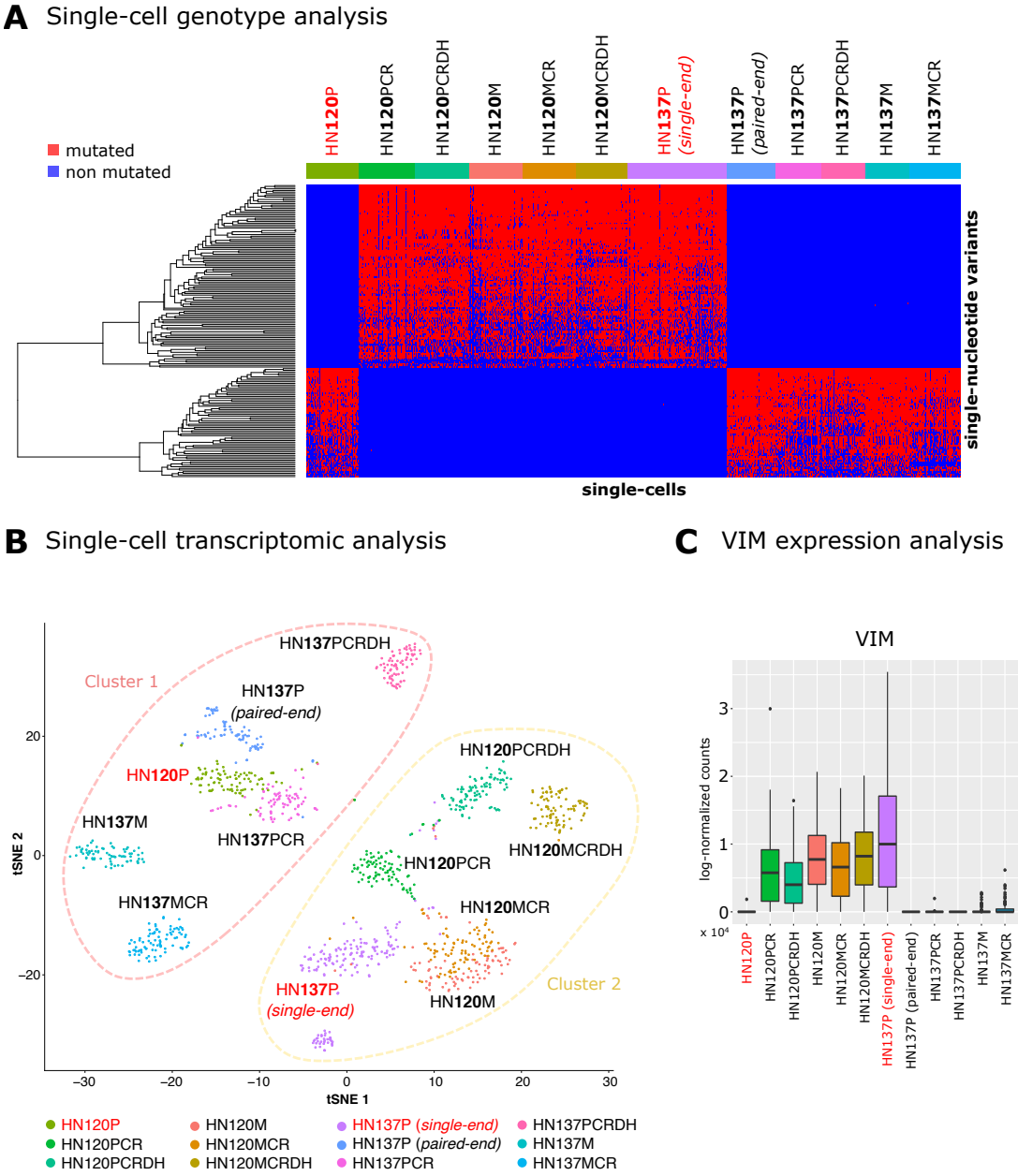


Figure 1: **(A)** The heatmap including the mutational profiles of all single cells of the HN120 and HN137 datasets is displayed (-P: primary line, -M: metastastic line, -CR: after cisplatin treatment, -CRDH: after drug-holiday). Red entries mark cells displaying a variant. For the ID of single-cells and SNVs please refer to Supplementary Table 1 and 2. **(B)** The t-SNE plot generated from the gene expression profiles of all single cells for all datasets is shown. **(C)** The distribution of the expression level of VIM on all single cells is shown with boxplots for all datasets.

in HN137P (*single-end*) and in all HN120 datasets, excluded HN120P, whereas is down-regulated (median = 0) in HN120P and in all HN137 datasets, excluded HN137(*single-end*).

Furthermore, in [13] the authors state that, in presence of cisplatin treatment, the heterogeneous HN137P cells demonstrate a progressive enrichment of ECAD, and the gradual depletion of VIM+ cells, until the latter get extincted. Conversely, from the supposedly homogeneous ECAD+ population of HN120P cells, the authors report the de novo emergence of VIM+ cells after two weeks of treatment. In order to explain this unexpected phenomenon, the authors

invoke the presence of a covert epigenetic mechanism that emerges under drug-induced selective pressure. Instead, we believe that this result might be easily explained by a label swap of HN120P and HN137P (*single-end*), as confirmed by the genotypic and transcriptomic analyses presented above (see Figure 1).

To conclude, the results presented in this work prove the efficacy of a joint analysis of the genotypic and transcriptomic identity of single-cells. Accordingly, this might represent a powerful instrument to uncover the elusive genotype-phenotype relation and to investigate the complex interplay underlying cancer evolution and drug resistance.

Data availability

A repository including data and scripts to replicate the analyses is available at this link: `https://github.com/BIMIB-DISCo/oral_squamous_longitudinal`.

References

[1] Shumei Chia, Joo-Leng Low, Xiaoqian Zhang, Xue-Lin Kwang, Fui-Teen Chong, Ankur Sharma, Denis Bertrand, Shen Yon Toh, Hui-Sun Leong, Matan T Thangavelu, et al. Phenotype-driven precision oncology as a guide for clinical decisions one patient at a time. *Nature communications*, 8(1):1–12, 2017.

[2] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491, 2011.

[3] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1):1–35, 2020.

[4] Devon A Lawson, Kai Kessenbrock, Ryan T Davis, Nicholas Pervolarakis, and Zena Werb. Tumour heterogeneity and metastasis at single-cell resolution. *Nature cell biology*, 20(12):1349–1360, 2018.

[5] Fenglin Liu, Yuanyuan Zhang, Lei Zhang, Ziyi Li, Qiao Fang, Ranran Gao, and Zemin Zhang. Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data. *Genome biology*, 20(1):1–15, 2019.

[6] Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular systems biology*, 15(6), 2019.

[7] Iain C Macaulay, Wilfried Haerty, Parveen Kumar, Yang I Li, Tim Xiaoming Hu, Mabel J Teng, Mubeen Goolam, Nathalie Saurat, Paul Coupland, Lesley M Shirley, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature methods*, 12(6):519, 2015.

[8] Anna S Nam, Kyu-Tae Kim, Ronan Chaligne, Franco Izzo, Chelston Ang, Justin Taylor, Robert M Myers, Ghaith Abu-Zeinah, Ryan Brand, Nathaniel D Omans, et al. Somatic mutations and cell identity linked by genotyping of transcriptomes. *Nature*, 571(7765):355–360, 2019.

[9] Lucrezia Patruno, Davide Maspero, Francesco Craighero, Fabrizio Angaroni, Marco Antoniotti, and Alex Graudenzi. A review of computational strategies for denoising and imputation of single-cell transcriptomic data. *Briefings in Bioinformatics*, 10 2020. bbaa222.

[10] Daniele Ramazzotti, Fabrizio Angaroni, Davide Maspero, Gianluca Ascolani, Isabella Castiglioni, Rocco Piazza, Marco Antoniotti, and Alex Graudenzi. Longitudinal cancer evolution from single cells. *bioRxiv*, 2020.

[11] Daniele Ramazzotti, Alex Graudenzi, Luca De Sano, Marco Antoniotti, and Giulio Caravagna. Learning mutational graphs of individual tumour evolution from single-cell and multi-region sequencing data. *BMC bioinformatics*, 20(1):210, 2019.

[12] Ankur Sharma. Hiding in plain sight: Epigenetic plasticity in drug-induced tumor evolution. *Epigenetics Insights*, 12:2516865719870760, 2019.

[13] Ankur Sharma, Elaine Yiqun Cao, Vibhor Kumar, Xiaoqian Zhang, Hui Sun Leong, Angeline Mei Lin Wong, Neeraja Ramakrishnan, Muhammad Hakimullah, Hui Min Vivian Teo, Fui Teen Chong, et al. Longitudinal single-cell RNA sequencing of patient-derived primary cells reveals drug-induced infidelity in stem cell hierarchy. *Nature communications*, 9(1):4931, 2018.

[14] Ankur Sharma and Ramanuj DasGupta. Tracking tumor evolution one-cell-at-a-time. *Molecular & cellular oncology*, 6(3):1590089, 2019.

[15] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

[16] Zilu Zhou, Bihui Xu, Andy Minn, and Nancy R Zhang. Dendro: genetic heterogeneity profiling and subclone detection by single-cell rna sequencing. *Genome Biology*, 21(1):1–15, 2020.

Acknowledgements

This work was partially supported by the Elixir Italian Chapter and the SysBioNet project, a Ministero dell'Istruzione, dell'Università e della Ricerca initiative for the Italian Roadmap of European Strategy Forum on Research Infrastructures by the AIRC-IG grant 22082. Support was also provided by the CRUK/AIRC Accelerator Award #22790, "Single-cell Cancer Evolution in the Clinic". We thank Giulio Caravagna, Chiara Damiani, Francesco Craighero and Lucrezia Patruno for helpful discussions.

Competing Interests

The authors declare that they have no competing financial interests.

Contributions

All authors performed the analyses, interpreted the results, drafted and approved the manuscript. AG and DR supervised the study.

Sequencing of viral samples. Viral samples, collected from infected hosts, can be sequenced using pipeline presented section 2.1.3. Like for single-cell data, different variant calling tools can be applied to retrieve mutation profiles. For examples, VarScan 2 [54] can be applied to produce VCF files for each viral samples including point mutations. The result is a continuous-value matrix which reports samples on the rows and mutations on the columns.

Like bulk data, which are a mixture of subpopulations, viral samples collected from human hosts are a mixture of virions (also called quasispecies) [50]. After variant calling, one obtains the frequency of the distinct viral mutations. Each mutation frequency reflects its presence in the corresponding portion of virions. Thus, we can distinguish two classes of mutations: (*i*) clonal (also named as fixed) mutations and (*ii*) intra-host minor mutations. The mutations in the former class display a high frequency (typically larger than 50%), and are included in the corresponding consensus sequences. Intra-host minor mutations, instead, are more likely to have emerged after the infections, so they are host-specific. Given the low frequency, such mutations are not included in the consensus sequences.

In this work, we designed methods to exploit both clonal and intra-host viral mutations profiles. In paper P#6, we used both mutation types to improve the reconstruction of the viral evolution, whereas in paper P#7 we used only intra-host minor mutations to reconstruct the mutational signatures underlying mutational processes.

In addition, we developed a dedicated pipeline to obtain VCF files from FASTQ files of viral samples, which is released as a Nextflow (i.e., workflow manager system [122]) file, so to provide an automated and user-friendly tool to call mutations from distinct deep sequencing protocols. The pipeline is part of the published protocol for the discovery of viral mutational signatures (named VirMutSig), included in the appendix P#A1. The pipeline file is also provided in the corresponding GitHub repository (see Appendix: Code repositories).

3

Methods for omics data analysis and integration

The research on new methods for omics data analysis and integration included in this work can be divided into two main branches: (*i*) methods to exploit gene expression profiles (section 3.1), (*ii*) methods to exploit mutational profiles (section 3.2).

An attempt to combine both omics data types in the context of cancer evolution is also provided in section 3.2.1. Details on data generation for both cases can be found respectively in Section 2.2 and 2.3.

3.1 Computational methods to exploit gene expression profiles

Stemming from my previous works on the topic [154, 175], during the PhD project important efforts were devoted to exploit gene expression profiles generated via either bulk or single-cell sequencing experiments (see section 2.2), for the evaluation of the metabolic heterogeneity of biological samples in distinct experimental scenarios. In the following, I discuss the results related to two main topics:

- Projection of gene expression profiles onto metabolic networks (section 3.1.1).
- Classification of cancer samples from the topological properties of metabolic networks (section 3.1.2).

Both topics are described in the following, together with the related articles.

3.1.1 Projection of gene expression profiles onto metabolic networks

As anticipated in the introduction 1.1, metabolic reprogramming of cancer cells is needed for the initiation and progression of the disease [123]. In fact, cancer cells alter their metabolic pathways to sustain their proliferative behaviour, by increasing the bioenergetic and biosynthetic demand. Cancer driver events and nutrient availability in the tumor microenvironment determine the fluxes through the metabolic pathways (i.e., the rate of turnover of molecules from substrates to products). The coupled action of driver genetic alterations and environmental perturbations determine the heterogeneity observed among oncological patients, and among the cancer cell subpopulations coexisting in single tumors [241].

Metabolome data can provide systematic information about the molecular mechanisms, and are suitable to help the identification of biomarkers for many diseases. Overall, metabolome data are complementary to genome and transcriptome data and allow us to fill the gap between genotype and phenotype, defined as the functional output of the biological interplay with the microenvironment. In principle, current high-throughput technologies allow researchers to collect systemic information about the state of the metabolism of a given tissue [145], even at a single-cell resolution [223].

Unfortunately, such technologies are still in their infancy, so publicly available metabolomic datasets are rare, or include a limited numbers of metabolites [233], in particular for single-cell experiments [216]. To overcome this limitation, one can exploit the gene expression profiles generated via either bulk or single-cell RNA-sequencing experiments (see Section 2.2) to characterize the metabolic states of a given biological sample. In this regard, it is first important to delineate how cellular metabolism is usually modelled.

Metabolic networks. Cellular metabolism is usually represented as a network of biochemical reactions either via genome-scale models [82, 143] or core models, which represent only a selected portion of the whole metabolism [118]. The network is represented with a bipartite digraph where nodes are either reactions or metabolites. More in detail, metabolites (substrates) are connected to a reaction, which is linked to other metabolites (products). Metabolic networks are integrated with further information, such as the Gene-Protein-Reaction (GPR) associations, i.e., logical formulas that describe how gene products concur to catalyze a given reaction. In particular, a functional enzyme can exist in different isoforms, which can be composed of many subunits. Isoforms and subunits are synthesized by specific genes. So, GPRs describe how gene products constitute a functional enzyme.

Projection of gene expression profiles. In [154] we propose a method called Metabolic Reaction Enrichment Analysis (*MaREA*) to project the gene expression

profiles, obtained as described in section 2.3, onto a metabolic network, so to estimate the metabolic activity of a given sample. The idea is to consider the gene expression levels associated with a given reaction as a proxy of its activity. More specifically, the rate of transformation from substrates to products via a biochemical reaction correlates with the abundance of the corresponding catalytic enzyme, which in turn, is a function of the corresponding gene expression levels. Thus, we proposed to exploit gene expression profiles to compute a Reaction Activity Score (RAS) for each reaction included in a given metabolic model, so to obtain a reaction score profile for each sample in the original dataset. Notice that this approach is general and can be applied to either bulk or single-cell transcriptomic data.

Metabolic stratification of cancer samples. In [154], we computed the RAS profiles of a large number of breast cancer patients obtained from The Cancer Genome Atlas TCGA Research Network: https://www.cancer.gov/tcga. Thanks to clustering analysis, we could identify homogeneous groups with similar survival expectation.

Performing such analyses is often difficult for people not skilled in computer science, as the necessary bioinformatics tools tend not to be easy to install or do not provide a clear interface. Fortunately, platforms such as Galaxy Project [142] are designed to provide a user-friendly graphical interface and a repository of pre-installed computational tools to a broad range of researchers with different backgrounds. We have decided to port our method MaREA on the platform, so that it can become part of workflow analyses of gene expression data. In addition, we have also recently installed a server in the Elixir infrastructure that one can freely use to perform our analysis. The related application note P#3 is presented below.

Computational and Structural Biotechnology Journal 18 (2020) 993–999

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

MaREA4Galaxy: Metabolic reaction enrichment analysis and visualization of RNA-seq data within Galaxy

Chiara Damiani [a,b,c,*,1], Lorenzo Rovida [b,1], Davide Maspero [b,d], Irene Sala [b], Luca Rosato [b], Marzia Di Filippo [c,e], Dario Pescini [c,e], Alex Graudenzi [f], Marco Antoniotti [b], Giancarlo Mauri [b,c]

[a] Dept. of Biotechnology and Biosciences, Universitá degli Studi di Milano-Bicocca, Milan, Italy
[b] Dept. of Informatics, Systems and Communication, Universitá degli Studi di Milano-Bicocca, Milan, Italy
[c] SYSBE-IT/SYSBIO Centre of Systems Biology, Universitá degli Studi di Milano-Bicocca, Milan, Italy
[d] Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[e] Department of Statistics and Quantitative Methods, Universitá degli Studi di Milano-Bicocca, Milan, Italy
[f] Institute of Molecular Bioimaging and Physiology, Consiglio Nazionale delle Ricerche (IBFM-CNR), Segrate, Milan, Italy

ARTICLE INFO

Article history:
Received 18 October 2019
Received in revised form 20 March 2020
Accepted 8 April 2020
Available online 17 April 2020

Keywords:
RNA-seq
Metabolism
Galaxy
Sample stratification
TCGA

ABSTRACT

We present MaREA4Galaxy, a user-friendly tool that allows a user to characterize and to graphically compare groups of samples with different transcriptional regulation of metabolism, as estimated from cross-sectional RNA-seq data. The tool is available as plug-in for the widely-used Galaxy platform for comparative genomics and bioinformatics analyses. MaREA4Galaxy combines three modules. The Expression2RAS module, which, for each reaction of a specified set, computes a Reaction Activity Score (RAS) as a function of the expression level of genes encoding for the associated enzyme. The MaREA (Metabolic Reaction Enrichment Analysis) module that allows to highlight significant differences in reaction activities between specified groups of samples. The Clustering module which employs the RAS computed before as a metric for unsupervised clustering of samples into distinct metabolic subgroups; the Clustering tool provides different clustering techniques and implements standard methods to evaluate the goodness of the results.

© 2020 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

In the last recent years, life sciences have witnessed a renewed focus on phenotype level phenomena. Accordingly, there has been an increasing attention towards cellular metabolism, which is regarded as the ultimate level of phenotype, reflecting the response of biological systems to regulatory and environmental changes.

Alteration of metabolic processes plays a pivotal role in many pathologies, such as cancer, metabolic syndromes, neurodegenerative diseases [10,16], as well as in aging processes [12].

Although quantification of metabolites has become more and more feasible [9], the difficulty of inferring changes in metabolic pathways based on metabolomics data [3] is pushing the need to understand metabolic alterations by leveraging gene expression data.

To this end, computational strategies are being proposed to integrate *–omics* data into metabolic networks [13,18,14]. Within this context, we have recently introduced the pipeline MaREA (Metabolic Reaction Enrichment Analysis) [8]. MaREA characterizes the metabolic disregulations that distinguish sets of individuals, by projecting RNA-seq data onto metabolic networks, without requiring explicit metabolic measurements.

MaREA computes a Reaction Activity Score (RAS) for each metabolic reaction and each sample/individual, based on the read count of the set of genes that encode the catalyzing enzyme(s). The scores are first used as features for cluster analysis and then to visualize the metabolic disregulations that distinguish the identified clusters, in a form understandable to life scientists.

In [8], we have demonstrated that MaREA can efficiently stratify cancer patients according to their metabolic activity, as proven by significantly different survival expectancy. Moreover, MaREA proved to be able to readily capture metabolic differences between

* Corresponding author at: Dept. of Biotechnology and Biosciences, Universitá degli Studi di Milano-Bicocca, Milan, Italy.
E-mail address: chiara.damiani@unimib.it (C. Damiani).
[1] Equal contributors.

https://doi.org/10.1016/j.csbj.2020.04.008
2001-0370/© 2020 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

two conditions, such as the properties that distinguish normal from tumor samples.

The MaREA pipeline is highly versatile and can be applied to virtually any study aiming at comparing the metabolism of samples in different conditions or experimental settings.

In [8], we released a MATLAB-based tool that implements the methodology. However, some technical barriers current limit the application of the pipeline: MATLAB is a proprietary software and many life scientists do not have the software licence; moreover, the tool is not web-based and, therefore, it requires local resources; also, most life scientists are not familiar with the MATLAB environment.

To overcome these limitations, we present here the freely available open source web-based tool MaREA4Galaxy which embeds the MaREA pipeline within the widely-used platform Galaxy [1].

Galaxy is a user-friendly web-based workflow system that allows biomedical researchers to use computational biology tools even without sophisticated computer science skills.

As compared to other user-friendly web-based metabolic network visualization tools, such as Escher [11] or Fame [2], which mainly focus on Flux Balance Analysis, MaREA4Galaxy specifically enables metabolic reaction enrichment analysis of gene expression data which may have been obtained directly within the Galaxy environment, for instance by using Galaxy tools to produce read counts from raw RNA-seq data. MaREA4Galaxy automatically recognizes most common gene nomenclature systems. It also enables cluster analysis of samples based on reaction activities, as well as on any other features. The cluster analysis module implements new functionalities, as compared to the previously released MATLAB-based MaREA tool. New clustering algorithms have been included, as well as new instruments for the evaluation of clustering goodness and for the selection of optimal number of clusters.

Moreover, MaREA4Galaxy inherits the benefits of Galaxy. In particular, Galaxy allows to build multi-step computational analyses. It allows users to upload their own data, as well as to interface with public databases, and enables researchers to perform the text manipulation required to properly format data for analysis without requiring advanced programming skills. Galaxy can be downloaded, customized and installed either locally or on a dedicated server. It also provides a comprehensive documentation.

In order to illustrate the functionalities of MaREA4Galaxy, we show a novel example on real data obtained from The Cancer Genome Atlas (TCGA) [17]. In particular, we perform an unsupervised cluster analysis of the gene expression of liver hepatocellular carcinoma tumors and we analyze the obtained clusters.

2. Implementation and availability

Following the recommendation by Galaxy's core developers' team, the back-end development of MaREA4Galaxy is based on Python and the front-end development on XML. The interaction between front-end and back-end is based on the template engine Cheetah. MaREA4Galaxy is built on top of the following libraries: `lxml`, `svglib`, `reportlab`, `pandas`, `scipy`, `python-libsbml`, `matplotlib`, `numpy` and `scikit-learn` for clustering analysis.

MaREA4Galaxy is stored in a versioned code archive in ToolShed, at: https://toolshed.g2.bx.psu.edu/repos/bimib/marea. ToolShed allows the administrators of the hundreds of public and private Galaxy servers worldwide to easily install MaREA4Galaxy, as well as any other Galaxy utility, into their instances.

Once installed, MaREA4Galaxy appears in the Galaxy toolbar (left bar) under the name MaREA (see Fig. 1 for an example). A demo of MaREA4Galaxy is available at:http://bimib.disco.unimib.it:5555.

3. Functionalities and workflow

MaREA4Galaxy processes datasets stored in the history panel of Galaxy. These datasets can be uploaded directly from the user's computer as structured text file by using, e.g., the Galaxy built-in tool `Get Data`, or obtained as output of intermediary analyses performed with other tools.

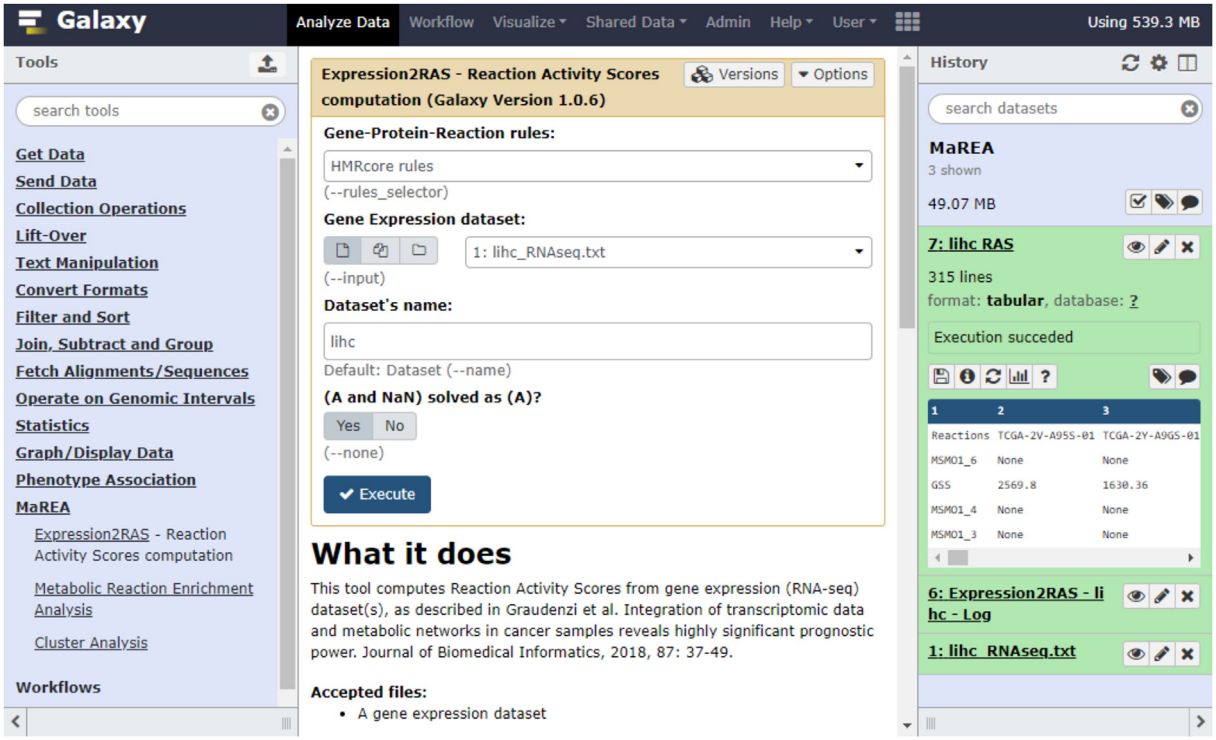


Fig. 1. Screenshot of the MaREA4Galaxy interface. The module for RAS computation is illustrated. In particular, the built-in (default) HMRcore GPR rules are chosen. In the 'add dataset' field there is the RNA-seq dataset which has been previously uploaded and that appears in green in the History panel on the right.

MaREA4Galaxy consists in three interconnected modules that may also work independently.

- The RASs computation module (Expression2RAS).
- The metabolic reaction enrichment analysis module (MaREA).
- The cluster analysis module (Clustering).

As better detailed in the following sections, the Expression2RAS module computes a RAS for each reaction in each sample. The MaREA module allows to visualize on a metabolic map the metabolic reactions that are up- or down- regulated in different groups of samples either defined *a priori* or identified by the Clustering module. The Clustering module allows to identify sample subgroups (or clusters) that share similar metabolic properties, ideally by employing the RAS computed by the Expression2RAS module. Any other data can however be used as feature for unsupervised clustering of samples. The metabolic differences between the clusters thus obtained can, in turn, be analyzed with the MaREA module.

3.1. Computation of reaction activity scores

The Expression2RAS tool selects and extracts metabolic genes from a gene-expression dataset and, by solving Gene-Protein-Reaction (GPR) association rules, computes a *Reaction Activity Score* (RAS) for metabolic reactions of interest, as illustrated in [8]. The assumption is that enzyme isoforms contribute additively to the overall activity of a given reaction, whereas enzyme subunits limit its activity, by requiring all the components to be present for the reaction to be catalyzed [8].

3.1.1. Input

The Expression2RAS tool (Fig. 1) takes two main inputs: 1) the list of GPRs; 2) the normalized read count of genes from a given cross-sectional RNA-seq dataset, as, e.g., RPKM (Reads per Kilobase per Million mapped reads) or TPM (Transcripts Per Kilobase Million).

The first input is a representation of the metabolic model being studied and it is basically a dictionary (key-value data structure), which associates a set of genes to each metabolic reaction. Both reactions and genes must be defined by a unique identifier.

Boolean operators AND and OR define the relationship between genes and enzymes. The AND operator joins genes that encode for different subunits of the same enzyme, whereas the OR operator joins genes that encode for isoforms of the same enzyme.

For the user's convenience, two human metabolic network models have been made directly available within the tool: HMRcore and Recon 2.2. HMRcore corresponds to the set of GPR rules included in the core model of central carbon metabolism introduced in [6] and was used and curated in [4,7,5,8], whereas Recon 2.2 [15] is a genome-wide model encompassing virtually all reactions encoded in human metabolism. However, the user can also opt to upload any custom metabolic network model of her/his choice.

The ID of genes in the dataset must of course coincide with the ID used in the GPR rules. In case built-in GPRs are used, the following gene nomenclatures are automatically recognized: HUGO ID, Ensemble ID, HUGO symbol, Entrez ID.

In case of missing expression value, referred to as NaN (Not a Number), for a gene joined with an AND operator in a given GPR rule, the user can choose to solve the rule 'A AND NaN' as A, or to disregard it tout-court (i.e., treated as NaN).

3.1.2. Output

The tool simply returns as output a dataset reporting the RAS computed for each sample for each reaction in the chosen metabolic network. The RAS dataset is displayed in the History panel.

3.2. Metabolic reaction enrichment analysis

The MaREA tool statistically compares the RAS of user-defined groups of samples [8] and visualizes the identified differences.

According to the user's preference the tool performs the following comparisons.

- Pairwise comparison of each group against all other groups.
- Comparison of each group against the rest of the samples.
- Comparison of each group against a user-defined control group.

3.2.1. Input

The MaREA tool (Fig. 2) takes as main input the Reaction Activity Scores of each sample, as computed by the Expression2RAS module and, if given, the eventual partition of samples/patients into distinct classes.

The input RAS dataset can be organized in two alternative ways: 1) two or more separate RAS datasets, each relative to a different set of samples/patients; 2) a unique RAS dataset for all samples/patients, plus a file that associates to each sample its affiliation to a set.

As (optional) input, the user may also supply a graphical map of the metabolic network for an efficient visualization of the analysis outputs. If the HMRcore model is chosen, the corresponding map is included within the tool and does have not to be uploaded. The metabolic map format is a `svg` file, reporting metabolites and products of each reaction linked with an arrow, whose ID matches the name of the reaction in the GPR file.

The following advanced options can also be displayed and specified.

- The P-Value threshold, used for significance Kolmogorov-Smirnov (KS) test, to verify whether the distributions of RASs over the samples in two sets are significantly different.
- The threshold of the fold-change between the average RAS of two groups. Among the reactions that pass the KS test, only fold-change values larger than the indicated threshold will be visualized on the output metabolic map.
- optional outputs to be displayed in the History panel.

The reader is referred to [8] for further theoretical aspects regarding the options above, whereas further technical details regarding formatting of input files are available in the help section in Galaxy.

3.2.2. Output

MaREA returns for each evaluated comparison, a collection output in the History including the following items.

- A table reporting the fold-change between RASs and p-value of the Kolmogorov-Smirnov test.
- The modified metabolic map (whenever supplied as input). Reactions up-regulated in the first class as compared to the second class are marked in red, whereas reactions down-regulated in the former are marked in blue. Thickness of arrows is proportional to the fold-change between the average RASs of the two classes. Non-Classified reactions, i.e., reactions without information about the corresponding gene-enzyme rule, are marked in black. Reactions that display a non-significant p-value or a RAS fold-change below the threshold are marked in gray color. The pdf of the map can be directly visualized within Galaxy. The user can also download the svg format of the map in order to apply changes.
- A log file, reporting possible warning or error messages. Problems that prevent the pipeline's functioning, such as wrong format of files, gene ID type not supported or duplicated IDs,

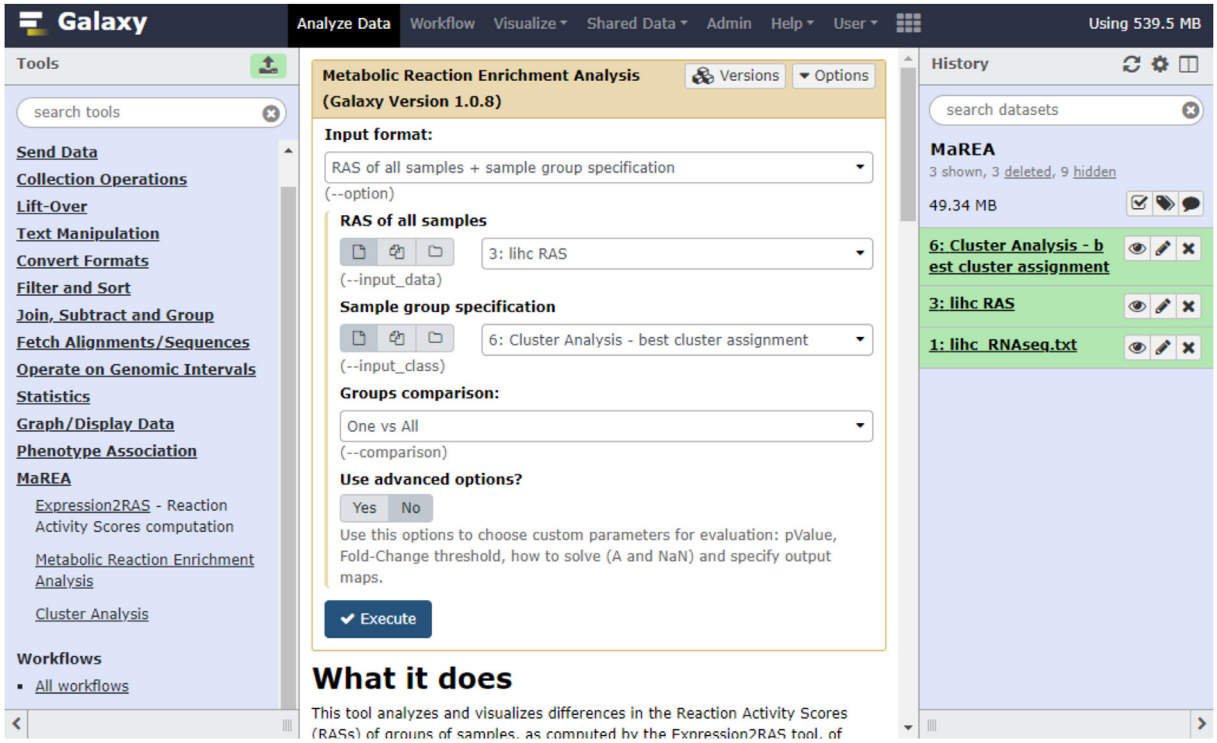


Fig. 2. Screenshot of the MaREA4Galaxy interface. The module for metabolic reaction enrichment analysis is illustrated. The input format option 'RNAseq dataset of all samples + sample group specification' has been selected and the best clustering obtained with the k-means algorithm in the History has been selected as sample group specification.

insufficient number of classes, as well as minor problems, such as extra-columns, duplicated labels, missing gene values, or empty classes, are properly notified in detail.

3.3. Cluster analysis

The Clustering tool has been conceived to cluster gene expression data, by using the RAS scores computed by MaREA4Galaxy as features, given its efficacy in stratifying cancer patients according to metabolic phenotype, as demonstrated in [8]. However, it is suited to cluster observations in any dataset in which rows indicate different variables/features and columns different observations.

The Clustering tool implements three of the main existing algorithms to cluster data, namely: K-means, agglomerative clustering and DBSCAN (Density Based Spatial Clustering of Applications with Noise). Parameters and outputs of the tool are specific of each algorithm as briefly described in the following. A screenshot of this feature of the tool is reported in Fig. 3.

3.3.1. K-means

Given that K-means clustering requires the number of clusters k to be set by the user, and that it is usually difficult to know the correct number of clusters *a priori*, the Clustering tool allows to evaluate different values of k and provides standard methods to estimate the goodness of each clustering in order to choose the best one. In particular, the elbow plot is generated, which allows to identify the "elbow" (the point of inflection on the curve) as the best candidate. The tool also generates a *silhouette* plot for each k, which reports the cohesion and the separation indexes of each element. The tool also computes the silhouette score of each element, as well as the average silhouette of each k, returning the k with the best (highest) silhouette. The user can specify the minimum and maximum number of clusters to be evaluated and whether elbow and dendrogram plots must be generated.

3.3.2. Agglomerative clustering

In the case of agglomerative clustering, the hierarchical output illustrated by the dendrogram facilitates the choice of the best clustering. The Clustering tool returns the set of clusters obtained when cutting the dendrograms at different points. The user can specify the minimum and maximum number of clusters to be tested and whether the dendrogram plot must be generated.

3.3.3. DBSCAN

The DBSCAN method automatically chooses the number of clusters, based on parameters that define when a region is to be considered dense. Custom parameters may be used, namely the maximum distance between two samples for one to be considered as in the neighborhood of the other and the number of samples in a neighborhood for a point to be considered as a core point.

4. Application example

To illustrate MaREA4Galaxy functioning, we show here an example of application. We analyze a liver hepatocellular carcinoma RNA-seq dataset taken from the TCGA pancancer study, including 372 patients. The original data is available at:http://download.cbioportal.org/blca_tcga.tar.gz.

The goal of our example is to identify patients' cohorts with different metabolic features, without assuming any prior knowledge about the dataset. To this end, we first upload our dataset in the History panel of Galaxy by means of the `Get Data` tool. Based on the HMRcore metabolic map, we then compute the RAS for each reaction in each patient, by means of Expression2RAS tool, see Fig. 1.

We then switch to the Clustering tool and we select as input dataset the RASs computed before. As a proof of principle, we choose K-means as clustering algorithm. We test a number of clusters k from 2 to 5, and we indicate that we want to generate both

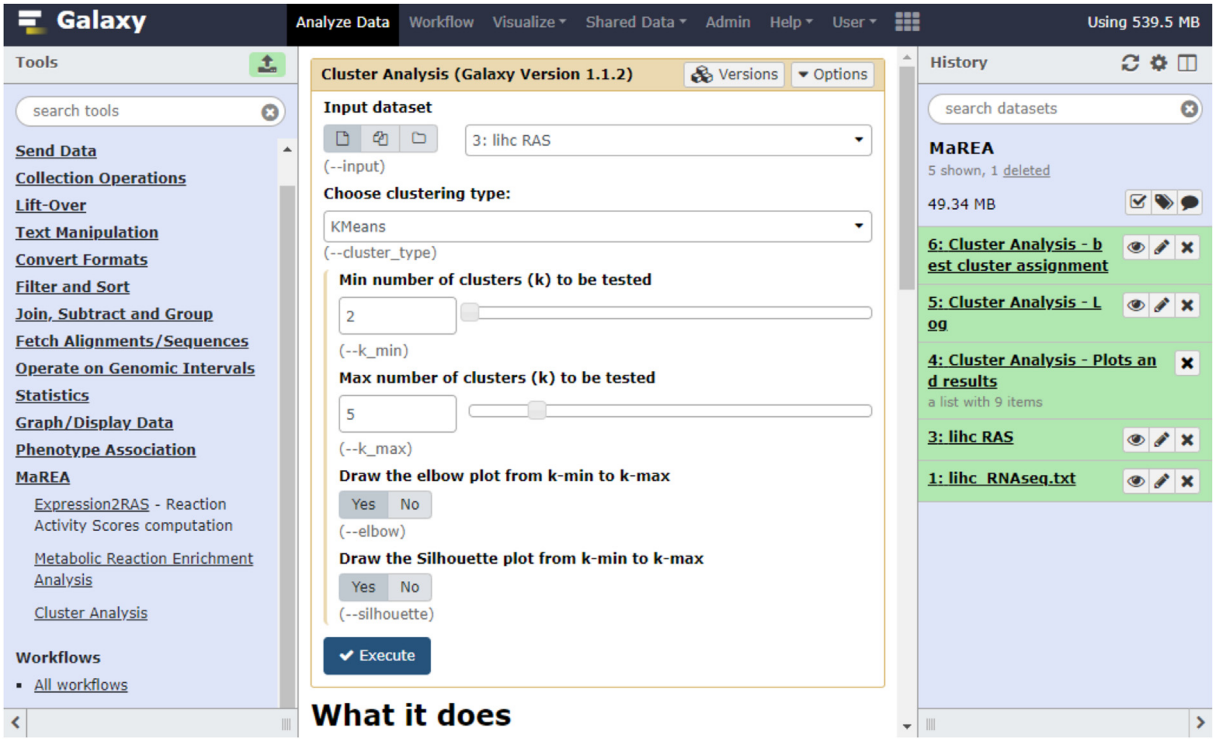


Fig. 3. Screenshot of the MaREA4Galaxy interface. The module for cluster analysis is illustrated. The RAS computed by the MaREA tool have been selected as input dataset and K-means has been chosen as clustering method. 2 to 5 number of clusters will be tested. The elbow and silhouette plots will be generated.

elbow and silhouette plots for each tested k (see a screenshot of the tool in Fig. 3).

The execution of the tool returns the clustering output for each value of k, and indicates as best clustering the one that maximizes the average silhouette coefficient. In this case, the tools returns $k = 2$ as the best clustering. The generated elbow and silhouette plots are reported in Fig. 4. It can be noticed that, although we have identified a good stratification of liver cancer patients, by qualitative observation of the elbow plot one may also choose $k = 3$ as best clustering.

Finally, we can use the MaREA tool to promptly analyze the differences between the two patients' cohorts. As shown in the screenshot in Fig. 2, we select this time the input format option 'RNAseq of all samples +sample group specification' and we select the best clustering output obtained with Clustering as sample group specification file. We flag the option of generating the pdf map and once we execute the tool we obtain the map reported in Fig. 5.

Although a few reactions significantly differ between the two cluster, an expert who is familiar with this classical representation of central carbon metabolism can immediately notice (Fig. 5) the main metabolic features that distinguish the two patients cohorts.

For example, in the first group, the upper glycolytic pathway is up-regulated, whereas lower glycolysis, upstream of lactate production, is down-regulated. Lactate production from pyruvate is instead up-regulated, indicating that pyruvate production derives from alternative routes, such as from the amino acid serine derived from glucose. The reactions that go from serine to pyruvate are indeed up-regulated. Other differences involve the utilization of glutamine and synthesis of amino acids derived from glutamine,

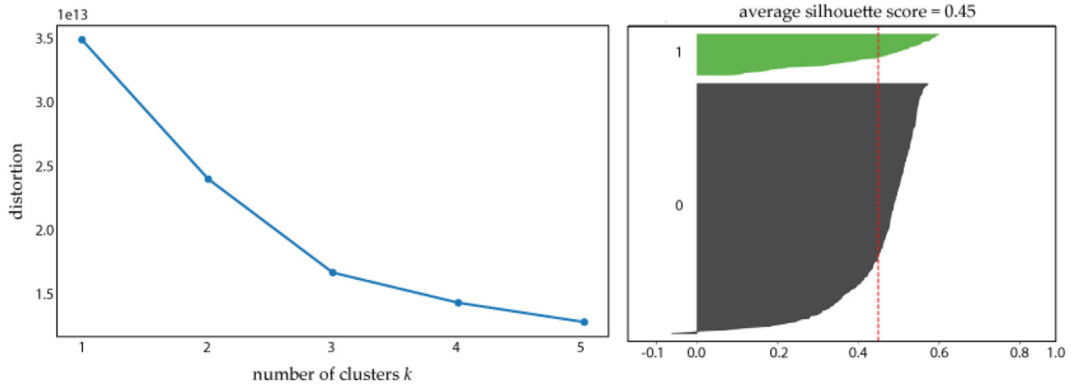


Fig. 4. Evaluation of clustering goodness by MaREA4Galaxy. Left panel: elbow plot generated by the Clustering module, showing an elbow for $k = 3$. Right panel: silhouette plot generated by the Clustering module for $k = 2$, which has been returned as best clustering according to the average silhouette score reported in the plot's title.

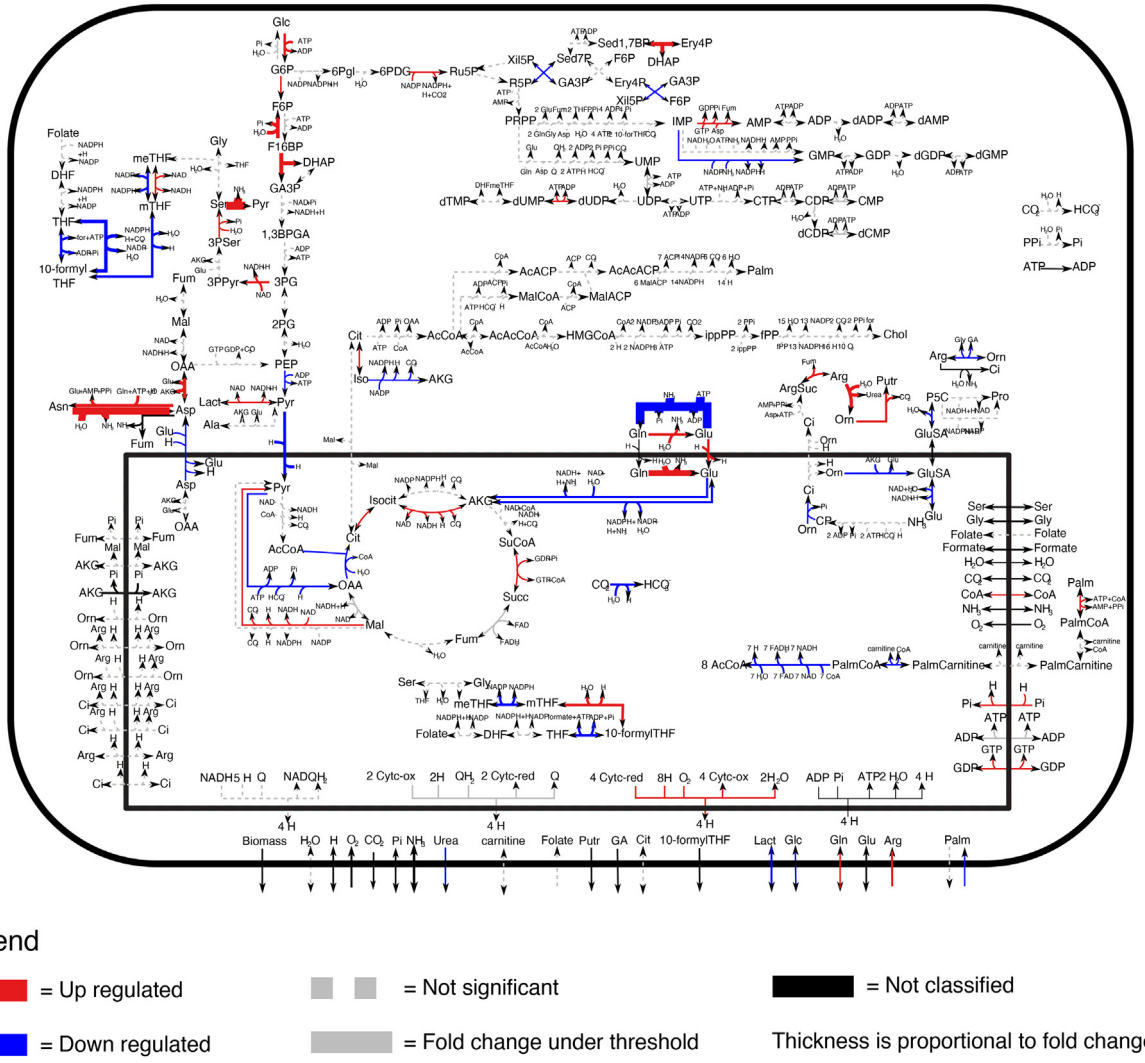


Fig. 5. Example of metabolic map generated by MaREA4Galaxy. In the example, red arrows indicate reactions up-regulated, whereas blue arrows reactions down-regulated, in a subgroup of liver hepatocellular carcinoma patients. Black arrows refer to reactions without information about the corresponding gene-enzyme rule. Dashed gray arrows refer to non significant disregulations according Kolmogorov-Smirnov test with p-value 0.01. Solid gray arrows refer to reactions with a variation lower then 20%. As output maps are provided as vector graphics (in svg/pdf file formats), they can be zoomed-in at will.

such as asparagine and aspartate, as well as production of putrescine and urea in the urea cycle.

5. Conclusion

We have shown how with a few intuitive steps, without the need to set technical parameters, and in a very short time, MaREA4Galaxy enables to uncover and characterize the differences in metabolic activity observed in different sample subgroups, as in the case of cancer patients.

Being empowered by the well-known open and web-based platform Galaxy for performing accessible, reproducible, and transparent bioinformatics science, MaREA4Galaxy can support many life scientists who may have little knowledge of computational methods for analyzing the metabolic variability underlying gene-expression datasets, no matter whether collected in their labs or available in public databases, thus paving the way to tackle metabolic plasticity and heterogeneity.

As an example, we have shown a novel application of the MaREA pipeline to liver hepatocellular carcinoma and we have identified two groups with well distinct metabolic features. Investigating the implications of these differences is out of the scope of this work. However, it would be interesting to analyze whether the two groups of patients differ in other aspects, such as their prognosis, (epi) genomic makeup or regulation of signaling pathways.

A better understanding of the fundamental causes of metabolic heterogeneity is important for personalised treatment of the many diseases involving metabolic alterations, as well as for targeted nutrition recommendations and intervention the field of personalized nutrition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Chiara Damiani: Conceptualization, Investigation, Supervision, Writing - original draft, Writing - review & editing. **Lorenzo Rovida:** Software. **Davide Maspero:** Conceptualization, Data curation. **Irene Sala:** Software. **Luca Rosato:** Software. **Marzia Di**

Filippo: Visualization. **Dario Pescini:** Visualization. **Alex Graudenzi:** Writing - review & editing. **Marco Antoniotti:** Writing - review & editing. **Giancarlo Mauri:** Writing - review & editing, Funding acquisition.

Acknowledgments

The institutional financial support to SYSBIO – within the Italian Roadmap for ESFRI Research Infrastructures – is gratefully acknowledged. CD and GM received funding from FLAG-ERA grant ITFoC.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.csbj.2020.04.008.

References

[1] Afgan E, Baker D, Van den Beek M, Blankenberg D, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Eberhard C, et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucl Acids Res 2016;44(W1):W3–W10.
[2] Boele J, Olivier BG, Teusink B. Fame, the flux analysis and modeling environment. BMC Syst Biol 2012;6(1):8.
[3] Damiani C, Colombo R, Di Filippo M, Pescini D, Mauri G. Linking alterations in metabolic fluxes with shifts in metabolite levels by means of kinetic modeling. In: Italian Workshop on Artificial Life and Evolutionary Computation. Springer; 2016. p. 138–48.
[4] Damiani C, Di Filippo M, Pescini D, Maspero D, Colombo R, Mauri G. popFBA: tackling intratumour heterogeneity with Flux Balance Analysis. Bioinformatics 2017;33(14):i311–8.
[5] Damiani C, Maspero D, Di Filippo M, Colombo R, Pescini D, Graudenzi A, et al. Integration of single-cell rna-seq data into metabolic models to characterize tumour cell populations. PLOS Computat Biol 2018;15(2):e1006733.
[6] Di Filippo M, Colombo R, Damiani C, Pescini D, Gaglio D, Vanoni M, Alberghina L, Mauri G. Zooming-in on cancer metabolic rewiring with tissue specific constraint-based models. Computat Biol Chem 2016;62:60–9.
[7] Graudenzi A, Maspero D, Damiani C. Modeling spatio-temporal dynamics of metabolic networks with cellular automata and constraint-based methods. In: Cellular Automata. ACRI 2018. Lecture Notes in Computer Science. vol. 11115. Springer, Cham; 2018, p. 16–29..
[8] Graudenzi A, Maspero D, Di Filippo M, Gnugnoli M, Isella C, Mauri G, Medico E, Antoniotti M, Damiani C. Integration of transcriptomic data and metabolic networks in cancer samples reveals highly significant prognostic power. J Biomed Inform 2018;87:37–149.
[9] Holmes E, Wilson ID, Nicholson JK. Metabolic phenotyping in health and disease. Cell 2008;134(5):714–7.
[10] Hotamisligil GS. Inflammation and metabolic disorders. Nature 2006;444 (7121):860.
[11] King ZA, Dräger A, Ebrahim A, Sonnenschein N, Lewis NE, Palsson BO. Escher: a web application for building, sharing, and embedding data-rich visualizations of biological pathways. PLoS Comput Biol 2015;11(8).
[12] López-Otín C, Galluzzi L, Freije JM, Madeo F, Kroemer G. Metabolic control of longevity. Cell 2016;166(4):802–21.
[13] Machado D, Herrgård M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. PLoS Comput Biol 2014;10(4):e1003580.
[14] Opdam S, Richelle A, Kellman B, Li S, Zielinski DC, Lewis NE. A systematic evaluation of methods for tailoring genome-scale metabolic models. Cell Syst 2017;4(3):318–29.
[15] Swainston N, Smallbone K, Hefzi H, Dobson PD, Brewer J, Hanscho M, Zielinski DC, Ang KS, Gardiner NJ, Gutierrez JM, et al. Recon 2.2: from reconstruction to model of human metabolism. Metabolomics 2016;12(7):1–7.
[16] Ward PS, Thompson CB. Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. Cancer Cell 2012;21(3):297–308.
[17] Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM, Network CGAR, et al. The cancer genome atlas pan-cancer analysis project. Nature Genet 2013;45(10):1113–20.
[18] Yizhak K, Chaneton B, Gottlieb E, Ruppin E. Modeling cancer metabolism on a genome scale. Mol Syst Biol 2015;11(6):817.

3.1.2 Classification of cancer samples from the topological properties of metabolic networks

The definition of accurate diagnostic methods leveraging omics data is a crucial objective of cancer data science. To this purpose, classification frameworks aim at exploiting the intrinsic heterogeneity of cancer samples via machine learning (ML) approaches, which have proven extremely effective in this context [197].

In section 3.1.1, we presented our method to calculate RASs by considering gene expression datasets and metabolic models. RASs are calculated for each reaction in the model to which a GPR is associated. Since reactions are nodes in the related metabolic network, we can consider RASs as weights for all the edges entering and leaving each reaction node. We can then obtain a sample-specific metabolic network by pruning edges with weights below a given relevance threshold and considering the remaining *giant component* (i.e., the biggest connected part of the pruned metabolic network [25]).

Our goal is twofold. We first want to evaluate the effectiveness of different machine learning methods in classifying samples considering the topological features of the corresponding giant component. Furthermore, we want to determine the best relevance threshold to prune the network and obtain the most representative giant component.

To this end, we tested 3 widely used ML methods (*i*) Multi-Layer Perceptrons (MLPs), (*ii*) Support Vector Machines (SVMs), and (*iii*) Random Forests (RFs). We also generated different giant components for each metabolic network scanning the relevant threshold values in the range between $[0, 1]$. Finally, for each of them, we computed the following topological features (*i*) Average degree, (*ii*) Average hierarchical degree level 2, (*iii*) Average hierarchical degree level 3, (*iv*) Average geodesic path length, and (*v*) Assortativity. Such metrics are described in the our paper included below (see paper P#4).

We tested the approach by classifying normal or breast cancer tissue samples included in a gene expression dataset (obtained from TCGA). Since we know the correct class of each entry, we can compute accuracy, precision and recall to evaluate the performance of the classifiers. In this regard, we highlight that more sophisticated metrics might be employed to evaluate the classification performance such as, e.g., the Brier's score [1] or the area under the ROC curve [15], and may be further investigated. We also compared the performances of the same ML approaches using the first 5 principal component of the metabolic gene expression profile matrix.

Our paper P#4, included in the following, shows that topological features of metabolic networks pruned considering a relevant threshold are sufficient to distinguish healthy from cancer samples. In particular, we have determined that applying the SVM method while setting the relevant threshold to 0.1 provides the best results in terms of accuracy, precision, and recall.

In other terms, our approach performs feature reduction, as we pass from an in-

put data matrix including thousands of features (i.e., metabolic genes) to 5 features only (i.e., the topological features of the metabolic networks), which are sufficient for a good classification. This result may, in turn, underlie important generic properties of metabolic component rearrangements after cancer initiation, and might deserve further investigation.

Send Orders for Reprints to reprints@benthamscience.net

88 *Current Genomics*, **2021**, *22*, **88-97**

RESEARCH ARTICLE

On the Use of Topological Features of Metabolic Networks for the Classification of Cancer Samples

Jeaneth Machicao[1,2,+,*], Francesco Craighero[3,+], Davide Maspero[3,4], Fabrizio Angaroni[3], Chiara Damiani[5,6,†], Alex Graudenzi[4,7,†,*], Marco Antoniotti[3,7,†] and Odemir M. Bruno[1,†,*]

[1]*São Carlos Institute of Physics, University of São Paulo, São Carlos, Brazil;* [2]*School of Engineering, University of São Paulo, São Paulo, Brazil;* [3]*Department of Informatics, Systems and Communication, University of Milan-Bicocca, Milan, Italy;* [4]*Institute of Molecular Bioimaging and Physiology, Consiglio Nazionale delle Ricerche (IBFM-CNR), Segrate, Milan, Italy;* [5]*Department of Biotechnology and Biosciences, University of Milan-Bicocca, Milan, Italy;* [6]*Sysbio Centre for Systems Biology, Milan, Italy;* [7]*Bicocca Bioinformatics, Biostatistics and Bioimaging Center (B4), University of Milan-Bicocca, Milan, Italy*

Abstract: ***Background*:** The increasing availability of omics data collected from patients affected by severe pathologies, such as cancer, is fostering the development of data science methods for their analysis.

***Introduction*:** The combination of data integration and machine learning approaches can provide new powerful instruments to tackle the complexity of cancer development and deliver effective diagnostic and prognostic strategies.

***Methods*:** We explore the possibility of exploiting the topological properties of sample-specific metabolic networks as features in a supervised classification task. Such networks are obtained by projecting transcriptomic data from RNA-seq experiments on genome-wide metabolic models to define weighted networks modeling the overall metabolic activity of a given sample.

***Results*:** We show the classification results on a labeled breast cancer dataset from the TCGA database, including 210 samples (cancer vs. normal). In particular, we investigate how the performance is affected by a threshold-based pruning of the networks by comparing Artificial Neural Networks, Support Vector Machines and Random Forests. Interestingly, the best classification performance is achieved within a small threshold range for all methods, suggesting that it might represent an effective choice to recover useful information while filtering out noise from data. Overall, the best accuracy is achieved with SVMs, which exhibit performances similar to those obtained when gene expression profiles are used as features.

***Conclusion*:** These findings demonstrate that the topological properties of sample-specific metabolic networks are effective in classifying cancer and normal samples, suggesting that useful information can be extracted from a relatively limited number of features.

ARTICLE HISTORY

Received: July 07, 2020
Revised: December 16, 2020
Accepted: December 18, 2020

DOI:
10.2174/1389202922666210301084151

Keywords: Metabolic networks, cancer sample classification, machine learning, RNA-seq data, topological properties, network pruning.

1. INTRODUCTION

The development of automated strategies for the classification of cancer samples in distinct categories (*e.g.*, subtypes, risk groups, *etc.*) is one of the key challenges in current biosciences [1]. On the one hand, this might lead to the discovery of efficient, personalized diagnostic, prognostic, and therapeutic strategies for cancer patients. On the other hand, it could allow unraveling some of the still undeciphered mechanisms and processes underlying cancer development, leading to a data-driven understanding of the disease.

It is known that effective classification and clustering of cancer samples can be achieved by employing the information on expression data [2–7], genomic alteration profiles [8, 9], interaction networks [10], and even signaling pathways [11, 12]. In this work, however, we specifically focus on the metabolic properties that may distinguish cancer from normal samples. In fact, metabolic deregulation is one of the key hallmarks of cancer [13–15], even if its underlying mechanisms are still partially unknown. In this respect, in recent years, an increasing number of computational strategies have been devised, in order to take advantage of the

*Address correspondence to this author at the São Carlos Institute of Physics, University of São Paulo, São Carlos, Brazil;
E-mails: alex.graudenzi@ibfm.cnr.it, bruno@ifsc.usp.br, machicao@usp.br
[+]*Co-first authors*; [†]*Co-senior authors*

 © 2021 Bentham Science Publishers

growing availability and reliability of -omics data to investigate the alterations of metabolism in cancer [16–19]. Very often, such data have been employed in constraint-based models, such as Flux Balance Analysis (FBA), in which metabolic fluxes are simulated to compare different experimental scenarios [20–24].

Moreover, more recently, approaches coupling constraint-based metabolic modeling with supervised machine learning algorithms have been proposed [25]. In our case, we explore for the first time the possibility of employing the topological properties of metabolic networks as input features of classification algorithms. To this end, we rely on an approach firstly introduced in [26,27] in which transcriptomic data, such as RNA-seq, are employed to determine the approximate activity value of the reactions included in a given metabolic network.

More in detail, by introducing a relevance threshold on the metabolic activity level, we pruned the original metabolic network to define individual-specific networks in which only the significantly active reactions are preserved. The topological properties of such individual-specific networks are then used as features to perform a supervised classification task via various algorithmic strategies and, in particular, Multi-Layer Perceptrons (MLPs), Support Vector Machines (SVMs) and Random Forests (RFs).

To investigate our hypothesis, this work presents the classification results in a simple scenario in which the sample categories are known a priori – cancer vs. normal – concerning the TCGA-BRCA breast cancer dataset [28], which includes 210 total samples.

We show that noteworthy classification performance can be achieved by using a few key topological properties of metabolic networks, *i.e.*, average degree, average hierarchical degree, average geodesic path length and assortativity. Interestingly, a similar pruning threshold (in the range $0.01 – 0.1$) is identified as optimal for all tested machine learning strategies, suggesting that it could be an effective choice to extract useful information from the "relevant" activity of metabolic networks, while discarding possible artifacts due to noisy observations. Overall, the best classification performance is obtained with SVMs and threshold 0.1, which exhibit 0.866 of (average) accuracy, 0.86 precision and 0.879 recall on the test set, after k-fold cross-validation and hyper-parameter estimation. Furthermore, we show that the best performing SVM classifier (with the optimal threshold) delivers similar classification performance with respect to an analogous classifier processing a reduced gene expression feature vector, as computed by selecting the 5 principal components on the list of 1673 metabolic genes from Recon2.2 [29].

These results prove that the projection of transcriptomic activity on metabolic networks provides useful information to efficiently classify cancer samples and might pave the way for the development of strategies for experimental hypothesis generation.

2. MATERIALS AND METHODS

2.1. Integration of RNA-seq and Metabolic Networks

As proposed earlier [26, 27], it is possible to project transcriptomic data onto human metabolic networks [30], to derive an approximate activity value for each metabolic reaction in any given sample.

We first employ an input metabolic network M such as the Human Metabolic Reaction (HMR) [31] or Recon [29, 32]. M is a bipartite-directed graph that includes two kinds of nodes: (*i*) metabolites (*i.e.*, substrates or products), and (*ii*) metabolic reactions. The edges in M connect either: (*i*) the substrates and the relative reaction, or (*ii*) a reaction and the relative products. The total number of nodes of M is N, whereas the total number of edges is E. Reaction nodes are associated with Gene-Protein-Reaction (GPR) rules, *i.e.*, logical formulas that describe the related catalyses *via* AND and OR logical operators. In particular, AND rules are employed when distinct genes encode different *subunits* of the same enzyme, whereas OR rules are used when distinct genes encode *isoforms* of the same enzyme.

RNA-seq data are then used to provide an approximate activity value to each reaction in the input network. In particular, our method takes as input a n (*genes*) × m (*samples*) matrix T in which each element $T_{g,s}$, $g = 1, \dots, n$, $s = 1, \dots, m$, includes the transcript level of gene g in sample s (the *Reads per Kilobase per Million* mapped reads – RPKM).

For each reaction in the input network $r \in G$ and for each sample $s = 1, \dots, m$, we define a *Reaction Activity Score* (RAS), by distinguishing two cases.

Reactions with GPR including an AND operator

$$RAS_{r,s} = \min\left(T_{g,s}: g \in \mathcal{A}_r\right), \quad (1)$$

where $\mathcal{A}_r$ is the set of genes that encode the subunits of the enzyme catalyzing reaction r.

Reactions with GPR including an OR operator

$$RAS_{r,s} = \sum_{g \in \mathcal{O}_r} T_{g,s}, \quad (2)$$

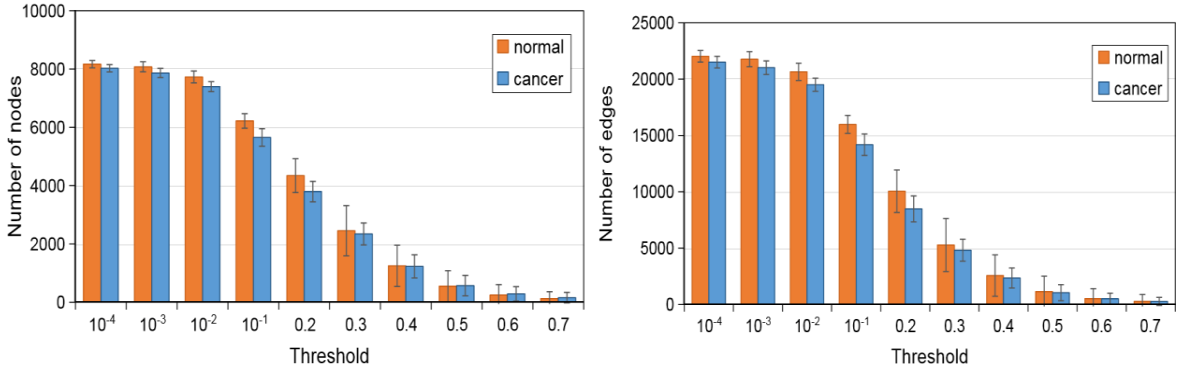
where $\mathcal{O}_r$ is the set of genes that encode isoforms of the enzyme that catalyzes reaction r.

In case of composite reactions, we respect the standard precedence of the two operators. The rationale underlying the definition of the RAS is that enzyme isoforms (OR) contribute *additively* to the overall activity of a certain reaction, whereas enzyme subunits (AND) *limit* its activity. RASs are finally normalized to obtain values in the range [0, 1] (with 0 meaning *no activity* and 1 meaning *maximum activity observed in the dataset*).

Even though this simplified approach neglects the heterogeneity of reaction kinetic constants, protein binding affinities and translation rates, it was proven effective in the investigation of cancer metabolic deregulation and in cancer sample stratification [26, 27].

2.2. Cancer Sample Classification via Metabolic Network Pruning

We define the sample-specific metabolic network of a given sample s as the weighted adjacency matrix W^s, which contains $N \times N$ elements, such that each element w_{ij}^s is equal to: (*i*) $RAS_{j,s}$ if i is a substrate of reaction j, (*ii*) $RAS_{i,s}$ if i is a reaction and j one of its products, (*iii*) 0 otherwise.



(a) (b)

Fig. (1). Number of nodes $\langle N^{T_l,S} \rangle$ (averaged on all samples) (a) and number of edges $\langle E^{T_l,S} \rangle$ (averaged on all samples) (b) of the giant component $G^{T_l,S}$ of the sample-specific metabolic network (computed from the Recon2.2 network (29)), in addition to their standard deviation (error bar), defined by different threshold T_l values either on normal and cancer samples.

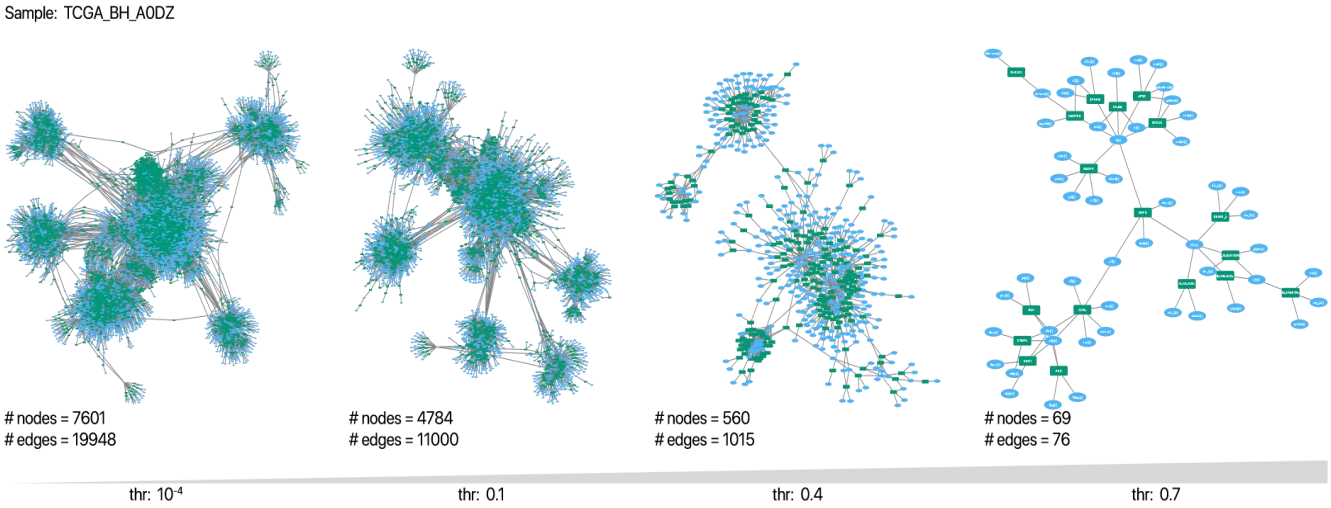


Fig. (2). The giant components of the metabolic network of the cancer sample of patient TCGA BH A0DZ obtained by projecting RNA-seq data on Recon2.2 metabolic network (29), are shown. 4 distinct giant components are shown, obtained with the following relevance thresholds: 10^{-4}, 0.1,0.4,0.7. Networks were drawn via Cytoscape (37).

Since we are interested in exploiting the topological properties of the "*giant component*" of the sample-specific metabolic network (as proposed, *e.g.*, in [33]), we employ a network pruning procedure to select the relevant metabolic reactions. This threshold criterion was employed earlier [34–36]. In detail, a threshold parameter $T_l \in [0,1]$ is used to obtain an unweighted and thresholded adjacency matrix $A^{T_l,S}$, the elements of which are defined as follows:

$$A_{ij}^{T_l,S} = \begin{cases} 1, \ if \ w_{ij}^s \geq T_l \\ 0, \ if \ w_{ij}^s < T_l \end{cases} \forall i,j = 1,\dots,N. \qquad (3)$$

It must be noted that we have focused on the *larger than* option, because we can hypothesize that only significantly active reactions (above the threshold) are responsible for the phenotypic/functional properties of cells. By scanning different values of the threshold, we can then evaluate the impact on the performance of classifiers that take as input certain topological measurements of the resulting giant component (see below), thus identifying an optimal threshold value.

Clearly the threshold parameter determines the size of the giant component, *i.e.*, the largest connected subgraph of the sample-specific metabolic network, which we define as $G^{T_l,S}$ and which includes $N^{T_l,S}$ nodes and $E^{T_l,S}$ edges.

For instance, in Fig. (**1**), one can see how the number of nodes and edges of the giant component of the sample-specific metabolic network (computed from the Recon2.2 network [29, 32]) is generally affected by the choice of distinct thresholds, regarding both cancer and normal samples. In greater detail, on the left side of Fig. (**1a**), smaller thresholds, such as $T_l \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, lead to a larger size of the giant component (on average), while on the right side, larger thresholds, such as $T_l \in \{0.2, 0.3, \dots, 0.7\}$, lead to a radical network reduction, with a threshold $T_l = 0.7$ retaining 127 nodes on average, which represents approximately 1.45% of the total number of nodes of the original metabolic network. As a representative example, the shrinking of the giant component for a specific sample is visually represented in Fig. (**2**).

We also note that this behavior occurs similarly on both cancer and normal samples, even if the size of the giant component of the former ones tends to be slightly smaller. One may speculate that cancer subpopulations engage in a relatively lower number of metabolic functions with respect to normal cells, given that their main objective is "selfish" proliferation. Further investigations are needed to validate this interesting hypothesis.

2.3. Algorithmic Methods for Classification

In general, the choice of adequate network descriptors is crucial for pattern recognition purposes. Typically, the feature extraction is based on well-established network structural measures (see details in Section 2.3.1). The concurrent use of well-known measures such as *degree*, *mean degree*, *clustering coefficient*, *mean hierarchical degree*, *centrality*, and even *spectral measurements*, can identify global properties shared by a large majority of empirical and synthetic networks such as random, small-world, scale-free networks, and geographic networks models [38, 39].

2.3.1. Features Based on Network Structural Measures

Networks measurements falling in various categories (*e.g.*, connectivity-related, distance-related, spectral, degree correlation measures) can be effectively used to characterize the topological properties of real-world networks [38, 40]. In our case, we are interested in determining whether certain topological measurements of the giant component of the sample-specific metabolic network obtained from RNA-seq data projection, and after opportune threshold-based pruning, can be effectively employed as features to classify cancer samples. In particular, we selected the following measures.

Average degree. Among the connectivity-related measurements, we here consider the degree (or connectivity) $k_i^{T_l,s}$ of node *i* of the giant component of sample *s*, given threshold T_l, as the number of neighbors of a node $i^{T_l,s}$ defined by:

$$k_i^{T_l,s} = \sum_{j=1}^{N^{T_l,s}} A_{ij}^{T_l,s}.$$

Accordingly, the average degree of the giant component is defined by Eq. (4), as follows:

$$\langle k^{T_l,s} \rangle = \frac{1}{N^{T_l,s}} \sum_{i=1}^{N^{T_l,s}} k_i^{T_l,s}. \quad (4)$$

Average hierarchical degree. The hierarchical degree $k_i^{T_l,s^h}$ of node i can also be measured considering the connectivity of the neighboring nodes constrained to a hierarchical level h. As an example, in social networks, the hierarchical degree of level 2 of given node i, k_i^2, is the sum of the degrees of the neighbors of its neighbors. Therefore, the mean hierarchical degree of the giant component of a sample-specific metabolic network is given by Eq. (5), as follows:

$$\langle k^{T_l,s^h} \rangle = \frac{1}{N^{T_l,s}} \sum_{i=1}^{N^{T_l,s}} k_i^{T_l,s^h}. \quad (5)$$

Average geodesic path length. A path is defined as the sequence of nodes visited to go from node *i* to *j*. The distance between them is the number of edges within the path, and d_{ij} is defined as the geodesic path, *i.e.*, the smallest path length. When there is no path between *i* and *j*, $d_{ij} = 0$. The average geodesic path length of the giant component of the sample-specific metabolic network is given by:

$$\langle l^{T_l,s} \rangle = \frac{1}{N^{T_l,s}(N^{T_l,s}-1)} \sum_{i \neq j} d_{ij}, \quad (6)$$

where *i* and *j* are two nodes of the giant component and $\frac{1}{N^{T_l,s}(N^{T_l,s}-1)}$ corresponds to a normalization factor, considering a fully connected network (40).

Assortativity. The *assortativity* $\Gamma^{T_l,s}$ [41], *i.e.*, the Pearson correlation coefficient of degree among all pairs of linked nodes *i* and *j* of the giant component, quantifies the tendency of the nodes of a given degree *k* to connect to nodes with a similar degree and, in our case, it is defined as follows:

$$\Gamma^{T_l,s} = \frac{\left(\frac{1}{N^{T_l,s}}\right)\sum_{j>i}\left(k_i^{T_l,s}k_j^{T_l,s}A_{ij}^{T_l,s}\right) - \left[(1/N^{T_l,s})\sum_{j>i}(1/2)\left(k_i^{T_l,s}+k_j^{T_l,s}\right)A_{ij}^{T_l,s}\right]^2}{\left(\frac{1}{N^{T_l,s}}\right)\sum_{j>i}(1/2)\left(k_i^{T_l,s^2}+k_j^{T_l,s^2}\right)A_{ij}^{T_l,s} - \left[(1/N^{T_l,s})\sum_{j>i}(1/2)\left(k_i^{T_l,s}+k_j^{T_l,s}\right)A_{ij}^{T_l,s}\right]^2}, \quad (7)$$

$\Gamma^{T_l,s}$ is a value within the range [−1, 1]. Values closer to 1 indicate a positive correlation (nodes with high degree tend to connect to nodes with high degree), while values closer to −1, indicate a negative correlation (nodes with a high degree tend to connect to nodes with low degree), whereas values close to 0 indicates the absence of linear dependence.

In the following, we will show how to compose a feature vector by considering a set of topological measurements[35, 36, 38]. In this respect, the giant component of a sample-specific metabolic network $G^{T_l,s}$can be characterized by a tuple containing: *(i)* the average degree $\langle k^{T_l,s} \rangle$ (Eq. (4)), *(ii)* the average hierarchical degree of level 2 $\langle k^{T_l,s} \rangle$ (Eq. (5)), *(iii)* the average hierarchical degree of level 3 $\langle k^{T_l,s^3} \rangle$((Eq. (5)), *(iv)* the average geodesic path length $\langle l^{T_l,s} \rangle$ (Eq. (6)) and *(v)* the assortativity $\Gamma^{T_l,s}$ (Eq. (7)). The vector is given by:

$$\vec{\phi}(T_l, s) = [\langle k^{T_l,s} \rangle, \langle k^{T_l,s^2} \rangle, \langle k^{T_l,s^3} \rangle, \langle l^{T_l,s} \rangle, \Gamma^{T_l,s}] \quad (8)$$

We notice that other measures such as the clustering coefficient might be employed as features. However, since in our case the input network is bipartite, there are no triangle neighborhoods and, accordingly, the clustering coefficient would always be 0. Since our framework is designed to be general, one can expect this feature to be relevant in different experimental scenarios, with distinct datasets and alternative representations of reaction graphs (see, *e.g.*, [42–44]).

2.4. Classification Setup

Given any relevance threshold T_l, the feature vectors are extracted for the resulting giant component of each sample *s*, and the classification step can be performed. The main goal of this analysis is to evaluate the classification performance of various classifiers $\mathcal{M}$, *i.e.*, MLPs, SVMs and RFs on the feature vector $\vec{\phi}(T_l, s)$. Furthermore, we tested the same classifiers on a reduced feature vector, including the 5 first principal components of the expression profiles of the 1673 metabolic genes present in the Recon2.2 model [29], in order to provide a comparison on the same number of features employed in our approach.

Table 1. Hyperparameters grid search for the tested classifiers, *i.e.*, MLPs, SVMs and RFs, executed via the scikit-learn Python library. Parameter names are the sklearn arguments of the related functions (default was used for the other parameters).

Method	Function	Parameter	Grid search values
MLP	neuralnetwork.MLPClassifier	solver	[adam, lbfgs]
		hidden_layer_sizes	[(50,),(100,),(50,50)]
		batch_size	[16, 32, 64]
		learning_rate_init	[0.1, 0.01, 0.001]
		learning_rate	[constant, adaptative]
		max_iter	10000
RF	ensemble.RandomForestClassifier	max_depth	[10, 20, 40, None]
		max_features	[auto, sqrt]
		min_samples_leaf	[1, 2, 3]
		min_samples_split	[2, 3, 5]
		n_estimator	[100, 200, 500, 1000]
SVM	svm.SVC	C	[2^{-5}, 2^{-4},..., 2^{12}]
		gamma	[2^{-15} 2^{-14},..., 2^{4}]
		tol	[10^{-3}, 10^{-4}]
		kernel	[rbf, sigmoid, linear]

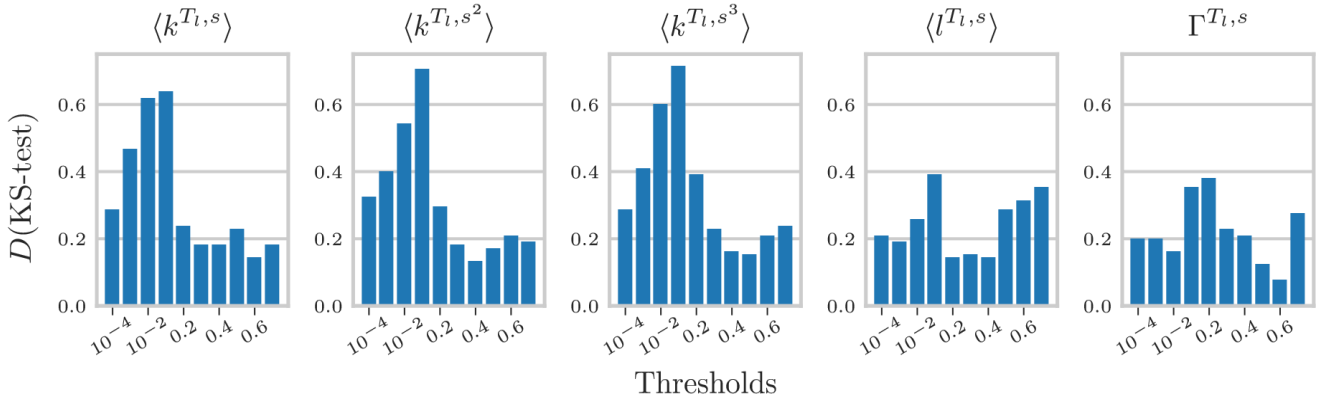


Fig. (3). Kolmogorov-Smirnov statistic (KS-test, (48)) between normal and cancer samples for each threshold and network topological measure: average degree $\langle k^{T_l,s} \rangle$, assortativity $\Gamma^{T_l,s}$ average hierarchical degree of level 2 $\langle k^{T_l,s^2} \rangle$ and 3 $\langle k^{T_l,s^3} \rangle$ and average geodesic path length $\langle l^{T_l,s} \rangle$. The higher the K-S test is, the more the distribution of the network measure is different between normal and cancer samples. The highest values are obtained with $\langle k^{T_l,s} \rangle$, $\langle k^{T_l,s^2} \rangle$, and $\langle k^{T_l,s^3} \rangle$ and thresholds equal to 10^{-2} and 0.1.

In order to prevent over-optimistic results, we performed for each classifier a nested cross-validation as proposed earlier [45] and detailed as follows.

The original dataset, including cancer and normal samples, is split into 5 folds, ensuring the balance between classes. 5-fold outer cross-validation is executed by using: (*i*) one fold as the test set to assess the model performance and (*ii*) 4 folds in an inner 5-fold cross-validation procedure to select the optimal hyperparameters h of the model $\mathcal{M}(h)$ via grid search (see Table **1**). The whole procedure is repeated 3 times to ensure robustness to the results. The performance of all classifiers is assessed on average accuracy, precision and recall with respect to ground-truth labels.

All the experiments described above were performed using the scikit-learn Python library [46].

2.5. Network Datasets

We tested our approach on the breast cancer dataset TCGA-BRCA published earlier (28). We downloaded the dataset via the cBioPortal [47]. This dataset includes the expression profile (RNA Seq V2 RSEM) of biopsies taken from 817 patients. We selected the 105 patients for which the expression profiles of both cancer and normal tissues are provided, for a total of 210 samples used in our analysis.

RNA-seq data were projected on the Recon2.2 metabolic network [29, 32] to obtain a dataset in which a Reaction Activity Score is assigned to each metabolic reaction in each sample (see above). The RASs were then normalized by dividing each reaction score by the maximum value of all samples. Finally, normalized RAS profiles are used to weigh the metabolic network as described above.

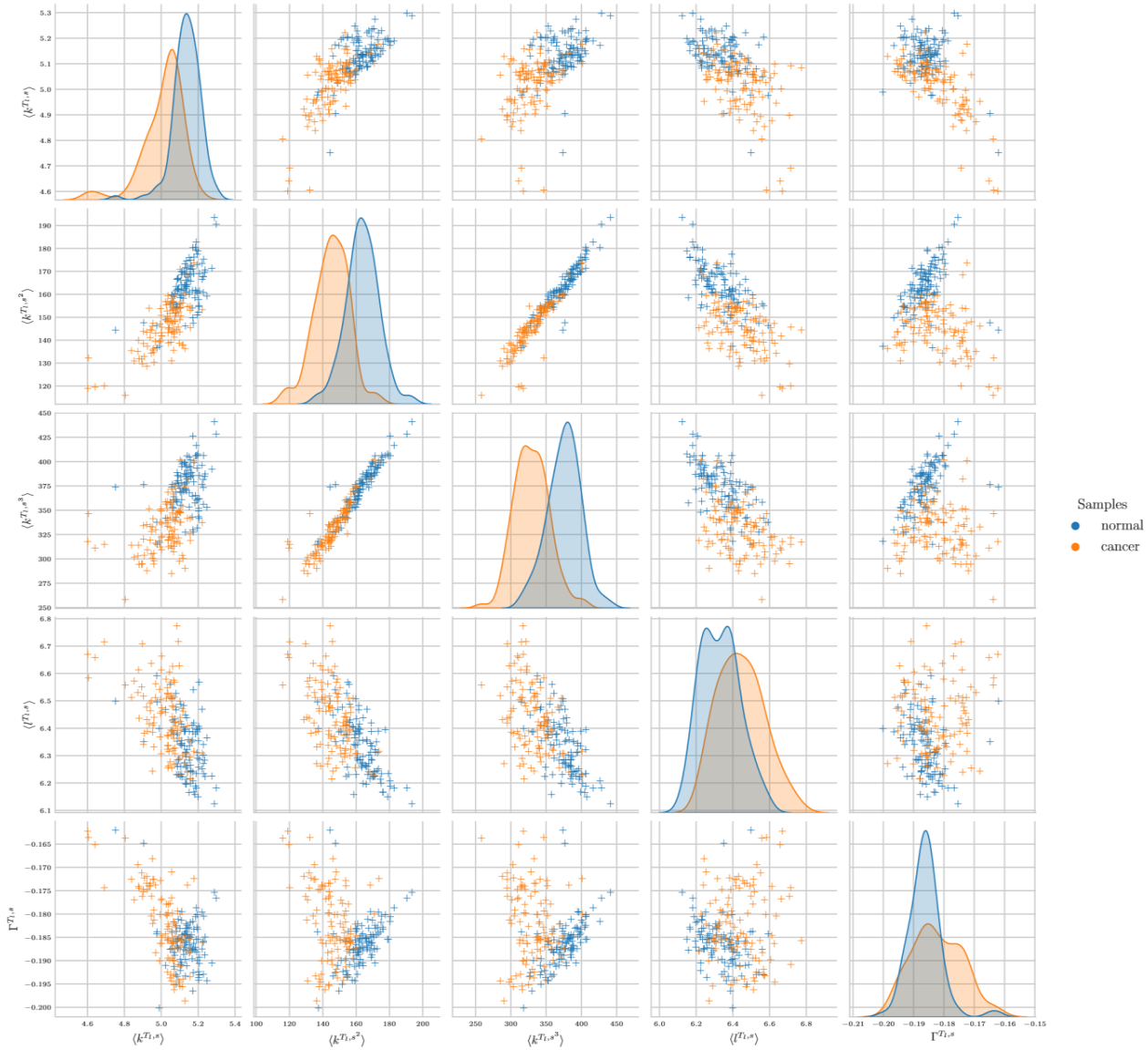


Fig. (4). Projection of cancer and normal samples on the space of topological measure pairs and (on the diagonal) the distribution for each measure and every sample category, for a selected threshold $T_l = 0.1$.

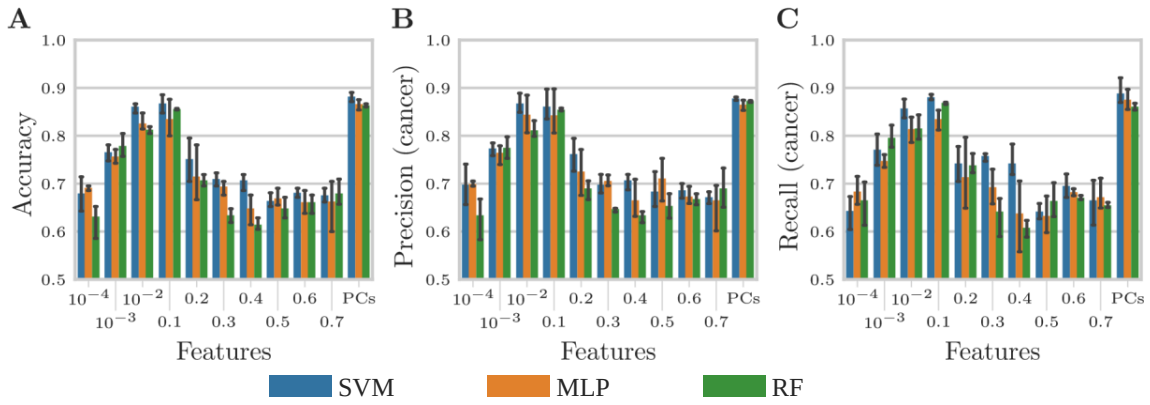


Fig. (5). From left to right: average accuracy (**A**), average precision on cancer samples (**B**) and average recall on cancer samples (**C**) with SVMs, MLPs and RFs. The average is computed on the test sets via a repeated nested cross-validation, for three different seeds, whereas the error bars represent the standard deviation (see Section 2.4 for additional details). The best thresholds are $T_l = 10^{-2}$ and $T_l = 0.1$.

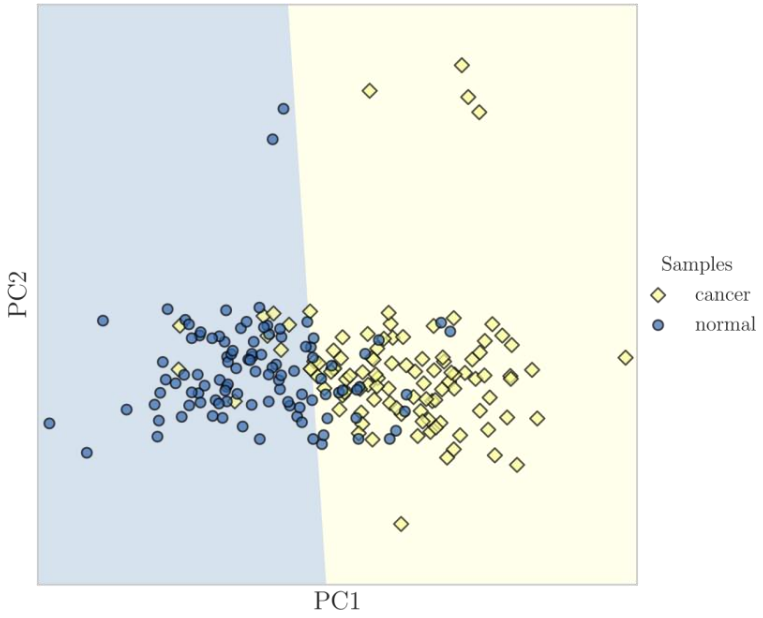


Fig. (6). Decision boundary of the SVM classifier with optimal hyperparameters and threshold $T_l = 0.1$ on the full dataset. The axes correspond to the first two principal components of the full feature vector $\vec{\phi}(T_l, s)$.

3. RESULTS

3.1. RAS Threshold Analysis

A small T_l will result in larger giant components while, in contrast, higher values of T_l will result in smaller giant components. To choose the best classifier, we evaluated the performance obtained by the following distinct threshold values:

$$T_l \in \{10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.2, 0.3, \ldots, 0.7\}. \quad (9)$$

Thus, each feature vector $\vec{\phi}(T_l, s)$, contains the five topological measures defined above as descriptors (see Section 2.3.1).

To test the discrimination power of the feature vectors $\vec{\phi}(T_l, s)$, in Fig. (**3**), we computed the Kolmogorov-Smirnov statistic [48] between normal and cancer samples for each threshold and topological measure. The KS statistic *D* (KS-test) is the distance between the cumulative probability distributions; hence the higher is the value, the more the network measures are different between normal and cancer samples.

As a result, in our dataset, degree statistics, *i.e.*, $\langle k^{T_l,s} \rangle$, $\langle k^{T_l,s^2} \rangle$ and $\langle k^{T_l,s^3} \rangle$, achieve the highest *D* (KS-test), in particular for thresholds equal to 10^{-2} and 0.1. In Fig. (**4**), we plotted the distributions of all pairs of features in $\vec{\phi}(T_l, s)$, for $T_l = 0.1$. In accordance with the results of Fig. (**3**), the degree statistics distributions and, in particular, $\langle k^{T_l,s^2} \rangle$ and $\langle k^{T_l,s^3} \rangle$, have the sharpest difference among normal and cancer samples.

3.2. Classification Performance

The classification performance was assessed for all classifiers (*i.e.*, MLPs, SVMs and RFs) on the feature vector $\vec{\phi}(T_l, s)$, with regard to all relevance thresholds, via the nested cross-validation procedure described above (see Section 2.4). In addition, we employed as benchmark three analogous classifiers (*i.e.*, MLPs, SVMs and RFs), which were provided as input with a feature vector including the 5 first principal components (PCs) of the expression profiles of the 1673 metabolic genes.

In Fig. (**5**), we report the average accuracy, precision and recall for all tested classifiers, with respect to all relevance thresholds, as well as the benchmark classifiers on gene expression PCs, by employing the ground-truth cancer sample labels (the error bars represent the standard deviation).

Interestingly, the best performance is achieved for all classifiers with thresholds in the small range $T_l = 10^{-2}$ and $T_l = 0.1$, and points at the existence of an effective pruning strategy to maintain the "relevant" active metabolic pathways that discriminate cancer from normal samples, while limiting the confounding effects possibly due to noisy observations and biological variability.

More in detail, the best performing classifier is provided by SVMs, which reach an average accuracy of 0.86 and 0.87, a precision of 0.87 and 0.86 and a recall of 0.86 and 0.88, for $T_l = 10^{-2}$ and $T_l = 0.1$, respectively.

Interestingly, such performance is extremely similar to that obtained with SVMs on the vector of gene expression PCs (average accuracy = 0.88, precision = 0.88 and recall = 0.89) and slightly superior to that of MLPs and RFs on the same vector. This result suggests that the information extracted from the few selected topological measures on the giant component of the sample-specific metabolic network is effective in discriminating cancer from normal samples, similarly to benchmark approaches processing gene expression data (5).

Finally, in Fig. (**6**), the decision boundary of the best performing SVM classifier, *i.e.*, obtained with $T_l = 0.1$ and optimal hyperparameters is displayed on the first two PCs of the feature vector $\vec{\phi}(T_l, s)$, from which one can see that the method is able to correctly classify also the outliers of both categories.

CONCLUSION

In this work, we have introduced a new computational framework for the classification of cancer samples, which combines the integration of transcriptomic data and metabolic networks with state-of-the-art machine learning approaches. This task is of practical relevance in many biomedical contexts and might pave the way for the development of automated strategies for experimental hypothesis generation. In particular, the introduction of our framework contributes to the emerging field of approaches combining sample-specific metabolic modeling with machine learning to classify cancer samples and/or to predict drug response, as recently reviewed [49, 50].

More in detail, we here proved that the information on the metabolic activity of single samples, derived via integration of highly accessible RNA-seq data, can be effectively used to classify healthy and pathological states, a result that appears to be robust when the original networks are significantly pruned via a relevance threshold. All in all, this result would suggest that the useful information to determine possibly aberrant states in a given sample can be derived from the high-level (topological) properties of a relatively limited number of active processes. The identification and characterization of such processes deserve further investigation.

Regarding our machine learning approach, we here relied on classical topological measures, such as degree, hierarchical degrees, average geodesic path length and assortativity, to encode the structural information of the metabolic network. Additional experiments may employ recent graph representation learning techniques [51, 52], including graph kernels [53] and convolutional neural networks on graphs [54], to automatically extract a low-dimensional feature vector of the input network.

We finally remark that extensions of the framework are currently ongoing to test its applicability to more complex scenarios, involving, for instance, multiclass and multi-label classification with respect to cancer subtypes and risk categories.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No animals/humans were used for studies that are the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The datasets generated and analyzed for this study can be found at this link: https://github.com/BIMIB-DISCo/METNET-CLASSIFICATION.

FUNDING

Financial support from the Italian Ministry of University and Research (MIUR) through grant “Dipartimenti di Eccellenza 2017” to University of Milano-Bicocca, Department of Biotechnology and Biosciences is acknowledged. Partial support was also provided by the CRUK/AECC/AIRC Accelerator Award #22790, “Single-cell Cancer Evolution in the Clinic”. J.M. is grateful for the support from the National Council for Scientific and Technological Development (CNPq grant #155957/2018-0) and São Paulo Research Foundation (FAPESP grant #2020/03514-9).

O. M. B. acknowledges support from CNPq (Grant #307897/2018-4) and FAPESP (grant #2014/08026-1 and 2016/18809-9).

This work was also partially supported by a Bicocca 2020 Starting Grant to F.A.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] Kourou, K.; Exarchos, T.P.; Exarchos, K.P.; Karamouzis, M.V.; Fotiadis, D.I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.*, **2014**, *13*, 8-17. [http: //dx.doi.org/10.1016/j.csbj.2014.11.005] [PMID: 25750696]

[2] Furey, T.S.; Cristianini, N.; Duffy, N.; Bednarski, D.W.; Schummer, M.; Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **2000**, *16*(10), 906-914. [http: //dx.doi.org/10.1093/bioinformatics/16.10.906] [PMID: 11120680]

[3] Sotiriou, C.; Neo, S-Y.; McShane, L.M.; Korn, E.L.; Long, P.M.; Jazaeri, A.; Martiat, P.; Fox, S.B.; Harris, A.L.; Liu, E.T. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci. USA*, **2003**, *100*(18), 10393-10398. [http: //dx.doi.org/10.1073/pnas.1732912100] [PMID: 12917485]

[4] Lu, J.; Getz, G.; Miska, E.A.; Alvarez-Saavedra, E.; Lamb, J.; Peck, D.; Sweet-Cordero, A.; Ebert, B.L.; Mak, R.H.; Ferrando, A.A.; Downing, J.R.; Jacks, T.; Horvitz, H.R.; Golub, T.R. MicroRNA expression profiles classify human cancers. *Nature*, **2005**, *435*(7043), 834-838. [http: //dx.doi.org/10.1038/nature03702] [PMID: 15944708]

[5] CP de Souto, M.; G Costa, I.; SA de Araujo, D.; B Ludermir, T.; Schliep, A. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, **2008**, *9*(1), 497. [http: //dx.doi.org/10.1186/1471-2105-9-497]

[6] Vanneschi, L.; Farinaccio, A.; Mauri, G.; Antoniotti, M.; Provero, P.; Giacobini, M. A comparison of machine learning techniques for survival prediction in breast cancer. *BioData Min.*, **2011**, *4*(1), 12. [http: //dx.doi.org/10.1186/1756-0381-4-12] [PMID: 21569330]

[7] Curtis, C.; Shah, S.P.; Chin, S.F.; Turashvili, G.; Rueda, O.M.; Dunning, M.J.; Speed, D.; Lynch, A.G.; Samarajiwa, S.; Yuan, Y.; Gräf, S.; Ha, G.; Haffari, G.; Bashashati, A.; Russell, R.; McKinney, S.; Langerød, A.; Green, A.; Provenzano, E.; Wishart, G.; Pinder, S.; Watson, P.; Markowetz, F.; Murphy, L.; Ellis, I.; Purushotham, A.; Børresen-Dale, A.L.; Brenton, J.D.; Tavaré, S.; Caldas, C.; Aparicio, S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **2012**, *486*(7403), 346-352. [http: //dx.doi.org/10.1038/nature10983] [PMID: 22522925]

[8] Caravagna, G.; Graudenzi, A.; Ramazzotti, D.; Sanz-Pamplona, R.; De Sano, L.; Mauri, G.; Moreno, V.; Antoniotti, M.; Mishra, B. Algorithmic methods to infer the evolutionary trajectories in cancer

progression. *Proc. Natl. Acad. Sci. USA*, **2016**, *113*(28), E4025-E4034.
[http: //dx.doi.org/10.1073/pnas.1520213113] [PMID: 27357673]

[9] Caravagna, G.; Giarratano, Y.; Ramazzotti, D.; Tomlinson, I.; Graham, T.A.; Sanguinetti, G.; Sottoriva, A. Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nat. Methods*, **2018**, *15*(9), 707-714.
[http: //dx.doi.org/10.1038/s41592-018-0108-x] [PMID: 30171232]

[10] Hofree, M.; Shen, J.P.; Carter, H.; Gross, A.; Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods*, **2013**, *10*(11), 1108-1115.
[http: //dx.doi.org/10.1038/nmeth.2651] [PMID: 24037242]

[11] Michael, L.G.; Joseph, E.L.; William, T.B.; Jong, W.K; Quanli, W.; Matthew, D.C; Michael, B.D; Michael, K.; Bernard Mathey, P.; Anil, P. A pathwaybased classification of human breast cancer. *Proceedings of the National Academy of Sciences*, , pp. 6994-6999.**2010**]

[12] Graudenzi, A.; Cava, C.; Bertoli, G.; Fromm, B.; Flatmark, K.; Mauri, G.; Castiglioni, I. Pathway-based classification of breast cancer subtypes. *Front. Biosci.*, **2017**, *22*, 1697-1712.
[http: //dx.doi.org/10.2741/4566] [PMID: 28410140]

[13] Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell*, **2011**, *144*(5), 646-674.
[http: //dx.doi.org/10.1016/j.cell.2011.02.013] [PMID: 21376230]

[14] Cantor, J.R.; Sabatini, D.M. Cancer cell metabolism: one hallmark, many faces. *Cancer Discov.*, **2012**, *2*(10), 881-898.
[http: //dx.doi.org/10.1158/2159-8290.CD-12-0345] [PMID: 23009760]

[15] Ward, P.S.; Thompson, C.B. Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. *Cancer Cell*, **2012**, *21*(3), 297-308.
[http: //dx.doi.org/10.1016/j.ccr.2012.02.014] [PMID: 22439925]

[16] Tomita, M.; Kami, K. Cancer. Systems biology, metabolomics, and cancer metabolism. *Science*, **2012**, *336*(6084), 990-991.
[http: //dx.doi.org/10.1126/science.1223066] [PMID: 22628644]

[17] Beverly, A Targeting cancer metabolism. **2012**.

[18] Hyduke, D.R.; Lewis, N.E.; Palsson, B.Ø. Analysis of omics data with genome-scale models of metabolism. *Mol. Biosyst.*, **2013**, *9*(2), 167-174.
[http: //dx.doi.org/10.1039/C2MB25453K] [PMID: 23247105]

[19] Lewis, N.E.; Abdel-Haleem, A.M. The evolution of genome-scale models of cancer metabolism. *Front. Physiol.*, **2013**, *4*, 237.
[http: //dx.doi.org/10.3389/fphys.2013.00237] [PMID: 24027532]

[20] Orth, J.D.; Thiele, I.; Palsson, B.Ø. What is flux balance analysis? *Nat. Biotechnol.*, **2010**, *28*(3), 245-248.
[http: //dx.doi.org/10.1038/nbt.1614] [PMID: 20212490]

[21] Machado, D.; Herrgård, M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLOS Comput. Biol.*, **2014**, *10*(4), e1003580.
[http: //dx.doi.org/10.1371/journal.pcbi.1003580] [PMID: 24762745]

[22] Jamialahmadi, O.; Hashemi-Najafabadi, S.; Motamedian, E.; Romeo, S.; Bagheri, F. A benchmark-driven approach to reconstruct metabolic networks for studying cancer metabolism. *PLOS Comput. Biol.*, **2019**, *15*(4), e1006936.
[http: //dx.doi.org/10.1371/journal.pcbi.1006936] [PMID: 31009458]

[23] Damiani, C.; Di Filippo, M.; Pescini, D.; Maspero, D.; Colombo, R.; Mauri, G. popFBA: tackling intratumour heterogeneity with Flux Balance Analysis. *Bioinformatics*, **2017**, *33*(14), i311-i318.
[http: //dx.doi.org/10.1093/bioinformatics/btx251] [PMID: 28881985]

[24] Damiani, C.; Maspero, D.; Di Filippo, M.; Colombo, R.; Pescini, D.; Graudenzi, A.; Westerhoff, H.V.; Alberghina, L.; Vanoni, M.; Mauri, G. Integration of single-cell RNA-seq data into population models to characterize cancer metabolism. *PLOS Comput. Biol.*, **2019**, *15*(2), e1006733.
[http: //dx.doi.org/10.1371/journal.pcbi.1006733] [PMID: 30818329]

[25] Damiani, C.; Gaglio, D.; Sacco, E.; Alberghina, L.; Vanoni, M. Systems metabolomics: from metabolomic snapshots to design principles. *Curr. Opin. Biotechnol.*, **2020**, *63*, 190-199. Nanobiotechnology Systems Biology.
[http: //dx.doi.org/10.1016/j.copbio.2020.02.013] [PMID: 32278263]

[26] Graudenzi, A.; Maspero, D.; Di Filippo, M.; Gnugnoli, M.; Isella, C.; Mauri, G.; Medico, E.; Antoniotti, M.; Damiani, C. Integration of transcriptomic data and metabolic networks in cancer samples reveals highly significant prognostic power. *J. Biomed. Inform.*, **2018**, *87*, 37-49.
[http: //dx.doi.org/10.1016/j.jbi.2018.09.010] [PMID: 30244122]

[27] Damiani, C.; Rovida, L.; Maspero, D.; Sala, I.; Rosato, L.; Di Filippo, M.; Pescini, D.; Graudenzi, A.; Antoniotti, M.; Mauri, G. MaREA4Galaxy: Metabolic reaction enrichment analysis and visualization of RNA-seq data within Galaxy. *Comput. Struct. Biotechnol. J.*, **2020**, *18*, 993-999.
[http: //dx.doi.org/10.1016/j.csbj.2020.04.008] [PMID: 32373287]

[28] Ciriello, G.; Gatza, M.L.; Beck, A.H.; Wilkerson, M.D.; Rhie, S.K.; Pastore, A.; Zhang, H.; McLellan, M.; Yau, C.; Kandoth, C.; Bowlby, R.; Shen, H.; Hayat, S.; Fieldhouse, R.; Lester, S.C.; Tse, G.M.; Factor, R.E.; Collins, L.C.; Allison, K.H.; Chen, Y.Y.; Jensen, K.; Johnson, N.B.; Oesterreich, S.; Mills, G.B.; Cherniack, A.D.; Robertson, G.; Benz, C.; Sander, C.; Laird, P.W.; Hoadley, K.A.; King, T.A.; Perou, C.M. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, **2015**, *163*(2), 506-519.
[http: //dx.doi.org/10.1016/j.cell.2015.09.033] [PMID: 26451490]

[29] Swainston, N.; Smallbone, K.; Hefzi, H.; Dobson, P.D.; Brewer, J.; Hanscho, M.; Zielinski, D.C.; Ang, K.S.; Gardiner, N.J.; Gutierrez, J.M.; Kyriakopoulos, S.; Lakshmanan, M.; Li, S.; Liu, J.K.; Martínez, V.S.; Orellana, C.A.; Quek, L.E.; Thomas, A.; Zanghellini, J.; Borth, N.; Lee, D.Y.; Nielsen, L.K.; Kell, D.B.; Lewis, N.E.; Mendes, P. Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics*, **2016**, *12*(7), 109.
[http: //dx.doi.org/10.1007/s11306-016-1051-4] [PMID: 27358602]

[30] Cazzaniga, P.; Damiani, C.; Besozzi, D.; Colombo, R.; Nobile, M.S.; Gaglio, D.; Pescini, D.; Molinari, S.; Mauri, G.; Alberghina, L.; Vanoni, M. Computational strategies for a system-level understanding of metabolism. *Metabolites*, **2014**, *4*(4), 1034-1087.
[http: //dx.doi.org/10.3390/metabo4041034] [PMID: 25427076]

[31] Mardinoglu, A.; Agren, R.; Kampf, C.; Asplund, A.; Uhlen, M.; Nielsen, J. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat. Commun.*, **2014**, *5*, 3083.
[http: //dx.doi.org/10.1038/ncomms4083] [PMID: 24419221]

[32] Thiele, I.; Swainston, N.; Fleming, R.M.; Hoppe, A.; Sahoo, S.; Aurich, M.K.; Haraldsdottir, H.; Mo, M.L.; Rolfsson, O.; Stobbe, M.D.; Thorleifsson, S.G.; Agren, R.; Bölling, C.; Bordel, S.; Chavali, A.K.; Dobson, P.; Dunn, W.B.; Endler, L.; Hala, D.; Hucka, M.; Hull, D.; Jameson, D.; Jamshidi, N.; Jonsson, J.J.; Juty, N.; Keating, S.; Nookaew, I.; Le Novère, N.; Malys, N.; Mazein, A.; Papin, J.A.; Price, N.D.; Selkov, E., Sr; Sigurdsson, M.I.; Simeonidis, E.; Sonnenschein, N.; Smallbone, K.; Sorokin, A.; van Beek, J.H.; Weichart, D.; Goryanin, I.; Nielsen, J.; Westerhoff, H.V.; Kell, D.B.; Mendes, P.; Palsson, B.Ø. A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.*, **2013**, *31*(5), 419-425.
[http: //dx.doi.org/10.1038/nbt.2488] [PMID: 23455439]

[33] Ma, H-W.; Zeng, A-P. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, **2003**, *19*(11), 1423-1430.
[http: //dx.doi.org/10.1093/bioinformatics/btg177] [PMID: 12874056]

[34] Backes, A.R.; Casanova, D.; Bruno, O.M. A complex network-based approach for boundary shape analysis. *Pattern Recognit.*, **2009**, *42*(1), 54-67.
[http: //dx.doi.org/10.1016/j.patcog.2008.07.006]

[35] Miranda, G.H.B.; Machicao, J.; Bruno, O.M. An optimized shape descriptor based on structural properties of networks. *Digit. Signal Process.*, **2018**, *82*, 216-229.
[http: //dx.doi.org/10.1016/j.dsp.2018.06.010]

[36] Machicao, J.; Filho, H.A.; Lahr, D.J.G.; Buckeridge, M.; Bruno, O.M. Topological assessment of metabolic networks reveals evolutionary information. *Sci. Rep.*, **2018**, *8*(1), 15918.
[http: //dx.doi.org/10.1038/s41598-018-34163-7] [PMID: 30374088]

[37] Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **2003**, *13*(11), 2498-2504. [http: //dx.doi.org/10.1101/gr.1239303] [PMID: 14597658]

[38] Costa, L.D.F.; Boas, P.R.V.; Silva, F.N.; Rodrigues, F.A. A pattern recognition approach to complex networks. *J. Stat. Mech.*, **2010**, *2010*(11), P11015. [http: //dx.doi.org/10.1088/1742-5468/2010/11/P11015]

[39] Banerjee, A.; Jost, J. Spectral plot properties: Towards a qualitative classification of networks. *NHM*, **2008**, *3*(2), 395-411. [http: //dx.doi.org/10.3934/nhm.2008.3.395]

[40] Costa, L da F.; Francisco, A. Rodrigues, Gonzalo Travieso, and Paulino Ribeiro Villas Boas. Characterization of complex networks: A survey of measurements. *Adv. Phys.*, **2007**, *56*(1), 167-242. [http: //dx.doi.org/10.1080/00018730601170527]

[41] Newman, M.E. Assortative mixing in networks. *Phys. Rev. Lett.*, **2002**, *89*(20), 208701. [http: //dx.doi.org/10.1103/PhysRevLett.89.208701] [PMID: 12443515]

[42] Filisetti, A.; Graudenzi, A.; Serra, R.; Villani, M.; De Lucrezia, D.; Rudolf, M. Füchslin, Stuart A Kauffman, Norman Packard, and Irene Poli. A stochastic model of the emergence of autocatalytic cycles. *J. Syst. Chem.*, **2011**, *2*(1), 2. [http: //dx.doi.org/10.1186/1759-2208-2-2]

[43] Filisetti, A.; Graudenzi, A.; Serra, R.; Villani, M.; Füchslin, R.M.; Packard, N.; Kauffman, S.A.; Poli, I. A stochastic model of autocatalytic reaction networks. *Theory Biosci.*, **2012**, *131*(2), 85-93. [http: //dx.doi.org/10.1007/s12064-011-0136-x] [PMID: 21979857]

[44] Serra, R.; Filisetti, A.; Villani, M.; Graudenzi, A.; Damiani, C.; Panini, T. A stochastic model of catalytic reaction networks in protocells. *Nat. Comput.*, **2014**, *13*(3), 367-377. [http: //dx.doi.org/10.1007/s11047-014-9445-6]

[45] Cawley, G.C.; Talbot, N.L.C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.*, **2010**, *11*, 2079-2107.

[46] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **2011**, *12*, 2825-2830.

[47] Cerami, Ethan; Gao, Jianjiong; Dogrusoz, Ugur The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. **2012**.

[48] Hodges, J.L. The significance probability of the smirnov two-sample test. *Ark. Mat.*, **1958**, *3*, 469-486. [http: //dx.doi.org/10.1007/BF02589501]

[49] Pacheco, M.P.; Bintener, T.; Sauter, T. Towards the network-based prediction of repurposed drugs using patient-specific metabolic models. *EBioMedicine*, **2019**, *43*, 26-27. [http: //dx.doi.org/10.1016/j.ebiom.2019.04.017] [PMID: 30979684]

[50] Zampieri, G.; Vijayakumar, S.; Yaneske, E.; Angione, C. Machine and deep learning meet genome-scale metabolic modeling. *PLOS Comput. Biol.*, **2019**, *15*(7), e1007084. [http: //dx.doi.org/10.1371/journal.pcbi.1007084] [PMID: 31295267]

[51] Cai, H.Y.; Zheng, V.W.; Chang, K.C-C. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Trans. Knowl. Data Eng.*, **2018**, *30*(9), 1616-1637. [http: //dx.doi.org/10.1109/TKDE.2018.2807452]

[52] Hamilton, W.L.; Ying, R.; Leskovec, J. Representation learning on graphs: Methods and applications. *IEEE Data Eng. Bull.*, **2017**, *40*(3), 52-74.

[53] Kriege, N.M.; Johansson, F.D.; Morris, C. A survey on graph kernels. *Applied Network Science*, **2020**, *5*(1), 6. [http: //dx.doi.org/10.1007/s41109-019-0195-3]

[54] Niepert, M.; Ahmed, M.; Kutzkov, K. Learning convolutional neural networks for graphs. *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016*, New York City, NY, USAJune 19-24, 2016*volume 48*, pp. 2014-2023.**2016**.

3.2 Computational methods to exploit mutational profiles

The mutational profiles generated from sequencing experiments described in section 2.3 can be employed for the investigation of a number of aspects related to multicellular heterogeneity. In this work, I mainly focused on two main topics:

- Inference of phylogenomic models (section 3.2.1), which resulted in:

  (*i*) a framework for the inference of longitudinal single-cell cancer evolution (see paper P#5).

  (*ii*) a comprehensive framework for the characterization of viral evolution from deep sequencing data of viral samples (see paper P#6).

  (*iii*) a new approach to improve the inference of evolution models applying a consensus approach (section 3.2.1.1).

- Decomposition of mutational profiles of viral samples into mutational signatures (section 3.2.2, paper P#7).

All topics are detailed in the following, alongside the selected related articles.

3.2.1 Inference of phylogenomic models

Phylogenetics. The primary scope of a phylogenetic analysis is to describe, via a graph, the pattern of descent among a group of individuals (e.g., species or cells), as inferred from similarities and differences in their morphological or genetic characteristics.

Standard phylogenetic analyses typically represent different evolutionary processes via binary trees, where the *root* is the lowest common ancestor of a given population, the *leaves* are the observed individuals or samples, and the internal nodes are the inferred coalescent events (e.g., unsampled ancestors).

In such representation, the length of the edges could be related to various quantities which underlay the *evolutionary distance* between ancestors and children, e.g., the number of different observable phenotypic traits. A renowned example of is that of Darwin's *tree of life* [109, 112], which represents the evolution of life and describes the temporal relationships between living and extinct organisms.

There exists a plethora of methods for the inference of phylogenetic models from data, which are reviewed in [12, 13] by Professor Felsenstein.

From phenotypic to genotypic phylogenetics. While classical phylogenetic studies exploit the phenotypic traits of individuals, recent studies can take advantage of the major improvement of sequencing technologies described in Section 2.1.

To this extent, phylogeny methods based on sequencing data shed new light on the different properties of the evolving system [17]. We can exploit such methods to (i) understand how the genetic sequences evolve (e.g., conserved regions or positively selected genomic variants) [], (ii) discover the most probable ancestral genomes, (iii) estimate the timings of evolutionary events, and (iv) test different hypotheses on evolutionary models [18]. Please see [58] for a comprehensive review of the most widely applied phylogenetic inference methods.

In fact, for many biological systems, such as cancer subpopulations and viral quasispecies, mutations are inherited and accumulated during the evolution of the system (despite noteworthy exceptions that are discussed later). Accordingly, the mutational profiles generated via either bulk or single-cell sequencing experiments, and described in Section 2.3 can be employed as input for phylogenetic inference. In this regard, I have been focusing on two distinct, yet analogous, biological systems, i.e., cancer and viral evolution.

Cancer phylogenetics. In its seminal work, Nowell states that cancer results from an evolutionary process that leads to the emergence, competition and (positive/negative) selection of genetically distinct subpopulations of cells called *clones* [10]. In the process, cancer subpopulations can acquire a relatively small number of phenotypic traits called hallmarks [46]. However, the combination of somatic mutations required to reach such hallmarks is possibly vast, tumor- and patient-specific, and, above all, still largely undeciphered.

Thus, one of the key challenges in cancer research is the implementation of effective computational strategies to exploit somatic mutations to reconstruct the evolution of single tumors (note that efforts to infer population-level models have been achieved, e.g., in [91, 137], but are not scope of the current work).

A fine characterization of the genetic evolutionary history of a tumor can serve, for instance, to (the list is not exhaustive): (*i*) assess the type of evolution, e.g., linear, branching, neutral, or punctuated [120]; (*ii*) assess the effect of therapies [259]; (*iii*) evaluate the fitness pressure of clones (e.g., selection coefficients or clonal prevalence variation) [251]; (*iv*) assess the presence of preferential temporal ordering (e.g, selective advantage relations)[34, 179, 228, 234]; (*v*) identify genetic events responsible for hallmark acquisition (e.g., metastasis) or prognostic bio-markers [247]; (*vi*) date the key cancer evolution events [213]; (*vii*) investigate the genotype-phenotype relation; (*viii*) predict the possible future evolution of the tumor [79, 176].

For these reasons, many effort have been recently carried out to reconstruct models of cancer evolution from available data. In detail, we focused our work on the design of computational methods: (*i*) to process single-cell mutational profiles, (*ii*) generated from longitudinal sequencing experiments, (*iii*) with the goal of reconstructing clonal

trees (see below). The rationale is that, given the high levels of noise and missing data of single-cell mutational profiles, it is fundamental to employ robust and expressive statistical frameworks.

Notice that the acquisition of cancer hallmarks is supposed to driven by the accumulation of relevant genomic mutations (i.e., driver events), which produce an increase in the cell's fitness, typically inducing a relative expansion of the relative offsprings. Accordingly, in the so-called *clonal trees* all nodes are clones and edges represent ancestral relations characterized by specific driver mutations [96]. Clonal trees are distinct from standard phylogenetic trees, in which only the leaves represent single cells.

Longitudinal Analisys of Cancer Evolution. In the paper P#5, we propose a new statistical framework that starts from somatic mutation profiles of cells sampled at multiple time points to retrieve the clonal dynamics during the disease progression. The method aims at solving a Boolean matrix factorization problem optimising a weighted-log-likelihood function via Markov Chain Monte Carlo search schema to account for uncertainty in the data (i.e., false positives (FP), false negatives (FN), and missing data in mutational profiles). Our method relies on the Infinite Sites Assumption (ISA), which assumes that mutations are only accumulated once and are never lost during the cancer evolution. With this assumption, we can add perfect phylogeny constraints to reduce the search space and correct the inconsistencies that may be present in the data [113, 114].

The method returns a longitudinal clonal tree, including information about the change in prevalence of the clonal subpopulations at different time points. Importantly, if the samples are taken before, during and after a therapy administration, our method allows one to compare the impact of a given drug among clones through time points. We finally showed that LACE is able to handle datasets including non-relevant mutations (i.e., passengers), by detecting co-occurrence mutation patterns.

Integration of omics data types. Importantly, thanks to the possibility of calling variant from scRNA-seq data and the robustness of the statistical framework, in the article we show an explicit mapping between the genotype and the gene expression (phenotype) of single cells. In particular, our case study involved BRAF-mutant melanoma PDX datasets generated before, during, and after a therapy administration. LACE provided a high-resolution picture of the evolutionary history of this tumor, including clones with different prevalence among time points. The differential analyses of their expression profile allowed us determine how the drug affected their proliferative behaviour in each time point. This result, highlight that an explicit mapping between the clonal evolution and the phenotypic properties with single-cell resolution proved to be a powerful and expressive approach to decipher intra-tumor heterogeneity on multiple scales delivering

experimental hypotheses with translational relevance.

Journal of Computational Science 58 (2022) 101523

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

LACE: Inference of cancer evolution models from longitudinal single-cell sequencing data

Daniele Ramazzotti [a,1], Fabrizio Angaroni [b,1], Davide Maspero [b,c,d,1], Gianluca Ascolani [b], Isabella Castiglioni [e,d], Rocco Piazza [a,f], Marco Antoniotti [b,f], Alex Graudenzi [d,f,*]

[a] Department of Medicine and Surgery, University of Milan-Bicocca, Monza, Italy
[b] Department of Informatics, Systems and Communication, University of Milan-Bicocca, Milan, Italy
[c] Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[d] Inst. of Molecular Bioimaging and Physiology, Consiglio Nazionale delle Ricerche (IBFM-CNR), Segrate, Milan, Italy
[e] Department of Physics "Giuseppe Occhialini", University of Milan-Bicocca, Milan, Italy
[f] Bicocca Bioinformatics Biostatistics and Bioimaging Centre – B4, Milan, Italy

ARTICLE INFO

Keywords:
Cancer evolution
Single-cell sequencing
Longitudinal data
Phylogenomics

ABSTRACT

The rise of longitudinal single-cell sequencing experiments on patient-derived cell cultures, xenografts and organoids is opening new opportunities to track cancer evolution, assess the efficacy of therapies and identify resistant subclones.

We introduce LACE, the first algorithmic framework that processes single-cell mutational profiles from samples collected at different time points to reconstruct longitudinal models of cancer evolution. The approach maximizes a weighted likelihood function computed on longitudinal data points to solve a Boolean matrix factorization problem, via Markov chain Monte Carlo sampling.

On simulations, LACE outperforms state-of-the-art methods for both bulk and single-cell sequencing data with respect to the reconstruction of the ground-truth clonal phylogeny and dynamics, also in conditions of unbalanced datasets, significant rates of sequencing errors and sampling limitations. As the results are robust with respect to data-specific errors, LACE is effective with mutational profiles generated by calling variants from (full-length) scRNA-seq data, and this allows one to investigate the relation between genomic and phenotypic evolution of tumors at the single-cell level.

Here, we apply LACE to a longitudinal scRNA-seq dataset of patient-derived xenografts of $BRAF^{V600E/K}$ mutant melanomas, dissecting the impact of BRAF/MEK-inhibition on clonal evolution, also in terms of clone-specific gene expression dynamics. Furthermore, the analysis of breast cancer PDXs from longitudinal targeted scDNA-sequencing experiments delivers a high-resolution temporal characterization of intra-tumor heterogeneity.

1. Introduction

The advent of single-cell omics measurements has fueled an exceptional growth of high-resolution quantitative studies on complex biological phenomena [1–3]. This is extremely relevant in the analysis of cancer evolution and in the characterization of intra-tumor heterogeneity (ITH), which is a major cause of drug resistance and relapse [4–8].

In recent years, a large array of targeted cancer therapies has been developed, such as, e.g., kinase inhibitors, monoclonal antibodies and, more recently, immunomodulatory agents and clinical grade CAR-T [9]. However, the availability of such personalized therapies requires comparably advanced diagnostic and monitoring tools to study the response of cancer cells under the selective pressure generated by the treatment. In this respect, a highly-awaited major experimental advancement is provided by *longitudinal* single-cell sequencing experiments on samples taken at different time points from the same tumor, or from patient-derived cell cultures, xenografts or organoids [10–12]. In most cases, single-cell transcriptomes are sequenced, e.g., via scRNA-seq experiments [13,14], yet studies employing data from whole-genome/exome and targeted scDNA-seq experiments have been successfully proposed [15,16]. Furthermore, approaches for the coupled

* Corresponding author at: Inst. of Molecular Bioimaging and Physiology, Consiglio Nazionale delle Ricerche (IBFM-CNR), Segrate, Milan, Italy.
E-mail address: alex.graudenzi@ibfm.cnr.it (A. Graudenzi).
[1] Equal contributors.

https://doi.org/10.1016/j.jocs.2021.101523
Received 20 May 2021; Received in revised form 22 October 2021; Accepted 2 December 2021
Available online 16 December 2021
1877-7503/© 2021 Elsevier B.V. All rights reserved.

analysis of genomes and transcriptomes at the single-cell level have been recently introduced [17,18].

Longitudinal single-cell sequencing data might allow one to track cancer evolution at unprecedented resolution, as they can be employed – in principle – to call somatic variants in each single cell of a tumor sample, at any given time point. Accordingly, this may allow one to draw a high-resolution picture of evolutionary history of that tumor, as well as to measure the effect of any possible external intervention, such as a therapeutic strategy. Yet, currently no technique can explicitly process longitudinal single-cell mutational profiles.

On the one hand, in fact, current approaches that process single-cell data and extend phylogenetic methods by handling data-specific errors [19–23] require the implementation of ad hoc modifications to handle longitudinal data, as proposed for instance in [24], and might struggle in both the reconstruction of the mutational tree and the assessment of the clonal composition and variation in time, especially in certain experimental scenarios (see the Results section). On the other hand, even though methods for longitudinal *bulk* sequencing data are starting to produce noteworthy results [25–27], they usually require complex computational strategies to deconvolve the signal coming from intermixed cell subpopulations. Furthermore, there is an ongoing debate whether multi-sample trees from bulk samples are indeed phylogenies or, conversely, if they might lead to erroneous evolutionary inferences [28].

We here propose LACE (Longitudinal Analysis of Cancer Evolution), a new computational method for the reconstruction of *longitudinal clonal trees* of tumor evolution from longitudinal single-cell somatic mutation profiles of tumor samples (see Fig. 1).

In the inferred tree, each vertex correspond to a *genotype*, which identifies a subset of single cells displaying the same set of somatic variants. Edges in the tree model both *parental* relations among genotypes and *persistence* relations through time (see Methods). Notice that genotypes represent *(sub)clones* if the considered variants are *drivers*. Alternatively, if knowledge about drivers is not available, genotypes can be grouped into candidate (sub)clones according to the co-occurrence similarity of mutations across single cells. LACE then estimates the prevalence of each candidate (sub)clone at each time point, hence allowing one to identify (sub)clones that emerge, expand, shrink or disappear during the history of the tumor, e.g., as a consequence of a therapy or a selection sweep. A formal definition of the single-cell longitudinal clonal tree returned by LACE is provided in the Supplementary Information (SI).

Our method manages noise in single-cell sequencing data by estimating false positive and false negative rates – which might be different in distinct time points – and returns the longitudinal clonal tree that maximizes a weighted likelihood function computed on all data points. In this way, our method is able to handle possible differences in quality, sample size and error rates of the experiments performed at distinct time points. It is well known, in fact, that extremely different error rates are observed in single-cell experiments performed via distinct experimental platforms, and that even experiments made with the same platform might display highly heterogeneous noise levels [29]. The search is then performed by solving a Boolean matrix factorization problem [30,31] via Markov Chain Monte Carlo (MCMC), which ensures high scalability and convergence (with a large number of samplings).

The robustness of our approach allows its application to the highly-available (full-length) scRNA-seq data – which are usually employed to characterize the gene expression patterns of single-cells in a variety of experimental settings [32] –, by calling somatic variants in transcribed regions with standard pipelines [33] and by selecting a set of confident variants, which might possibly include putative drivers. This allows one to overcome the limitation of relying on longitudinal single-cell whole genome/exome sequencing experiments, which are currently rarer and significantly more expensive. In addition, by applying standard data analysis pipelines for the analysis of transcriptomes [32], our method allows one to investigate the relation between somatic evolution and gene expression profiles in cancer clones and in single cells, especially in relation with possible external interventions, such as therapies.

There are several advantages in employing a formulation of the problem based on clonal trees, instead of standard phylogenetic trees in which single cells are placed as leaves. First, currently available single-cell data cannot guarantee the identifiability of a unique and reliable phylogenetic tree, mostly due to noise, to insufficient information and to the huge number of features of the output model [19,34]. Despite the uniqueness of the solution might not be guaranteed for clonal trees as well, in such case the computational problem is mitigated, as the model dimension is considerably reduced, especially with respect to the number of samples (i.e., cells). Second, from the biological perspective, the resolution at the clone level is an effective choice to explain and predict cancer evolution and generate hypotheses with translational relevance [34,35], in particular with regard to the possible effect of therapies, as it may allow one to pinpoint aggressive or resistant clones, investigate their molecular properties and possibly target their weak spots [36].

In order to assess the accuracy and robustness of the results produced by LACE, we performed extensive simulations, and compared with: SCITE [19] and TRaIT [22], two state-of-the-art tools for the inference of mutational trees from single-cell sequencing data, Sifit [21] a widely-used tool for tumor phylogenetic tree inference, SiCloneFit [23]) a Bayesian approach for the reconstruction of clonal trees from single-cell sequencing data, and with CALDER,myers2019calder, a recent method for the reconstruction of longitudinal phylogenetic trees from bulk samples.

We applied LACE to longitudinal single-cell datasets of tumor samples generated with distinct experimental setups and, in particular, via either scRNA-seq or targeted scDNA-seq experiments. In the first case, we employed a longitudinal scRNA-seq dataset of patient-derived xenografts (PDXs) of $\text{BRAF}^{\text{V600E/K}}$ mutant melanomas discussed in [37], by first performing mutational profiling via the widely-used GATK pipeline [38] and by selecting a panel of highly-confident somatic variants. In this respect, we here show that LACE can be used to produce robust results on longitudinal cancer evolution, even with noisy and incomplete data, and in particular, that it can characterize the efficacy of BRAF/MEK-inhibitor therapy on the clonal dynamics, also allowing one to portray the phenotypic properties of the distinct (sub)clones. In the second case, we applied LACE to a longitudinal single-cell targeted DNA-seq dataset of PDXs generated from triple-negative breast tumors, presented in [39]. Our approach was applied on the set of somatic variants identified in the original work and allowed us to refine the existing analysis, by producing a high-resolution picture of the evolutionary history of the tumor, thus proving the applicability of LACE to a wide range of existing data types.

2. Results

2.1. The LACE framework

LACE is a computational framework that processes multiple temporally ordered mutational profiles of single cells, collected from cancer samples or patient-derived cell cultures, xenografts or organoids, even in distinct experimental settings (e.g., pre- and post-treatment). Such profiles can be derived from whole-genome/exome or targeted scDNA-seq experiments, but also by calling variants from (full-length) scRNA-seq data.

LACE takes as input a binary matrix for each time point/experiment, in which an entry is 1 if a somatic variant (e.g., single-nucleotide variants — SNVs, structural variants, etc.) is present in a given cell, 0 if not present and NA if the entry is missing. In order to select highly-confident variants, one can benefit from standard practices for variant-calling and from statistical filters. Furthermore, as one might be interested in selecting a set of putative drivers, which might possibly

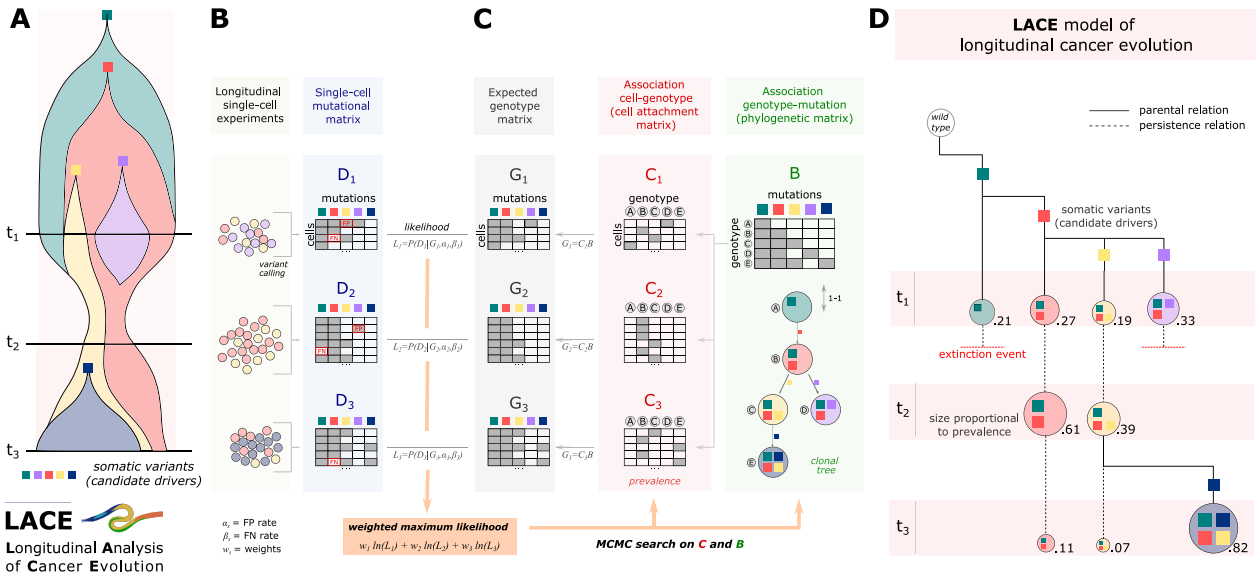


Fig. 1. The LACE framework. (A) The branching evolution of a single tumor is described via a fishplot, and is characterized by the accumulation of 5 candidate driver mutations in distinct (sub)clones. **(B)** Single cells are sampled and sequenced at three subsequent time points: t_1, t_2 and t_3, e.g., via scRNA-seq or (targeted) scDNA-seq experiments. By applying pipelines for variant calling, somatic mutation profiles for each time point are generated (matrices $\mathbf{D}_i$), which might include false positives, false negatives and missing entries. **(C)** LACE solves a Boolean matrix factorization problem $\mathbf{C}_i \cdot \mathbf{B} = \mathbf{G}_i$, in which $\mathbf{B}$ is the perfect phylogenetic matrix, $\mathbf{C}_i$ is the cell attachment matrix and $\mathbf{G}_i$ is the expected genotype matrix. It then maximizes a weighted likelihood function computed on all time points, comparing $\mathbf{G}_i$ and $\mathbf{D}_i$, via a MCMC search scheme. **(D)** As a result, a single-cell longitudinal clonal tree is returned, which may include both observed and unobserved (sub)clones and the ancestral relations between them, as well as the prevalence variation, as measured at distinct time points. Solid lines represent parental relations between (sub)clones, each one characterized by a unique somatic variant (colored squares), whereas dashed lines represent persistence relations, which connect (sub)clones through time points or to extinction events.

characterize (sub)clones, one can leverage on standard approaches for driver selection and on prior biological knowledge regarding the specific tumor type.

Our method can be input with false positive and false negative rates for each time point, whenever such information can be derived from the technical features of the experiments. Otherwise, LACE includes different noise estimation procedures, either via a parameter grid search or by including the error rate learning within the model search.

LACE then solves a Boolean matrix factorization problem with perfect phylogeny constraints [40], by maximizing a weighted likelihood function on all time points. The rationale is that experiments collected at distinct time points may include even extremely different sample sizes and technical errors: LACE allows one to balance such differences, by setting proper weights on the likelihood function. As default, the weights are set to be inversely proportional to the sample size of each dataset, in order to have comparable likelihood values through distinct experiments. LACE employs a MCMC search scheme on the phylogenetic matrix, which is defined by the association between genotypes/(sub)clones and sets of mutations, and which identifies a unique clonal tree. The weighted likelihood is maximized by exhaustively scanning the attachment of single cells to the genotypes/(sub)clones of the tree.

LACE returns:

- The *single-cell longitudinal clonal tree* describing the longitudinal evolution of a tumor, in which nodes represent genotypes and edges are either parental relations among them or persistence relations through time (see Methods for further details). Notice that genotypes represent *candidate (sub)clones* if the characterizing somatic variants are drivers.
- *The attachment of single cells to genotypes/(sub)clones*, which can be used to estimate the prevalence at any considered time point, as well as to identify macro evolutionary events, such as extinction or the emergence of new (sub)clones.

  If the input mutational profiles are generated from scRNA-seq data, this allows one to investigate the gene expression profiles of the cells mapped to distinct genetic (sub)clones, thus providing a natural link between genotypic and phenotypic properties of single cells.
- The *expected genotype of single cells*, i.e., the genotype assigned to each cell after removing noise and missing data.
- The *error rates as estimated from data*, whether either the parameter grid search or the error rate learning are employed.
- When no information on drivers is available, mutations can be grouped according to their similarity with respect to the expected genotype matrix (see Methods). LACE returns the matrix of *pairwise distances among mutations* and that can be used, in turn, in standard clustering algorithms to identify which mutations mostly occur together in the cells of the considered dataset (as default, LACE returns the results of hierarchical clustering). Accordingly, the resulting groups of mutations can be used to identify candidate (sub)clones in the output model.

All the outputs are returned by the LACE R tool, which is currently available on Bioconductor at this link: https://bioconductor.org/packages/release/bioc/html/LACE.html.

2.2. Performance evaluation via simulations

In order to assess the performance of LACE and compare it with competing approaches, we executed extensive tests on synthetic datasets, generated with the cancer population dynamics simulator from [41]. To simulate sequencing errors, all datasets were inflated with false positives with rate α (i.e., the probability of flipping a "0" entry to "1"), false negatives with rate β (i.e., the probability of flipping a "1" entry to "0"), and missing data with rate γ (i.e., the probability of setting an entry to NA). Furthermore, distinct sampling schemes were adopted (further details on the simulations are provided in the SI).

We generated a total of 1200 independent longitudinal datasets for 4 distinct experimental scenarios (see below) and compared LACE with four state-of-the-art methods that process single-cell sequencing

data from single time points, for the reconstruction of either mutational trees (SCITE [19] and TRaIT [22]), tumor phylogenetic trees (Sifit,zafar2017sifit), or clonal trees (SiCloneFit) [23]) –, and with a further method for phylogenetic tree inference from longitudinal bulk sequencing data (CALDER [27]). As synthetic datasets need to be adapted to be processed by such tools, with respect to CALDER, we computed the cancer cell fraction of driver mutations from the observed single-cell genotypes, and by assuming a uniform sampling of single cells and a read depth of 200X, which is typical for whole-exome sequencing experiments. Input data for SCITE, TRaIT, Sifit and SiCloneFit were generated by concatenating the longitudinal datasets in a unique mutational profile matrix. Notice also that the true error rates of the simulated datasets were provided to all methods that take such parameters as input (see Supplementary Table 2 for the complete parameter settings of the simulations).

We compared the performance of the methods by assessing: (i/ii) precision: $\frac{TP}{TP+FP}$ and recall: $\frac{TP}{TP+FN}$ with respect to the ground-truth topology (in this case, the True Positives are the rightly inferred edges, the False Positives are the wrongly inferred edges, the False Negatives are the edges that are present in the ground-truth topology, but are not inferred); (iii) the Adjusted Rand Index (ARI) on attachment of single-cells to clones [42]: ARI $= \frac{\sum_{ij}\binom{q_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{q}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{q}{2}}$, where q_{ij} is an element of the contingency table between the ground-truth and the inferred cell attachment matrices, a_i is the sum of the ith row and b_j the sum of the jth column of the contingency table and q is the dimension of the contingency table. The ARI, which is the corrected-for-chance version of the Rand index, is commonly used to compare partitions and here allows us to evaluate how single cells are correctly associated to ground-truth clones. Note that in the SI we show additional results regarding the overall accuracy computed with respect to the ground-truth topology, the ground-truth cell attachment matrix and the ground-truth genotypes.

Setting A — Simulations with unbalanced longitudinal datasets. We designed a first in-silico scenario to account for sequencing experiments involving highly unbalanced datasets, as generated, for example, by sequencing experiments performed via distinct platforms at different time points, a plausible setting for studies involving patient-derived cell cultures or organoids. In particular, we simulated the case in which the first and the third time points are characterized by a smaller number of cells ($n_{t_1} = n_{t_3} = 100$) and a lower noise rate (false positive rate $\alpha_{t_{1,3}} = 0.01$, false negative rate $\beta_{t_{1,3}} = 0.1$), thus modeling a plausible setting resembling a Smart-seq protocol and the second time point in which a much larger number of cells is sequenced ($n_{t_2} = 1000$), yet with a significantly higher noise rate ($\alpha_{t_2} = 0.03, \beta_{t_2} = 0.3$), therefore modeling the features of a typical droplet-based experiment. We also assume that standard pipelines for mutation profiling from scRNA-seq data are employed, and we set a 0.5 probability for each dataset to have $\gamma = 0.1$ of missing entries[2].

In Fig. 2A one can see the distribution of precision and recall with respect to the ground-truth topology and of the ARI with respect to the ground-truth cell attachment matrix, as well as the related empirical cumulative distribution functions, for the six approaches in this scenario. In addition, we display the percentage gain of LACE with respect to other methods (computed on the average value) and the p-value of the one-sided Mann–Whitney U test computed comparing the distribution of LACE against other methods, on all metrics (the distribution of the overall accuracy with respect to the ground-truth tree, cell attachment matrix and genotype, the precision–recall scatterplots, for this and the remaining experimental settings, are shown in the SI).

By managing different error rates in distinct time points and by weighting the likelihood function with respect to the sample size, LACE is able to achieve consistently superior performances with respect to all competing methods, in terms of precision/recall on the original topology and of ARI on cell attachment, as also proven by the highly statistically significant p-values of the Mann–Whitney U test on all comparisons. Moreover, LACE shows in all cases the lowest variance, which demonstrate the higher robustness of its inference framework (the additional metrics included in the SI exhibit consistent results).

This proves that LACE produces robust results when dealing with experiments from distinct protocols and with high differences in sample size and noise rates, as it might be common in real-world settings.

Settings B/C – Simulations with longitudinal datasets with technical variability. To assess the consequences of noise and of technical variability in a larger experimental setting, we generated 300 independent datasets with a number of cells chosen at each time point with uniform probability in the set $n = \{100, 300, 600, 1000\}$. We modeled a first setting with low noise (i.e., α and β randomly chosen in the set $\{0.01, 0.02\}$ and $\{0.1, 0.2\}$, respectively) and a second setting with higher values of noise (i.e., α and β in the range $\{0.01, 0.02, 0.03\}$ and $\{0.1, 0.2, 0.3\}$, respectively).

In Fig. 2B-C one can see that also in this case LACE performs better than all competing methods in precision/recall with respect to the ground-truth topology (with significant p-values of the Mann–Whitney U test on most tests) and, especially, in the ARI on cell attachment, for which all p-values are highly significant and a percentage increase of around +20% is observed with respect to the second-best performing tool, SCITE, in both settings. We recall that a better cell-clone attachment implies a better estimation of the clonal architecture, which can be then mapped to gene expression data, when variants are called from scRNA-seq data, as it will be shown in the case studies. The genotype–phenotype mapping of single cells is a major novelty of our approach and the simulations prove that LACE is the most effective approach in assigning cells to ground-truth clones.

Note that also in Settings B and C, LACE exhibit the lowest variance on all metrics, proving that the results are robust with increasing noise levels (results on further metrics are consistent and are shown in the SI).

Setting D – Simulations with longitudinal datasets with sampling limitations. Solid tumors are typically characterized by a complex spatio-temporal dynamics, which is affected by the distinct modes of tumor growth and by the complex interplay with the micro-environment [48–50]. For instance, cells close to the tumor boundary might have a certain proliferative advantage due to the relative abundance of nutrients and space, thus affecting the overall clonal dynamics.

As a consequence, in single-cell analyses, relevant sampling biases can arise from that fact that, usually, cells are not uniformly (or exhaustively) sampled from the whole tumor, but from biopsies capturing the composition of a spatially delimited portion of it [51]. In such cases, the prevalence of single (sub)clones might be either over- or under-estimated, whereas certain subclones might be unobserved, likely affecting downstream analyses. Even though an in-depth investigation of the impact of spatial tumor sampling on phylogenetic inference is out of the scope of this work, we designed a simulated experiment to assess the performance of LACE and competing methods in this specific scenario.

In particular, in this setting all methods were applied to longitudinal datasets generated with significant sampling limitations, according to which only a subset of existing clones is sampled at any time point, in

[2] Clearly, the estimation of error rates in real-world data is a complex topic, which is influenced by the combination of technical issues related to data types (e.g., DNA or RNA), to the features of sequencing protocols/platforms and to the many practical choices during the variant calling phase, for instance related to quality check [44–47]. In this work, we assessed the performance of LACE and of competing approaches with simulated data including false positives, false negative and missing data in ranges adequate to cover a wide spectrum of realistic experimental scenarios, in line with what done in similar works on the subject [19–21,23].

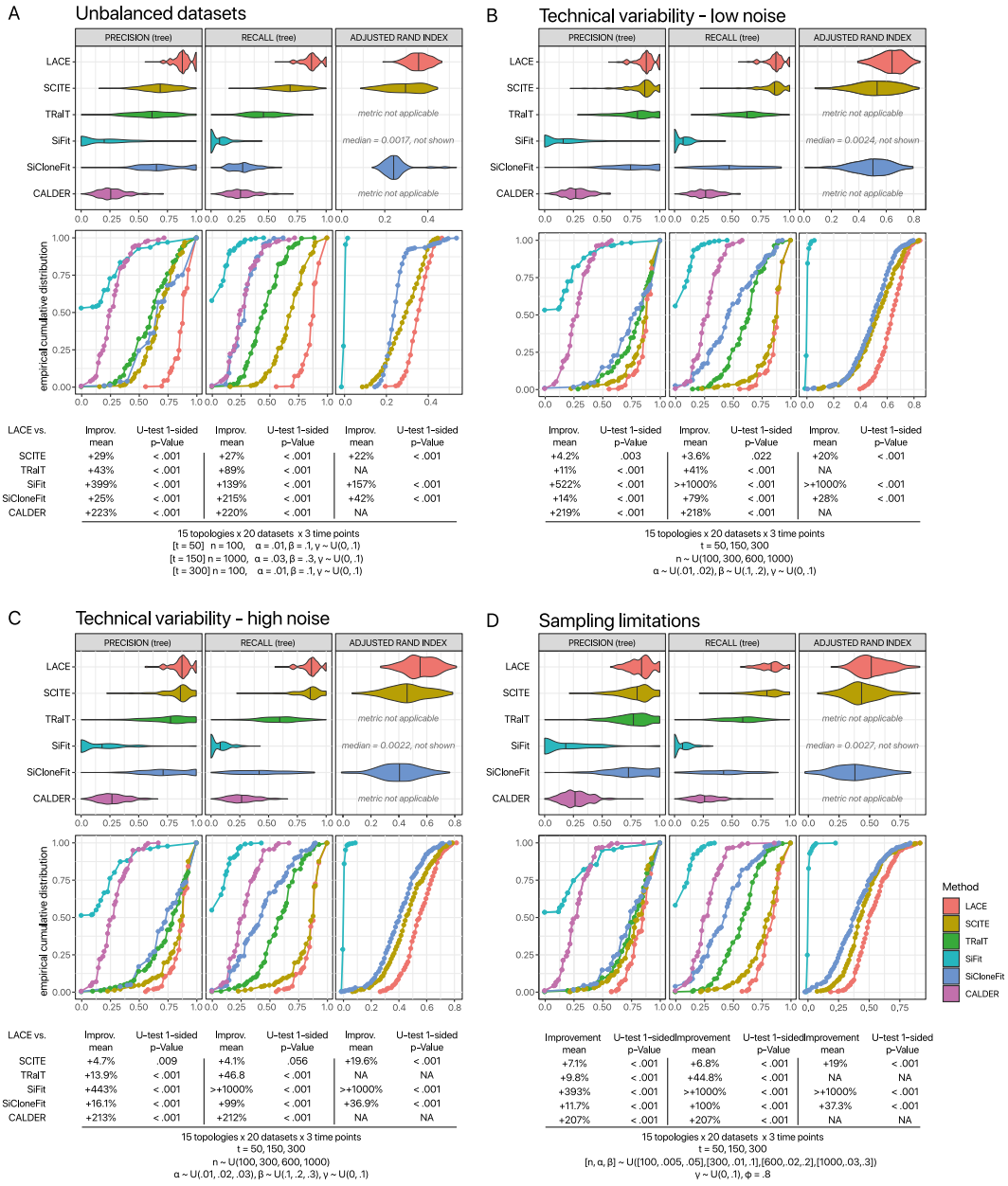


Fig. 2. Comparison on simulated data. The cancer population dynamics was simulated with the tool from [43] (see the SI for further details). 15 simulations were selected in which a number of drivers between 7 and 15 was observed, resulting in branching evolution topologies. For each topology, a number of independent single-cell mutational profile datasets were sampled at 3 distinct time points of the dynamics, for a total of 1200 longitudinal datasets. LACE was compared with SCITE [19], TRaIT,ramazzotti2017learning, Sifit [21], SiCloneFit,zafar2019siclonefit and CALDER [27], on precision and recall with respect to the ground-truth topology and the Adjusted Rand Index (ARI) with respect to the ground-truth cell attachment matrix. Upper panels return the distributions on all metrics via violin plots, whereas lower panels return the empirical cumulative distribution functions. The percentage increase of LACE with respect to competing methods was computed on average values for each metric, whereas the p-value of the one-sided Mann–Whitney U test was computed on the distributions (further metrics are provided in Supplementary Fig. 4 and 5). Notice that the ARI cannot be computed for TRaIT and CALDER, as such methods do not return cell attachments, while it is not shown for Sifit because of the low values. **(A)** To simulate unbalanced longitudinal datasets, we generated 20 independent datasets for each topology, including $n = 100$, $\alpha = 0.01$, $\beta = 0.1$ for time points $t = 50$ and $t = 300$, and $n = 1000$, $\alpha = 0.03$, $\beta = 0.3$ for $t = 150$. Each dataset in this and the next scenarios was inflated with 10% of missing data, with a 0.5 probability. **(B)** To model technical variability, 20 independent datasets were sampled for each topology, in which at each time point ($t = 50, 150, 300$), each dataset includes with uniform and independent probability: a number of cells in the set $\{100, 300, 600, 1000\}$, α in the set $\{0.01, 0.02\}$ and β in the set $\{0.1, 0.2\}$. **(C)** To account for higher error rates, 20 independent datasets for each topology were generated as in (B), with values of α in the set $\{0.01, 0.02, 0.03\}$ and β in the set $\{0.1, 0.2, 0.3\}$. **(D)** To model datasets generated with sampling limitations, we generated 20 independent datasets for each topology, in which at each time point ($t = 50, 150, 300$), a number of cells in the set $\{100, 300, 600, 1000\}$ was sampled from $\phi = 80\%$ of the existing clones. Datasets were inflated with noise proportional to sample size: $\alpha = 0.005$, $\beta = 0.05$ for $n = 100$, $\alpha = 0.01$, $\beta = 0.1$ for $n = 300$, $\alpha = 0.02$, $\beta = 0.2$ for $n = 600$ and $\alpha = 0.03$, $\beta = 0.3$ for $n = 1000$.

order to estimate the effectiveness in inferring both the ground-truth generative topology and the underlying clonal dynamics.

In detail, we generated 300 independent datasets from 15 topologies, in which we sampled a number of cells in the range $n = \{100, 300, 600, 1000\}$ from $\phi = 80\%$ of the existing clones at any given time point ($t = 50, 150, 300$). For instance, if at time $t = 150$ 9 clones were present in the simulated data, all single cells were sampled from 7 randomly selected clones. Datasets were then inflated with noise rates

proportional to the sample size, and a 0.5 probability was set for each dataset to have $\gamma = 0.1$ of missing entries, as for the other settings (for the complete parameter settings please refer to the SI).

Also in this scenario, LACE outperforms all competing methods with respect to all metrics, with highly significant p-values of the Mann–Whitney U test and noteworthy percentage gains, also displaying the lowest variance. Such results prove that LACE is able to deliver a robust temporal representation of intra-tumor heterogeneity from longitudinal data even in conditions of relevant sampling limitations, as expected when processing sequencing data from biopsies of solid tumors (see the case studies in the next subsection).

All in all, the results on simulated data show the effectiveness of LACE in delivering accurate and robust results in a wide range of realistic experimental scenarios, in which temporally ordered datasets from single-cell sequencing experiments are characterized by a high variability of sample size, error rates and sampling modes. We also recall that LACE's output is more expressive than those of methods designed to process single-time point datasets, as it allows one to quantify the clonal prevalence variation in time, as well as to estimate the temporal positioning of phenomena such as clone emergence or extinction, for instance as a consequence of a therapy.

Further experiments on simulated data. We further tested the robustness of LACE in a variety of additional simulated scenarios. In particular, we assessed the inference accuracy when the real error rates are not provided as input and the grid search is employed. In Supplementary Fig. 8 one can see that LACE is robust and produces reliable results when employing the grid search on false positive/negative rates (please refer to the SI, in particular Supplementary Table 3, for details on the settings of this and the following experiments). Analogously, the results are stable even when wrongly specified error rates are provided as input to LACE (Supplementary File 3).

We also tested a number of additional in-silico scenarios which include violations of the Infinite Sites Assumption (ISA) [40] and involving, in particular: (*i*) back mutations, due, e.g., to loss of heterozygosity (LOH), and (*ii*) homoplasies (i.e., identical variants observed in independent branches of the model), which underlie scenarios of convergent/parallel evolution. In Supplementary Fig. 6 one can see that LACE achieves optimal performances also in presence of relatively high proportions of variants affected by back mutations, whereas in Supplementary Fig. 7 the robustness of the results delivered by LACE is proven also with respect to convergent evolution scenarios.

2.3. Application of LACE to real-world datasets

Application of LACE to longitudinal scRNA-seq dataset from PDXs of BRAF-mutant melanomas — Dataset #1. We applied LACE to a longitudinal dataset from [37]. In the study, the authors analyze multiple omics data generated from both bulk and single-cell experiments, to investigate minimal residual disease (MRD) in patient-derived xenografts from BRAF-mutant melanomas. In particular, they expose PDXs to $\text{BRAF}^{\text{V600E/K}}$ inhibitor (i.e., *dabrafenib*), either alone or in combination with a MEK inhibitor (i.e., *trametinib*), and they perform multiple sequencing experiments at different time points.

Despite finding de novo mutations in known oncogenes (e.g., MEK1 and NRAS) in resistant cells, the authors' analyses of the copy number alteration profiles, performed via massively parallel sequencing of single-cell genomes, was not sufficient to effectively characterize the clonal architecture and evolution of the tumor, whose composition appear to be similar prior to and after the treatment. Conversely, by analyzing transcriptomic data from both bulk and single-cell RNA-seq experiments, the authors were able to identify four distinct cell subpopulations, characterized by specific transcriptional states (i.e., *neural crest stem cell* – NCSC, *invasive*, *pigmented* and *starved-like melanoma cell* – SMC), which are insensitive to treatment and eventually lead to relapse, whereas the remaining cell subpopulations get quickly extinct. Based on these findings, the authors hypothesize that the co-emergence of drug-tolerant states within MRD is predominantly due to the phenotypic plasticity of melanoma cells, which results in transcriptional reprogramming.

We here aim at refining the analysis of the clonal evolution of the tumor, by employing single-cell mutational profiles, as generated by calling variants from scRNA-seq data. In particular, we employed the GATK Best Practices [38] to identify good-quality variants and we finally selected 6 somatic SNVs, by applying a number of filters based on statistical and biological significance, and which might be considered as candidate drivers for this tumor (see Methods). We finally executed LACE with 50000 MCMC iterations, 100 restarts and by employing the grid-search on error rates.

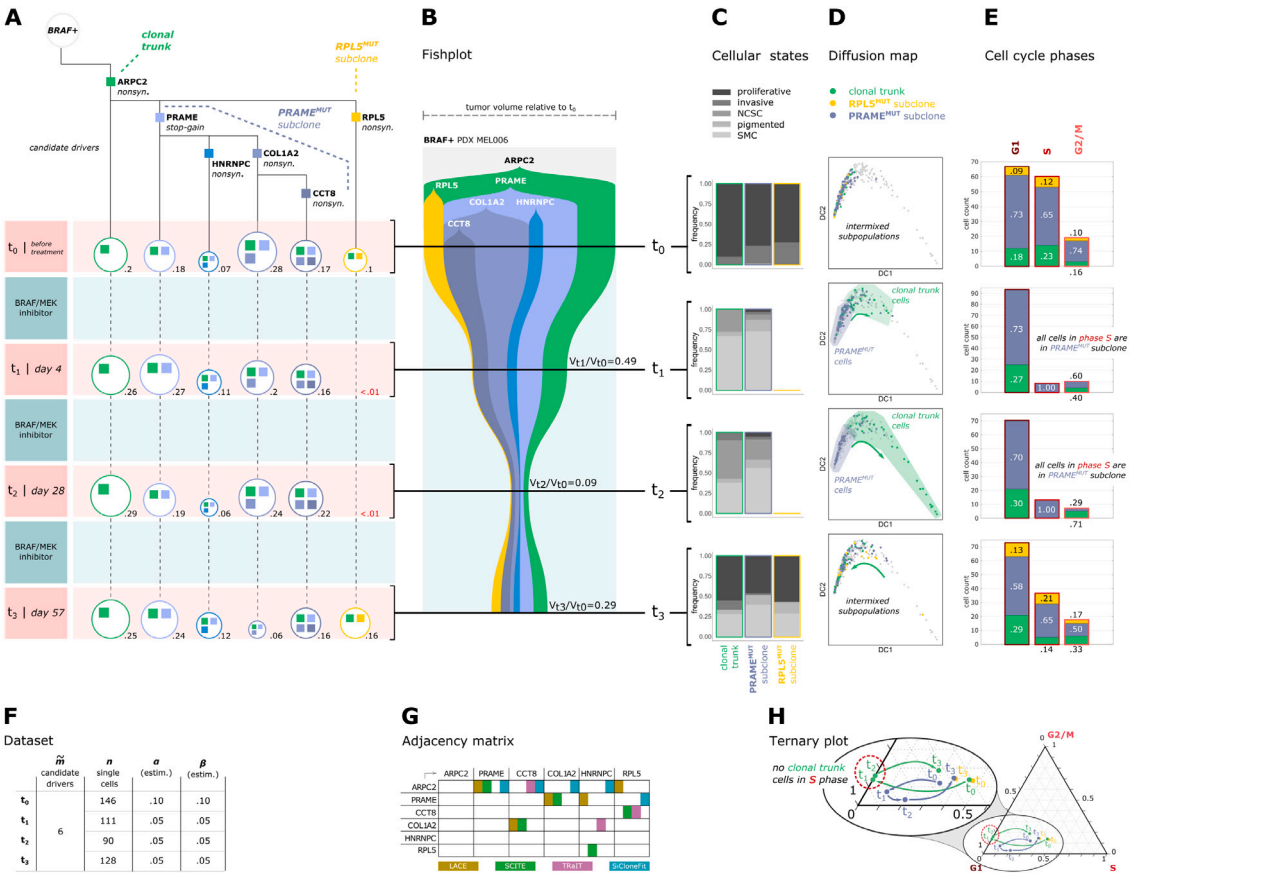
The LACE model shown in Fig. 3A–B reveals the presence of a clonal trunk including a nonsynonymous somatic mutation on ARPC2 – a known melanoma marker [54] –, and of two distinct subclones, with somatic mutations on PRAME and RPL5 as initiating events. In particular, PRAME is a melanoma-associated antigen and known prognostic and diagnostic marker, which was recently targeted for immunotherapy [55]. RPL5 is a candidate tumor suppressor gene for many tumor types and displays inactivating mutations or focal deletions in around 28% of melanomas, which usually result in somatic ribosome defects [56].

The $\text{PRAME}^{\text{MUT}}$ subclone is characterized by the accumulation of further nonsynonymous mutations in HNRNPC, COL1A2 and CCT8 and displays an overall prevalence around $\sim 70\%$ before treatment (t_0). In particular, HNRNPC is a heterogeneous ribonucleoprotein involved in mRNA processing and stability, and its regulation appears to be involved in the PLK1-mediated P53 expression pathway, as it was shown via quantitative proteomic analysis on $\text{BRAF}^{\text{V600/E}}$ mutant melanoma cells treated with PLK1-specific inhibitor [57]. COL1A2 is an extracellular matrix protein that is supposed to maintain tissue integrity and homeostasis, and was identified as a melanoma marker [58], as well as a candidate prognostic factors in several cancer types [59]. CCT8 encodes the theta subunit of the CCT chaperonin, and was found as up-regulated or mutated in several cancer types [60].

After 4 days of therapy (t_1), the tumor volume halves. In particular, the RPL5^{MUT} subclone seems to disappear, whereas both the clonal subpopulation and the $\text{PRAME}^{\text{MUT}}$ subclone maintain a stable prevalence. A further reduction of tumor volume is observed after 28 days of treatment ($\sim$ 9% of the volume at t_0), whereas at day 57 (t_3) a significant growth is observed, in which the tumor reaches $\sim$ 29% of the initial volume, hinting at a possible relapse likely due to cells developing resistance. RPL5^{MUT} subclone reappears at time t_3 with a prevalence around $\sim$ 16%, suggesting that its absence at time points t_1 and t_2 may be due to sampling limitations, which however do not affect the capability of LACE of inferring a correct evolution model. At time t_3, the prevalence of all (sub)clonal subpopulations is similar to that of time t_0, hinting at the absence of significant clonal selection. In Fig. 3G one can find the adjacency matrix representing the parental edges of the models inferred by LACE, SCITE, TRaIT and SiCloneFit (the complete models are shown in Supplementary Figures 10, 11 and 12). One can notice that, despite noteworthy differences, several phylogenetic relations, e.g., ARPC2 – PRAME, are retrieved by distinct methods.

We then analyzed the composition of (sub)clones with respect to the cellular states identified in [37] via single-cell transcriptomics analysis (Fig. 3C). Consistently with the findings in the article, the majority of cells in all (sub)clones are in a proliferative state before treatment, whereas at time point t_1 and t_2 no proliferative cells are left and all (sub)clones undergo transcriptional reprogramming, by displaying heterogeneous cell states (mostly starving-like melanoma cells, SMC), which result in acquired resistance at time t_3, when cells restart proliferating in all (sub)clones.

As only minor differences are observed among (sub)clones with respect to cellular states, we refined the transcriptomic analysis, by first



	$\tilde{m}$ candidate drivers	n single cells	α (estim.)	β (estim.)
t_0	6	146	.10	.10
t_1		111	.05	.05
t_2		90	.05	.05
t_3		128	.05	.05

Fig. 3. LACE model — PDX MEL006. (A) The longitudinal evolution of the PDX MEL006 derived from a BRAF mutant melanoma [37] returned by LACE is displayed. Single-cells were isolated and sequenced via scRNA-seq at four subsequent time points: (t_0) before treatment ($n = 146$ single cells); (t_1) after 4 days of concurrent treatment with BRAF inhibitor (i.e, dabrafenib) and MEK inhibitor (i.e, trametinib) ($n = 111$), (t_2) after 28 days of treatment ($n = 90$), (t_3) after 57 days of treatment ($n = 128$). Single-cell mutational profiles from scRNA-seq datasets were generated by applying the GATK Best Practices [38], and $m = 6$ somatic variants were selected as candidate drivers to be used as input in LACE, with the procedure described in the main text. Each node in the output model represents a candidate (sub)clone, characterized by a set of somatic variants (colored squares). Solid edges represent parental relations, whereas dashed lines represent persistence relations. The clonal prevalence, as measured by normalizing the single-cell count, is displayed near the nodes and marked in red if lower than 1%. **(B)** The representation via a standard fishplot, generated via TimeScape [52], is displayed. The volume size at different time points is taken from [37]. **(C)** The composition of the clonal subpopulation (green) and of both the RPL5^{MUT} (yellow) and PRAME^{MUT} (blue) subclones with respect to the cellular states identified via single-cell transcriptomics analysis in [37] is shown for each time point (cells labeled by the authors as "other" were excluded from the analysis). **(D)** The diffusion maps [53] computed on 58 differentially expressed genes identified via ANOVA test (FDR adjusted $p < 0.10$) is shown; plots are generated via SCANPY,wolf2018scanpy. A distinct diffusion map is shown for each time point, in which only the cells sampled at each time point are colored according to the clonal identity (i.e., clonal trunk, RPL5^{MUT} subclone or PRAME^{MUT} subclone). **(E)** The proportion of cells in G1, S and G2/M phases with respect to the distinct (sub)clones is shown with barplots for each time point. Cell phases are estimated on 97 cell cycle genes via SCANPY. At time point t_1 and t_2 all cells in phase S belong to the PRAME^{MUT} subclone. **(F)** In the table, the number of somatic variants $\tilde{m}$, the number of single cells n and the estimated false positive and false negative rates, α and β, are returned for each time point. **(G)** The adjacency matrix describing the edges of the models returned by LACE, SCITE, TRaIT and SiCloneFit is shown. **(H)** A ternary plot showing the trajectories of the (sub)clonal subpopulations in the cell cycle space is shown. In the barycentric plot the three variables represent the ratio of cells belonging to phase G1, S and G2/M, respectively, and sum up to 1.

focusing on differentially expressed genes. The differential expression analysis, performed via standard ANOVA among all (sub)clones on all time points (data normalized by library size, FDR adjusted $p < 0.10$), allowed us to identify PRAME as significantly up-regulated in PRAME^{MUT} cells, with $\log_2$-fold-change = 0.71; several other genes are found as differentially expressed among (sub)clones, yet with larger values of FDR p-value (the list of differentially expressed genes and the relative FDR p-values and fold-change values are provided in Supplementary File 1; please refer to the SI for further details on the analysis).

In order to analyze in depth the transition leading cells to resistance, we performed the same analysis with respect to the distinct time points. Interestingly, the list of differentially expressed genes among (sub)clones (FDR $p < 0.10$) includes no genes at time t_0, t_1 or t_3, but it includes 58 genes at time t_2 (see the Supplementary File 1). Within such group, 5 genes are significantly up-regulated in PRAME^{MUT} cells and display a $\log_2$-FC larger than 3, namely NGLY1 ($\log_2$-FC = 4.28), CDCA7 ($\log_2$-FC = 3.45), HK1 ($\log_2$-FC = 3.27), DNAJB4 ($\log_2$-FC = 3.27), ISOC2 ($\log_2$-FC = 3.11; the distribution of gene expression values at time t_2 in PRAME^{MUT} and PRAME^{WT} cell subpopulations is shown in Supplementary Fig. 13). The results of this analysis suggest that, in addition to shifting their cellular states, distinct (sub)clones may differently respond to the therapy, and this would result in a transient increase of phenotypic heterogeneity, especially at time t_2.

This aspect is particularly evident by looking at the projection of single cells in the space of the 58 most significant differentially expressed genes (FDR $p < 0.10$), represented via diffusion maps [53] in Fig. 3D. Before treatment (t_0), almost all cells are positioned in the left region of the map and appear to be highly intermixed, proving the existence of a homogeneous phenotypic behavior. At time t_1, the RPL5^{MUT} subclone (yellow) disappears, whereas the clonal subpopulation (green) undertakes an apparent shift toward the right region of the map, which is characterized by transcriptional patterns that progressively diverge from those observed prior to the therapy, and

which may possibly indicate high levels of cellular stress. This effect is notably amplified at time t_2, where an explicit split of the clonal and the PRAME^{MUT} subpopulations can be observed, also in correspondence of the maximum dispersion of the cells on the map.

This outcome would further prove that distinct genetic clones may indeed suffer the effects of BRAF/MEK inhibition in different ways, during the resistance development phase, and this would result in different transcriptional patterns. At time point t_3, when cells have achieved resistance and restart proliferating, RPL5^{MUT} subclone expands, and all cells appear to be intermixed on the left portion of the diffusion map once again.

We analyzed the cell cycle phase of the single cells at different time points, as estimated on 97 cell cycle genes via SCANPY [61]. In Fig. 3E one can see that cells are distributed across phases G1, S and G2/M in expected proportions in all (sub)clones before therapy (t_0). Strikingly, at time point t_1 and t_2 all cells in phase S belong to subclone PRAME^{MUT}, whereas all cells of the clonal subpopulation are found in phase G1 or G2/M. At time t_3 the scenario resembles that of time t_0 and all (sub)clones include cells in all cell cycle phases. By looking at the ternary plot representing the proportion of cells in different cell cycle phases (Fig. 3H), it is possible to notice that the clonal and the PRAME^{MUT} subpopulations indeed undertake distinct trajectories in presence of therapy, before returning to a state similar to the initial one.

This major result proves that the concurrent BRAF/MEK inhibition indeed affects in distinct ways different genetic clones during the resistance development stage. Apparently, cells lacking the PRAME mutation would be prevented from proceeding into S phase, whereas this effect would be highly mitigated in PRAME^{MUT} cells. All in all, these results cast a new light on the relation between clonal evolution and phenotype at the single-cell level, and suggest that distinct genetic clones may respond to therapy in significantly different ways.

In the SI we present a detailed analysis of the stability of LACE's results for this case study. In particular, we show that the consistency of the output model is preserved when either subsets or supersets of the selected somatic variants are considered and, similarly, when performing downsampling and oversampling of the original dataset (see Supplementary Figs. 15 and 16).

Application of LACE to longitudinal targeted scDNA-seq dataset from PDXs of triple-negative breast tumors — Dataset #2. We applied LACE to a second longitudinal dataset from single-cell targeted DNA-sequencing experiments, presented in [39]. In the work, primary and metastatic tissues of breast cancer patients were serially transplanted into immunodeficient mice to generate xenograft lines, and the clonal dynamics was investigated by tracking single nucleotide and structural variants from bulk sequencing experiments. In addition, the authors performed targeted deep SNV re-sequencing to validate the clonal expansion of two xenograft series at the single-cell resolution. Sample SA501, in particular, was generated from a triple-negative breast primary tumor and was selected by the authors because of the interesting post-engraftment clonal dynamics.

After pre-processing and QC, the mutational profile matrix provided as input to LACE includes $\tilde{m} = 20$ somatic variants and $n = 27, 36, 27$ single cells for time points t_0 = X1, t_1 = X2 and t_2 = X4, respectively (Fig. 4B, please refer to the Methods section for further details on this dataset).

As no procedure for driver identification was employed in the original article, we here applied a key feature included in LACE, which allows one to cluster variants with respect to the similarity on expected genotype profiles of the single cells. The intuition is that the somatic variants displaying similar patterns of co-occurrence across single cells may indicate a single (sub)clone, composed of cells with distinct genotypes (Fig. 4C; see Methods for further details). In Fig. 4A the LACE model of sample SA501 is displayed, which delivers a picture of the longitudinal evolution of the tumor at the resolution of both genotypes and candidate (sub)clones (LACE was executed with 50000 MCMC iterations, 100 restarts and by employing the grid-search on error rates).

6 candidate (sub)clones are identified by LACE, including a varying number of genotypes, and describing a complex clonal architecture. More in detail, a branching evolution scenario is observed, revealing the presence of two main subclones (#1 and #5), which appear from the clonal subpopulation prior to the first passage (t_0). From subclone #1, further subclones emerge and are progressively selected through time, with a late subclone (#4) originating after time t_1 and leading to a selection sweep at time t_2 (clonal prevalence $\sim 88\%$). Interestingly, the evolution model returned by LACE is consistent with that proposed by the authors of the original work, who employed PyClone [62] to cluster mutations from bulk sequencing data and MrBayes [63] to reconstruct a phylogenetic model on single-cells (Fig. 4D). In particular, the cascading acquisition of mutations from parental to descendant clones observed by the authors is here refined by the explicit temporal ordering among genotypes provided by LACE, and which might used to generate hypotheses on driver identification (the cell attachment inferred by our analysis and that returned in the original work are provided in Supplementary File 2).

The application of LACE to a further dataset from [39] (sample SA494) is discussed in the SI (Supplementary Fig. 17), in which we show that our method is able to identify a rare clone at the metastatic level, prior to its expansion observed after transplanting in xenografts.

3. Discussion

Cancer is an evolutionary process in which cells progressively accumulate somatic variants and undergo selection, while competing in a complex microenvironment [64]. Such variants can be used to track the evolution of a single tumor, to characterize intra-tumor heterogeneity and to identify the genomic makeup of (sub)clones responsible for therapy resistance or phenotypic switches [65].

LACE is the first algorithmic framework that can process single-cell datasets collected at different time points to produce longitudinal clonal trees of tumor evolution. Our approach can explicitly model different error rates and sample size in distinct experiments, which is typical in longitudinal studies. Accordingly, it can leverage the information extracted from possibly biased or non-exhaustive samplings of the tumor's cells, which, instead, might lead to erroneous evolutionary inference when using single-time point datasets. Moreover, as the results are noise-tolerant, LACE can deliver reliable results even with extremely imperfect mutational profiles, as those derived by calling variants from transcriptome. This allows one to exploit transcriptomic data, commonly available for most single-cell studies, to assess the clonal composition and the history of a tumor, and to directly investigate for the first time the relation between genomic clonal evolution and phenotype at the single-cell level, for instance in response to a certain treatment.

We note that new reliable protocols for high-quality genotyping of single cells from transcriptomic data are starting to appear, as proposed, for instance, with the Genotyping of Transcriptomes approach [18], in which the 10x Genomics platform is modified to amplify the targeted transcript and locus of interest. The improvement of the confidence on variant calling will accordingly enhance the accuracy of the results delivered by LACE. Furthermore, it was recently shown [66] that somatic mutations in mitochondrial DNA can be tracked by single-cell RNA sequencing and used for efficient lineage tracing. LACE might be easily employed with data on somatic mtDNA mutations from longitudinal cancer samples, once available. On a side note, the recent impressive advancements in spatial genomics/transcriptomics [67,68] may soon reach the single-cell resolution, allowing for a prompt employment of our theoretical framework on longitudinal samples, hence paving the way for an integrative analysis of the spatial and temporal properties of cancer evolution. In any scenario, it is reasonable to expect a surge

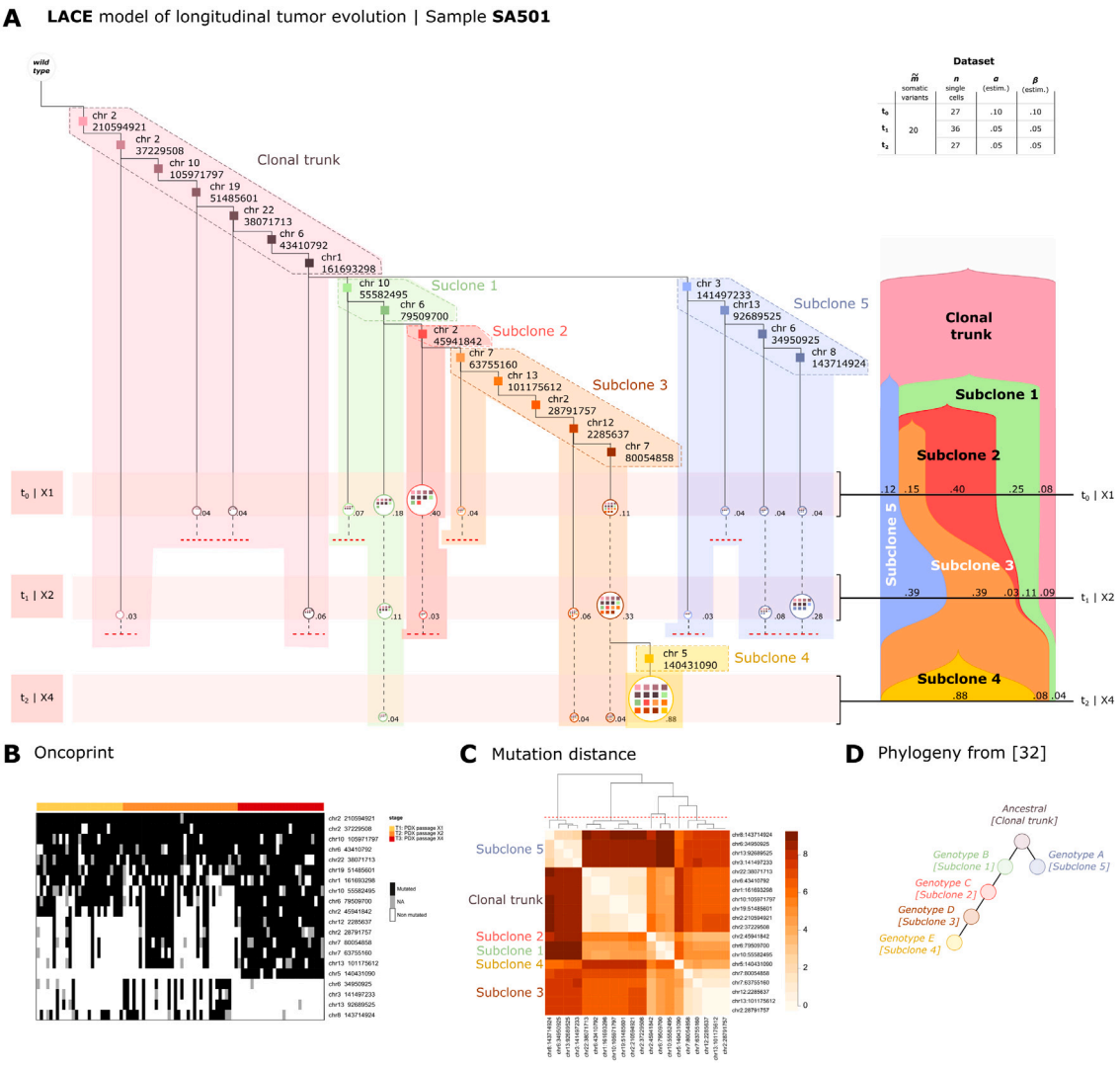


Fig. 4. LACE model — Sample SA501. (A) The longitudinal evolution inferred by LACE on 3 PDXs generated from triple-negative breast tumor sample SA501 [39] is displayed. Targeted DNA sequencing experiments were performed on single cells of PDXs at three subsequent passages – $t_0 = X1$ ($n = 27$ single cells), $t_1 = X2$ ($n = 36$) and $t_2 = X4$ ($n = 27$), by selecting a panel of 10 germline SNVs and 45 somatic SNVs. Single-cell mutational profiles were generated by discretizing the alternative allele ratio as proposed in the original work. Somatic SNVs displaying a ratio of missing data > 0.15 and germline mutations were filtered-out prior to the inference. As a result, $m = 20$ somatic variants were selected to be used as input in LACE. Each node in the output model represents a genotype characterized by a set of somatic mutations (colored squares). Solid edges represent parental relations, whereas dashed lines represent persistence relations. The prevalence of each genotype, as measured by normalizing the single-cell count, is displayed near the nodes. Colored shades mark 6 groups of mutations, which are identified by clustering variants according to the mutation distance defined in the text and shown in panel **C**, and which might represent candidate (sub)clones (the number of clusters was selected by following the analysis in [39]). In upper-right box, the number of somatic variants $\tilde{m}$, the number of single cells n and the estimated false positive and false negative rates, α and β, are returned for each time point. The representation via a standard fishplot, generated via TimeScape [52] at the resolution of candidate (sub)clones, is displayed on the right. **(B)** The oncoprint returning the single-cell mutational profile used in the analysis is shown. Black cells represent single cells in which the SNV is present, whereas gray cells represent missing values. **(C)** A heatmap returning the mutation distance among the 20 somatic variants used in the analysis is displayed (see Methods). 6 clusters identify candidate (sub)clones, and correspond to the 6 genotypes discussed in [39]. **(D)** The phylogenetic clonal model presented in [39] is displayed. .

of high-quality cancer longitudinal single-cell datasets, which might be used in LACE to track the evolution of tumors at unprecedented resolution.

Even though the application of longitudinal single-cell sequencing in clinical settings is still in its infancy, we expect a rapid diffusion of these techniques, for instance in the context of hematological clonal disorders, where cancer cells are readily accessible over time. The availability of tools dedicated to the analysis of longitudinal single-cell data will allow the study of the detailed molecular mechanisms triggered by the therapies directly in primary cancer cells [36]. Notably, in many hematological clonal disorders, such as Chronic Myeloid Leukemia, the leukemic stem cells and in particular the quiescent subset proved to be resistant even to targeted therapies such as Imatinib or second generation BCR-ABL1 inhibitors [69,70]. Unfortunately, the nature of this resistance is still elusive, owing to the technical challenges involved in studying rare cell populations during treatment by using conventional approaches. In this scenario, LACE may be employed to characterize the resistance mechanisms selectively occurring in small subsets of the cancer cells pool.

As shown in the case studies, the innovative features of LACE allow one to deliver experimental hypotheses with translational relevance. For instance, the model inferred from the BRAF-mutant melanoma PDX dataset produced a high-resolution picture of the evolutionary history of the tumor, as shaped by events such as the administration of a

therapy. Furthermore, LACE allowed us to identify (sub)clones that show different sensitivity to the therapy. In particular, we detected an unexpected behavior of the cell cycle machinery in different (sub)clones upon treatment, with only $\text{PRAME}^{\text{MUT}}$ subclone maintaining a fraction of cells in S phase. These findings suggest that longitudinal single-cell analyses are effective in dissecting the mechanisms by which cancer cells react upon treatment, at least for clonal disorders where tumor cells are readily accessible. All in all, by explicitly allowing a mapping between the clonal evolution and the phenotypic properties of single cells, LACE proved to be a powerful and expressive tool to decipher intra-tumor heterogeneity on multiple scales.

Our method was proven to be robust with respect to violations of the ISA on due to, e.g., back mutations and homoplasies. Yet, the theoretical framework might be easily extended to account for possible violations of the ISA, as proposed, e.g., in [21,71,72], and by possibly leveraging the information on copy-number alterations. Furthermore, as both bulk and single-cell sequencing data may be increasingly available in longitudinal studies, the integration of both data types within our framework might allow one to improve the clone identification and the inference quality, as proposed in [35,73]. Finally, once a significant number of longitudinal single-cell datasets would be available on specific cancer types, techniques based on transfer learning might be applied to our models to identify possible patterns of recurrent evolution across tumors, as proposed by some of the authors in [74].

4. Methods

4.1. Inputs

LACE requires longitudinal single-cell mutational profiles, as computed on a panel of selected somatic variants, e.g., single-nucleotide variants (SNVs), indels or structural variants. Mutational profiles can be generated via variant calling, e.g., from single-cell DNA-sequencing data – either whole-genome/exome or targeted sequencing –, or single-cell RNA-sequencing data.

In particular, the latter represents a cost-effective and increasingly widespread option [75], despite known technological issues, such as the impossibility of calling variants from non-transcribed regions, the typically low coverage and the high levels of data-specific noise [76,77]. However, we have shown that the results of LACE are noise-tolerant (see Fig. 2) and, therefore, it can be efficiently applied to possibly incomplete or noisy mutational profiles.

Notice that, as any statistical inference method, the reliability of the results is higher when the sampling bias is limited, as single-cell samplings should be ideally significant of the tumor's composition at any time point. However, by exploiting the information extracted from multiple datasets sampled from the same evolutionary history, LACE can deliver statistically significant output models even with possibly biased, incomplete or imperfect samplings, as opposed to standard methods for single-time point data.

In our framework, a *genotype* defines a subset of single cells sharing the same set of somatic variants. In order to select high-quality variants, filters on statistical significance (e.g., recurrence thresholds) and on clinical/functional features should be employed, aimed at reducing the impact of noise of single-cell measurements and excluding non functional variants (e.g., rare polymorphisms).

Moreover, it may be sound to select somatic variants that might be candidate drivers for that specific tumor, i.e., mutations impacting the phenotypic variation among clonal populations. To this end, the presence of known variants involving oncogenes and tumor suppressor genes should be first verified. Previously uncharacterized mutations might be also selected, if significantly present in the samples. The identification of putative drivers allows one to define *(sub)clones*, which we here consider as subsets of single cells sharing the same set of drivers. However, as driver identification is a typically hard task, at least when relying on data of single patients, one should prudently refer to *candidate* (sub)clones when interpreting the output model.

Furthermore, in case no reliable driver selection is possible, after the inference LACE allows one to group genotypes according to the similarity of co-occurrence of mutations across single-cells (see below); in fact, groups with similar patterns of co-occurrence may indicate a candidate (sub)clone. In any case, the clonal evolution of the tumor inferred by our method is consistent at the resolution of both genotypes and (sub)clones, as LACE's theoretical framework relies on the existence of a coherent process of accumulation of somatic variants.

4.2. The algorithmic framework

Longitudinal single-cell data factorization problem. Consider the case of y single-cell (DNA or RNA) sequencing experiments, taken at different time points $t_1 \leq t_2 \leq \cdots \leq t_y$. After variant calling and pre-processing (see above), for each time point t_s (with $s = 1 \dots y$), a *data matrix* $\mathbf{D}_s$ is generated, including the binary mutational profiles of n_s different single cells on the set of $\tilde{m}$ somatic mutations observed in at least one experiment. In detail, each row of $\mathbf{D}_s$ represents a single cell of the sth experiment and each column a mutation: an element $d^s_{i,j}$ is 1 if we observe mutation j in cell i of the sth experiment, is 0 if the mutation is not detected, is labeled with NA if the coverage is below a user-defined threshold. Notice that $\mathbf{D}_s$ can include false positives and false negatives.

As a pre-processing step, LACE first merges the columns that have identical values on all the rows, i.e., related to those mutations occurring exactly in the same single cells, because statistically indistinguishable (as originally proposed in [22]). In such cases, the final model will include aggregate mutational events (e.g., mutation A *and* mutation B).

We can then define the following binary matrices:

1. The *expected genotype matrix* $\mathbf{G}_s$: a binary $n_s \times \tilde{m}$ matrix where each row represents a single cell of the sth experiment (taken at time t_s) and each column a mutation or an aggregate mutational event (for the sake of clarity, in the following we will refer to single mutations only). An element $g^s_{i,j}$ is 1 if the mutation j is theoretically expected in cell i of the sth experiment, given the assumptions of our modeling framework, otherwise $g^s_{i,j}$ is 0.
2. The *phylogenetic matrix* $\mathbf{B}$: a binary $k \times \tilde{m}$ matrix where each row represents a genotype and each column a mutation; $b_{ij} = 1$ if mutation j is present in genotype i, otherwise $b_{i,j} = 0$. Each $\mathbf{B}$ can uniquely be represented by a tree and vice versa [78,79]. Moreover, if we assume that the phylogenetic process is *perfect* and that the Infinite Sites Assumption (ISA) holds [40], i.e., mutations are never lost, there are no homoplasies and there is only one root. $\mathbf{B}$ has the following properties: (i) $\mathbf{B}$ is a square matrix ($k = \tilde{m}$), i.e., the number of genotypes is equal to the number of mutations; (ii) the rank of $\mathbf{B}$ is k; (iii) the Hamming distance between any pair of rows of $\mathbf{B}$ is ≥ 1 and there is a column of $\mathbf{B}$ where all entries are 1; (iv) $\mathbf{B}$ is a lower triangular matrix and all the elements of its diagonal are equal to 1 [78,79].
3. The *cell attachment matrix* $\mathbf{C}_s$: a binary $n_s \times k$ matrix, where each row represents a single cell of the sth experiment and each column a genotype; $c^s_{i,j} = 1$ if cell i is associate to genotype j, otherwise $c^s_{i,j} = 0$. We notice that each cell is attached exactly to one genotype, i.e., the sum of any row of $\mathbf{C}_s$ is equal to 1. Note that this will allow one to easily compute the prevalence of any genotype in the sth time point.

Assuming that, in a given tumor, there exists a unique generative phylogenetic matrix $\mathbf{B}$, given a set of expected genotype matrices $\{\mathbf{G}_s\}_{s=1}^{y}$, the following factorizations hold:

$$\mathbf{G}_s = \mathbf{C}_s \cdot \mathbf{B}, \quad s = 1, 2, \dots, y, \tag{1}$$

where $\{\mathbf{C}_s\}_{s=1}^{y}$ is the set of cell attachment matrices.

However, as specified above, real-world data matrices typically include false positives, false negatives and missing data. Thus, Eq. (1)

might not hold when using $\mathbf{D}_s$ instead of $\mathbf{G}_s$, and standard phylogenetic inference methods needs to be extended to model error rates and missing data (see, e.g., [19,20]). Notice that the formulation of an analogous problem for longitudinal bulk sequencing data was introduced in [27].

Weighted likelihood function for the grid search formulation of LACE. Here, we describe the case in which a parameter grid search is employed for false positive rates $\{\alpha_s\}_{s=1}^{y}$ and false negative rates $\{\beta_s\}_{s=1}^{y}$, including the specific case of single values provided as input (please refer to the SI for the formulation of LACE when a Gaussian distribution is assumed as prior for the error rates).

LACE employs a number of assumptions on the evolutionary processes underlying the generation of single-cell mutational profiles, which are detailed in the SI, in addition to the thorough description of the statistical model (the probabilistic graphical model of LACE is depicted in Supplementary Figure 1). Thanks to such assumptions, the problem of maximizing the posterior probability is equivalent to the maximization of the weighted log-likelihood objective function defined as follows:

$$\ln(P(\{\mathbf{D}_s\}_{s=1}^{y}|\{\mathbf{G}_s\}_{s=1}^{y},\{\alpha_s\}_{s=1}^{y},\{\beta_s\}_{s=1}^{y})) = \sum_{s=1}^{y} w_s \ln(P(\mathbf{D}_s|\mathbf{G}_s,\alpha_s,\beta_s)), \quad (2)$$

where w_s are *weights* aimed at modeling possible idiosyncrasies of multiple longitudinal experiments, e.g., due to possible differences in sequencing quality and/or in the number of sampled cells. The definition of a weighted likelihood function allows us to explicitly account for experimental and technological differences among experiments collected at distinct time points and represents one of the major novelties of our approach. The choice of appropriate weights is problem- and data-specific and can benefit from a broad literature in statistical inference (see, e.g., [80]). In general, uniform weights would bias the solution toward datasets with larger sample sizes. For this reason, if no prior is available on the quality of the single experiments, we suggest as default weight for the sth dataset composed by n_s cells: $w_s = (1 - \frac{n_s}{n_T})/(y-1)$, where n_T is the total number of cells of all the experiments, and $y \geq 2$ is the number of experiments.

To compute every terms of Eq. (2) we employ the following formula:

$$P(\mathbf{D}_s|\mathbf{G}_s,\alpha_s,\beta_s) = \prod_{i=1}^{n_s}\prod_{j=1}^{\tilde{m}} P(d_{i,j}^s|g_{i,j}^s,\alpha_s,\beta_s), \quad (3)$$

where $d_{i,j}^s$ and $g_{i,j}^s$ are elements of $\mathbf{D}_s$ and $\mathbf{G}_s$, and:

$$P(d_{i,j}^s|g_{i,j}^s,\alpha_s,\beta_s) = \begin{cases} \alpha_s, & if \quad d_{i,j}^s = 1 \quad and \quad g_{i,j}^s = 0, \\ 1-\alpha_s, & if \quad d_{i,j}^s = 1 \quad and \quad g_{i,j}^s = 1, \\ \beta_s, & if \quad d_{i,j}^s = 0 \quad and \quad g_{i,j}^s = 1, \\ 1-\beta_s, & if \quad d_{i,j}^s = 0 \quad and \quad g_{i,j}^s = 0, \\ 1, & if \quad d_{i,j}^s = \text{NA}, \end{cases} \quad (4)$$

with α_s being the false positive rate of the sth experiment, β_s the false negative rate of the sth experiment, and NA label missing entries. Notice that the values of α_s and β_s might be even extremely different among datasets – e.g., due to technological features of the experiments, i.e., LACE can explicitly model different error rates (please refer to Supplementary Table 1 for a summary of the notation).

Search scheme. LACE's output model is composed by the following components: the phylogenetic matrix $\mathbf{B}$, the cell attachment matrices $\{\mathbf{C}_s\}_{s=1}^{y}$, the expected genotype matrices $\{\mathbf{G}_s\}_{s=1}^{y}$ and the estimated error rates $\{\alpha_s\}_{s=1}^{y}$ and $\{\beta_s\}_{s=1}^{y}$. The search space of the possible solutions is huge, in fact given $\tilde{m}$ mutations, it includes a discrete term of dimension $\frac{(2\tilde{m}-3)!}{(\tilde{m}-2)!2^{\tilde{m}-2}}$ for $\mathbf{B}$, another discrete term of dimension $\prod_{s=1}^{y} n_s^{\tilde{m}}$ for $\mathbf{C}_s$, times the dimension of the error rates grid search.

As an exhaustive search can be achieved only for very small models, LACE employs a Markov Chain Monte Carlo (MCMC) scheme via a Metropolis–Hastings algorithm, to find the model that maximizes the weighted likelihood function defined in (2), for each combination of error rates included in the grid search. In particular, the MCMC includes three ergodic moves on $\mathbf{B}$: (*i*) pairwise node relabeling (default probability 0.55), (*ii*) prune and reattach of a single node and its descendants to one of its predecessors (default probability 0.40), (*iii*) full node relabeling (default probability 0.05). For each proposed configuration $\mathbf{B}'$, we find the maximum likelihood $\{\mathbf{C}'_s\}_{s=1}^{y}$ via an exhaustive search and then, by evaluating the products between $\mathbf{B}'$ and $\{\mathbf{C}'_s\}_{s=1}^{y}$, we obtain the proposed expected genotype matrices, i.e., $\{\mathbf{G}'\}_{s=1}^{y}$.

Thus, the acceptance ratio ρ is given by:

$$\rho_{\{\mathbf{G}'_s\}_{s=1}^{y}} = \min\left\{\left(\frac{P(\{\mathbf{D}_s\}_{s=1}^{y}|\{\mathbf{G}'_s\}_{s=1}^{y},\{\alpha_s\}_{s=1}^{y},\{\beta_s\}_{s=1}^{y})}{P(\{\mathbf{D}_s\}_{s=1}^{y}|\{\mathbf{G}_s\}_{s=1}^{y},\{\alpha_s\}_{s=1}^{y},\{\beta_s\}_{s=1}^{y})}\right)^{1/T}, 1\right\}, \quad (5)$$

where T is a learning rate parameter, which could be used to speed up convergence as proposed in [19] (default value $T = 1$). With proper move probabilities, acceptance ratio and an infinite number of moves, the MCMC is ensured to converge to the maximum weighted likelihood solution [19,81] (see the Supplementary Figures 2 and 3 for the MCMC convergence test and the computation time analysis).

LACE finally runs multiple parallel MCMC searches with the fixed combinations of error rates included in the grid search and returns the model that maximizes the weighted likelihood function in (2) and which is marked with $\hat{}$ from now on. Notice that, as there might be multiple models with the same maximum weighted likelihood, in this case LACE returns all equivalent models as output (please refer to the SI for the pseudocode of the algorithm and for the alternative formulation in which the learning of the error rates is included in the MCMC search).

Grouping genotypes via mutation distance. LACE also allows one to assess the presence of groups/clusters of mutations with similar co-occurrence patterns in the expected genotype matrix, after the inference. Such groups might likely indicate (sub)clones and allow one to provide a coarse-grained resolution to the evolution analysis, and might be particularly useful when a reliable driver identification is not possible (see, e.g., the analysis shown in Fig. 4).

Given the concatenation $\hat{\mathbf{G}}$ of the maximum weighted likelihood expected genotype matrices $\{\hat{\mathbf{G}}_s\}_{s=1}^{y}$, the Euclidean distance between two variants i and j (i.e., columns of $\hat{\mathbf{G}}$), is given by:

$$\mathcal{E}(i,j) = \left(\sum_{k=1}^{n}(\hat{g}_{ki} - \hat{g}_{kj})^2\right)^{1/2}, \quad (6)$$

where n is the number of single cells in all experiments.

$\mathcal{E}(i,j)$ can be then employed with standard clustering methods to identify groups of co-occurring mutations in the expected genotype matrix $\hat{\mathbf{G}}$ and indicating candidate (sub)clones.

4.3. Outputs

Given the maximum weighted likelihood solution, LACE provides as output:

- the single-cell longitudinal clonal tree $\hat{\mathcal{T}}$, i.e., a node-weighted Steiner tree [82], where each vertex represents a specific (sub)clone in a given time point and its weight quantifies the prevalence. The edges can represent either parental relations (i.e., linking a given genotype with its offspring(s)) or persistence relations (i.e., linking the same genotype at different time points). A formal definition of the single-cell longitudinal clonal tree returned by LACE is provided in the SI.
- The perfect phylogenetic matrix: $\hat{\mathbf{B}}$.
- The cell attachment matrices for each time point: $\{\hat{\mathbf{C}}_s\}_{s=1}^{y}$.
- The expected genotype matrices for each time point: $\{\hat{\mathbf{G}}_s\}_{s=1}^{y}$.
- The error rates of the maximum likelihood solution for each time point: $\{\hat{\boldsymbol{\alpha}}_s\}_{s=1}^{y}$ and $\{\hat{\boldsymbol{\beta}}_s\}_{s=1}^{y}$.
- The matrix of the pairwise distances among variants, computed via (6).

4.4. Real-world datasets description

Dataset #1 – *Longitudinal scRNA-seq dataset from PDXs of BRAF-mutant melanomas.* We applied LACE to a longitudinal scRNA-seq dataset originally analyzed in [37]. In the study, a number of PDXs were derived from $\text{BRAF}^{\text{V600E/K}}$ mutant melanoma patients and were treated with concurrent BRAF/MEK-inhibition. In our analysis, we selected PDX MEL006, for which four temporally-ordered scRNA-seq datasets are available: (*i*) pre-treatment, (*ii*) after 4 days of treatment, (*iii*) after 28 days of treatment, (*iv*) after 57 days of treatment. In the study, whole transcriptome amplification was made with a modified Smart-seq2 protocol and libraries preparation were performed using the Nextera XT Illumina kit. Samples were sequenced on the Illumina NextSeq 500 platform, by using 75bp single-end reads. Low-quality cells were filtered-out based on library size, number of genes expressed per cell, ERCCs, house-keeping gene expression and mitochondrial DNA reads, and a total of 674 cells was finally included in the dataset [37].

We applied further filters to remove cells displaying a fraction of counts on mitochondrial genes larger than 20% and cells displaying outlier values with respect to library size. As a result, we selected 475 single cells for downstream analysis.

In order to call SNVs and indels from such dataset, we employed the GATK Best Practices [38], which are proven to be effective even with single-cell data [75] (see the SI for a detailed description of the GATK pipeline). A VCF file including 272674 unique variants was generated and subsequently annotated with Annovar [83].

A first filtering step was applied to discard low-quality or non functional variants. First, synonymous and unknown mutations were removed (196320 unique mutations left); second, variants observed in less than two reads in each single cell were filtered-out (195931 unique mutations left); third, we employed a threshold of 1% on minor allele frequency, to remove possible germline mutations, as no normal tissue was included in the study (191599 unique mutations left).

A further filtering step was then employed to identify a list of putative drivers to be used in LACE. We first selected the mutations showing a frequency greater than 5% in at least one time point (595 unique mutations left). We then marked as missing entries (i.e., NA) the variants in a position with coverage lower than 3, as they might be miscalled due to gene expression down-regulation. We kept the mutations displaying less than 40% of missing data on all time points (151 unique mutations left), a median coverage larger than 10 and a median alternative read count larger than 4 (82 unique mutations left). As a final step, we manually curated the list of 82 remaining variants, to verify the possible presence of errors due to amplification (i.e., strand slippages) or alignment artifacts.

As a result, we selected the following 6 candidate drivers to be provided as input to LACE: ARPC2 (chr2: 218249894, C>T, nonsynonymous substitution), CCT8 (chr21: 29063389, G>A, nonsyn.), COL1A2 (chr7: 94422978, C>A, nonsyn.), HNRNPC (chr14: 21211843, C>T, nonsyn.), PRAME (chr22: 22551005, T>A, stop-gain), RPL5 (chr1: 92837514, C>G, nonsyn.). The oncoprint including the mutational profiles of the single cells is displayed in Supplementary Fig. 9.

Dataset #2 – *Longitudinal targeted scDNA-seq dataset from PDXs of triple-negative breast tumors.* LACE was applied to two scDNA-seq datasets from PDXs of triple-negative breast tumors from [39], namely samples SA501 and SA494. The analysis of sample SA501 is presented in the main text (Fig. 4), whereas that of sample SA494 is presented in the SI (Supplementary Figure 17).

For sample SA501, we processed the allelic frequency matrix including a panel of 10 germline and 45 somatic variants selected by the authors, on the three engraftment passages for which single-cell deep re-sequencing was performed (i.e., X1, X2 and X4).

In particular, we excluded germline variants and defined a binary mutational profile, by considering each somatic SNV in each single cell as: (*i*) present (1) if the allelic frequency ≥ 0.10, (*ii*) absent (0) if the allelic frequency is ≤ 0.01, (*iii*) missing entry (NA) if the allelic frequency is > 0.01 and < 0.10 or if already marked as uninformative (< 25 mapped reads).

Finally, we filtered out all the variants displaying a value of missing data (NA) ≥ 0.15 in all time points, to focus on highly confident SNVs. As a result, the mutational profile matrix provided as input to LACE includes $\tilde{m} = 20$ somatic variants and $n = 27, 36, 27$ single cells for time points $t_0 = \text{X1}$, $t_1 = \text{X2}$ and $t_2 = \text{X4}$, respectively (Fig. 4B).

For all details regarding the analysis of sample SA494, please refer to Section 3.2 of the SI.

4.5. Simulations

To generate synthetic datasets we employed the tool from [43], which simulates a branching process modeling the population dynamics of cancer subpopulations, characterized by the accumulation of random mutations, which can either be drivers – i.e., inducing a certain proliferative advantage –, or passengers – i.e., with no effect. The simulator eventually returns the list of existing genotypes (including both passengers and drivers) and the relative prevalence at any time point, where a single time step corresponds to a generation (i.e., a replication event — see the SI for further details and the parameter settings).

We selected 15 simulation scenarios in which a number of drivers between 7 and 15 was observed, from which we sampled a large number of independent longitudinal single-cell mutational profile datasets including the drivers only, on three distinct time points, i.e., $t = 50, 150$ and 300. Single cells were randomly sampled with a probability proportional to cellular prevalence and we finally inflated the resulting binary mutational profiles with various rates of: false positives, α, false negatives, β and missing entries γ.

CRediT authorship contribution statement

Daniele Ramazzotti: Designed the approach, Defined the method, Implemented it, Performed the simulations, Executed the experimental data analysis pipeline, Analyzed the data and interpreted the results, Supervised the study. **Fabrizio Angaroni:** Designed the approach, Defined the method, Implemented it, Analyzed the data and interpreted the results. **Davide Maspero:** Designed the approach, Defined the method, Implemented it, Executed the experimental data analysis pipeline, Analyzed the data and interpreted the results. **Gianluca Ascolani:** Executed the experimental data analysis pipeline. **Isabella Castiglioni:** Analyzed the data and interpreted the results. **Rocco Piazza:** Executed the experimental data analysis pipeline, Analyzed the data and interpreted the results. **Marco Antoniotti:** Analyzed the data and interpreted the results. **Alex Graudenzi:** Designed the approach, Defined the method, Implemented it, Performed the simulations, Executed the experimental data analysis pipeline, Analyzed the data and interpreted the results, Supervised the study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

LACE is available as an open source R tool on Bioconductor, at this link: https://bioconductor.org/packages/release/bioc/html/LACE.html and at the GitHub repository: https://github.com/BIMIB-DISCo/LACE.

The scRNA-seq dataset used in the first case study (PDX MEL006) was downloaded from GEO: https://www.ncbi.nlm.nih.gov/geo/, accession code: GSE116237. The targeted scDNA-seq datasets used in the

second case study (PDXs SA501 and SA494) were retrieved from the Supplementary Information of the original article (Eirew et al., 2015). The source code used to replicate all our analyses, including synthetic and real datasets, is available at this link: https://github.com/BIMIB-DISCo/LACE-UTILITIES.

Acknowledgments

This work was partially supported by the Elixir Italian Chapter and the SysBioNet project, a Ministero dell'Istruzione, dell'Università e della Ricerca, Italy initiative for the Italian Roadmap of European Strategy Forum on Research Infrastructures and by the AIRC-IG grant 22082. Support was also provided by the CRUK, United Kingdom/AIRC, Italy Accelerator Award #22790, "Single-cell Cancer Evolution in the Clinic". We thank Giulio Caravagna, Chiara Damiani, Francesco Craighero and Lucrezia Patruno for helpful discussions. All authors approved the final version of the manuscript.

Appendix A. Supplementary data

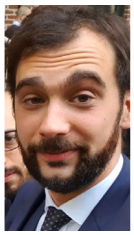
Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jocs.2021.101523.

References

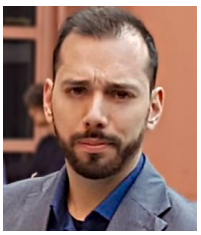
[1] E. Shapiro, T. Biezuner, S. Linnarsson, Single-cell sequencing-based technologies will revolutionize whole-organism science, Nature Rev. Genet. 14 (9) (2013) 618.
[2] Y. Wang, N.E. Navin, Advances and applications of single-cell sequencing technologies, Mol. Cell 58 (4) (2015) 598–609.
[3] M. Efremova, S.A. Teichmann, Computational methods for single-cell omics across modalities, Nature Methods 17 (1) (2020) 14–17.
[4] R.J. Gillies, D. Verduzco, R.A. Gatenby, Evolutionary dynamics of carcinogenesis and why targeted therapy does not work, Nat. Rev. Cancer 12 (7) (2012) 487–493.
[5] R.A. Burrell, N. McGranahan, J. Bartek, C. Swanton, The causes and consequences of genetic heterogeneity in cancer evolution, Nature 501 (7467) (2013) 338–345.
[6] B. Vogelstein, et al., Cancer genome landscapes, Science 339 (6127) (2013) 1546–1558.
[7] A. Sottoriva, et al., Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics, Proc. Natl. Acad. Sci. 110 (10) (2013) 4009–4014.
[8] N.E. Navin, The first five years of single-cell cancer genomics and beyond, Genome Res. 25 (10) (2015) 1499–1507.
[9] H.J. Jackson, S. Rafiq, R.J. Brentjens, Driving CAR T-cells forward, Nat. Rev. Clin. Oncol. 13 (6) (2016) 370.
[10] A. Goodspeed, L.M. Heiser, J.W. Gray, J.C. Costello, Tumor-derived cell lines as molecular models of cancer pharmacogenomics, Mol. Cancer Res. 14 (1) (2016) 3–13.
[11] H. Clevers, Modeling development and disease with organoids, Cell 165 (7) (2016) 1586–1597.
[12] S.F. Roerink, et al., Intra-tumour diversification in colorectal cancer at the single-cell level, Nature 556 (7702) (2018) 457–462.
[13] A.P. Patel, et al., Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma, Science 344 (6190) (2014) 1396–1401.
[14] D. Lähnemann, et al., Eleven grand challenges in single-cell data science, Genome Biol. 21 (1) (2020) 1–35.
[15] M. Pellegrino, et al., High-throughput single-cell DNA sequencing of acute myeloid leukemia tumors with droplet microfluidics, Genome Res. 28 (9) (2018) 1345–1352.
[16] L. Zhang, et al., Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan, Proc. Natl. Acad. Sci. 116 (18) (2019) 9014–9019.
[17] I.C. Macaulay, et al., G&T-seq: parallel sequencing of single-cell genomes and transcriptomes, Nature Methods 12 (6) (2015) 519.
[18] A.S. Nam, et al., Somatic mutations and cell identity linked by genotyping of transcriptomes, Nature 571 (7765) (2019) 355–360.
[19] K. Jahn, J. Kuipers, N. Beerenwinkel, Tree inference for single-cell data, Genome Biol. 17 (1) (2016) 1.
[20] E.M. Ross, F. Markowetz, OncoNEM: inferring tumor evolution from single-cell sequencing data, Genome Biol. 17 (1) (2016) 1.
[21] H. Zafar, A. Tzen, N. Navin, K. Chen, L. Nakhleh, SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models, Genome Biol. 18 (1) (2017) 178.
[22] D. Ramazzotti, A. Graudenzi, L. De Sano, M. Antoniotti, G. Caravagna, Learning mutational graphs of individual tumour evolution from single-cell and multi-region sequencing data, BMC Bioinformatics 20 (1) (2019) 210.
[23] H. Zafar, N. Navin, K. Chen, L. Nakhleh, SiCLonefit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data, Genome Res. 29 (11) (2019) 1847–1859.
[24] K. Morita, et al., Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics, BioRxiv (2020).
[25] G. Siravegna, et al., Radiologic and genomic evolution of individual metastases during HER2 blockade in colorectal cancer, Cancer Cell 34 (1) (2018) 148–162.
[26] K.H. Khan, et al., Longitudinal liquid biopsy and mathematical modeling of clonal evolution forecast time to treatment failure in the PROSPECT-c phase II colorectal cancer clinical trial, Cancer Discov. 8 (10) (2018) 1270–1285.
[27] M.A. Myers, G. Satas, B.J. Raphael, CALDER: Inferring phylogenetic trees from longitudinal tumor samples, Cell Systems 8 (6) (2019) 514 – 522.e5.
[28] J.M. Alves, T. Prieto, D. Posada, Multiregional tumor trees are not phylogenies, Trends Cancer 3 (8) (2017) 546–550.
[29] S.C. Hicks, F.W. Townes, M. Teng, R.A. Irizarry, Missing data and technical variability in single-cell RNA-sequencing experiments, Biostatistics 19 (4) (2017) 562–578.
[30] R. Schachtner, G. Pöppel, A. Tomé, E. Lang, From binary NMF to variational Bayes NMF: A probabilistic approach, in: Non-Negative Matrix Factorization Techniques, Springer, 2016, pp. 1–48.
[31] D. Ramazzotti, et al., VERSO: a comprehensive framework for the inference of robust phylogenies and the quantification of intra-host genomic diversity of viral samples, Patterns 2 (3) (2021) 100212.
[32] M.D. Luecken, F.J. Theis, Current best practices in single-cell RNA-seq analysis: a tutorial, Mol. Syst. Biol. 15 (6) (2019).
[33] F. Liu, et al., Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data, Genome Biol. 20 (1) (2019) 1–15.
[34] H.-J. Bandelt, P. Forster, A. Röhl, Median-joining networks for inferring intraspecific phylogenies, Mol. Biol. Evol. 16 (1) (1999) 37–48.
[35] S. Salehi, et al., ddClone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data, Genome Biol. 18 (1) (2017) 44.
[36] I. Dagogo-Jack, A.T. Shaw, Tumour heterogeneity and resistance to cancer therapies, Nat. Rev. Clin. Oncol. 15 (2) (2018) 81–94.
[37] F. Rambow, et al., Toward minimal residual disease-directed therapy in melanoma, Cell 174 (4) (2018) 843–855.
[38] M.A. DePristo, et al., A framework for variation discovery and genotyping using next-generation DNA sequencing data, Nature Genet. 43 (5) (2011) 491.
[39] P. Eirew, et al., Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution, Nature 518 (7539) (2015) 422–426.
[40] M. Kimura, The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations, Genetics 61 (4) (1969) 893.
[41] M. El-Kebir, G. Satas, L. Oesper, B.J. Raphael, Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures, Cell Syst. 3 (1) (2016) 43–53.
[42] L. Hubert, P. Arabie, Comparing partitions, J. Classification 2 (1) (1985) 193–218.
[43] M. El-Kebir, G. Satas, B.J. Raphael, Inferring parsimonious migration histories for metastatic cancers, Nature Genet. 50 (5) (2018) 718.
[44] X. Xu, et al., Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor, Cell 148 (5) (2012) 886–895.
[45] Y. Hou, et al., Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm, Cell 148 (5) (2012) 873–885.
[46] Y. Li, et al., Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer, GigaScience 1 (1) (2012) 12.
[47] N.E. Navin, Cancer genomics: one cell at a time, Genome Biol. 15 (8) (2014) 452.
[48] M. Gerlinger, et al., Intratumor heterogeneity and branched evolution revealed by multiregion sequencing, N. Engl. J. Med. 366 (10) (2012) 883–892.
[49] A. Graudenzi, G. Caravagna, G. De Matteis, M. Antoniotti, Investigating the relation between stochastic differentiation, homeostasis and clonal expansion in intestinal crypts via multiscale modeling, PLoS One 9 (5) (2014) e97272.
[50] B. Waclaw, et al., A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity, Nature 525 (7568) (2015) 261–264.
[51] K. Chkhaidze, et al., Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data, PLoS Comput. Biol. 15 (7) (2019) e1007243.
[52] M.A. Smith, et al., E-scape: interactive visualization of single-cell phylogenetics and cancer evolution, Nature Methods 14 (6) (2017) 549.

[53] L. Haghverdi, F. Buettner, F.J. Theis, Diffusion maps for high-dimensional single-cell analysis of differentiation data, Bioinformatics 31 (18) (2015) 2989–2998.
[54] M. Kashani-Sabet, et al., A multi-marker assay to distinguish malignant melanomas from benign nevi, Proc. Natl. Acad. Sci. 106 (15) (2009) 6268–6272.
[55] D. Orlando, et al., Adoptive immunotherapy using PRAME-specific T cells in medulloblastoma, Cancer Res. 78 (12) (2018) 3337–3349.
[56] J. Pelletier, G. Thomas, S. Volarević, Ribosome biogenesis in cancer: new players and therapeutic avenues, Nat. Rev. Cancer 18 (1) (2018) 51.
[57] B.D. Cholewa, M.C. Pellitteri-Hahn, C.O. Scarlett, N. Ahmad, Large-scale label-free comparative proteomics analysis of polo-like kinase 1 inhibition via the small-molecule inhibitor BI 6727 (volasertib) in BRAFV600e mutant melanoma cells, J. Proteome Res. 13 (11) (2014) 5041–5050.
[58] Y. Koga, et al., Genome-wide screen of promoter methylation identifies novel markers in melanoma, Genome Res. 19 (8) (2009) 1462–1470.
[59] J. Li, Y. Ding, A. Li, Identification of COL1A1 and COL1A2 as candidate prognostic factors in gastric cancer, World J. Surg. Oncol. 14 (1) (2016) 297.
[60] X. Huang, et al., Chaperonin containing TCP 1, subunit 8 (CCT 8) is upregulated in hepatocellular carcinoma and promotes HCC proliferation, Apmis 122 (11) (2014) 1070–1079.
[61] F.A. Wolf, P. Angerer, F.J. Theis, SCANPY: large-scale single-cell gene expression data analysis, Genome Biol. 19 (1) (2018) 15.
[62] A. Roth, et al., PyClone: statistical inference of clonal population structure in cancer, Nature Methods 11 (4) (2014) 396–398.
[63] F. Ronquist, et al., MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space, Syst. Biol. 61 (3) (2012) 539–542.
[64] P.C. Nowell, The clonal evolution of tumor cell populations, Science 194 (4260) (1976) 23–28.
[65] G. Caravagna, et al., Algorithmic methods to infer the evolutionary trajectories in cancer progression, Proc. Natl. Acad. Sci. 113 (28) (2016) E4025–E4034.
[66] L.S. Ludwig, et al., Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics, Cell 176 (6) (2019) 1325–1339.
[67] S.M. Lewis, et al., Spatial omics and multiplexed imaging to explore cancer biology, Nature Methods (2021) 1–16.
[68] A. Lomakin, et al., Spatial genomics maps the structure, character and evolution of cancer clones, BioRxiv (2021).
[69] A. Perl, M. Carroll, BCR-ABL kinase is dead; long live the CML stem cell, J. Clin. Invest. 121 (1) (2011) 22–25.
[70] R. Kinstrie, et al., CD93 is expressed on chronic myeloid leukemia stem cells and identifies a quiescent population which persists after tyrosine kinase inhibitor therapy, Leukemia (2020) 1–13.
[71] M. El-Kebir, SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error, Bioinformatics 34 (17) (2018) i671–i679.
[72] P. Bonizzoni, S. Ciccolella, G. Della Vedova, M.S. Gomez, Does relaxing the infinite sites assumption give better tumor phylogenies? An ILP-based comparative approach, IEEE/ACM Trans. Comput. Biol. Bioinform. (2018).
[73] S. Malikic, K. Jahn, J. Kuipers, S.C. Sahinalp, N. Beerenwinkel, Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data, Nature Commun. 10 (1) (2019) 2750.
[74] G. Caravagna, et al., Detecting repeated cancer evolution from multi-region tumor sequencing data, Nature Methods 15 (9) (2018) 707.
[75] P.M. Schnepp, M. Chen, E.T. Keller, X. Zhou, SNV identification from single-cell RNA sequencing data, Hum. Mol. Gen. (2019).
[76] J.K. Kim, A.A. Kolodziejczyk, T. Ilicic, S.A. Teichmann, J.C. Marioni, Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression, Nature Commun. 6 (2015) 8687.
[77] L. Patruno, et al., A review of computational strategies for denoising and imputation of single-cell transcriptomic data, Brief. Bioinform. 22 (4) (2021) bbaa222.
[78] D. Gusfield, Efficient algorithms for inferring evolutionary trees, Networks 21 (1) (1991) 19–28.
[79] M. El-Kebir, L. Oesper, H. Acheson-Field, B.J. Raphael, Reconstruction of clonal trees and tumor composition from multi-sample sequencing data, Bioinformatics 31 (12) (2015) i62–70.
[80] F. Hu, J.V. Zidek, The relevance weighted likelihood with applications, in: Empirical Bayes and Likelihood Inference, Springer, 2001, pp. 211–235.
[81] R. Tibshirani, T. Hastie, Local likelihood estimation, J. Amer. Statist. Assoc. 82 (398) (1987) 559–567.
[82] P.N. Klein, R. Ravi, A nearly best-possible approximation algorithm for node-weighted steiner trees, J. Algorithms 19 (1) (1995) 104–115.
[83] K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, Nucleic Acids Res. 38 (16) (2010) e164.

Daniele Ramazzotti received his Ph.D. in Computer Science at the University of Milano-Bicocca in February 2016. Since then, he has been Research Associate at several important Research Centers and Universities such as Stanford University where he spent more than 4 years. He is currently a Postdoctoral Research Fellow at the School of Medicine and Surgery at the University of Milano-Bicocca. His current research interests involve Bioinformatics with specific focus on Cancer Evolution, Statistics, Bayesian Learning, Machine Learning, Algorithmics and Theories of Causality.



Fabrizio Angaroni is a Postdoc Research Fellow at the Department of Informatics, Systems and Communication of the University of Milano-Bicocca. He holds a Ph.D. in Physics and Astrophysics from the University of Insubria. His current research is devoted to mathematical modeling and the development of statistical methods to analyze omics data. His main research topics are population genetics, control theory applied to biological systems, and viral and cancer evolution.



Davide Maspero graduated in Industrial Biotechnology and, currently, he is a Ph.D. candidate at the Department of Computer Science at the University of Milano-Bicocca. His main research interests are devoted to the development of new computational frameworks to analyze and integrate multiple omics data generated from complex biological systems, in order to depict their emerging properties and their spatiotemporal evolution.

Gianluca Ascolani is a postdoctoral researcher at the Dept. of Informatics, Systems and Communication of the University of Milan-Bicocca. He received his Ph.D. in physics in 2010 at the Center for Non-linear Science in the University of North Texas, US. He is author of numerous publications on international journals, organizer and program committee member of several international conferences. He works on complex systems, ecological evolutionary systems and physics of cancer, and he has developed various mathematical models on cancer interactions and evolution.

Isabella Castiglioni is a Full Professor of the Dept. of Physics "Giuseppe Occhialini" of the Univ. of Milan-Bicocca. Formerly, she was a Researcher at the Institute of Molecular Bioimaging and Physiology of the National Research Council (IBFM-CNR), where she was also a Member of the Council of Institute. She obtained scientific training in the field of medical physics, and molecular imaging and diagnostics, completed by a management training. She has 150+ scientific publications on indexed journals and conference proceedings, and she has coordinated numerous research projects funded on competitive calls and many research projects in agreement with IRCCS and Hospitals. Her research interest interests are mainly related to Physics applied to Medicine and Biology, Molecular Bioimaging and Artificial Intelligence, with a particular focus on radiomics, genomics and radiogenomics, in the broad field of (cancer) theranostics.

Rocco Piazza is an Associate Professor of Hematology in the Department of Medicine and Surgery of the University of Milan-Bicocca, Milan, Italy. He also works as clinical hematologist in the hematology unit of the San Gerardo Hospital, Monza, Italy. Before that he was a Research Scientist at the University of Pavia, Italy. He received his Ph.D. in molecular biology in 2002 from the Department of Biochemistry of the University of Pavia. His main research interest is the dissection of the molecular mechanisms responsible for cancer onset and evolution, using both wet cell and molecular biology techniques as well as bioinformatics. He is author of 80+ publications on indexed international journals and he is recipient of several national and international research grants, among them AIRC, PRIN and ERARE.

Marco Antoniotti is a Full Professor of Computer Science in the DISCo of the Università of Milano – Bicocca, Milan, Italy. From 2000 to 2006 Senior Research Scientist of the NYU Courant Bioinformatics Group. Before that he was a Research Scientist at PARADES EEIG. From 1996 to 1997 Post-Doctoral fellow at University of California at Berkeley PATH Institute. He received his Ph.D. in computer science from Courant Institute of Mathematical Sciences of New York University. His main research topics are Bioinformatics and Systems Biology, Simulation, Verification and Language Design Issues in Hybrid Systems. Marco Antoniotti is the author of several journal and conference papers and of several software projects; he has received support for his research from Regione Lombardia, NSF, the Elixir European network, Cancer Research UK and AIRC under the Accelerator Award Scheme and the European Commission under the MCSA and COST actions.

Alex Graudenzi is a Tenured Researcher of the Institute of Bioimaging and Molecular Physiology, Consiglio Nazionale delle Ricerche (IBFM-CNR), Milan. He received his Ph.D. in computational modeling and simulation in 2010. Since then, he has been Research Associate, Assistant Professor and Visiting Scientist at several important Research Centers and Universities. He is author of 60+ publications on indexed international journals and conference proceedings, recipient of several research awards, scientific collaborator in many international projects, co-developer of 15+ tools for computational biology, and organizer and program committee member of numerous international conferences. He works at the boundaries of Bioinformatics, Complex Systems and Artificial Intelligence to investigate the properties of biological systems and, especially, of cancer evolution.

Characterization of viral evolution. Phylogenetic analyses are also commonly applied to monitor and characterize the evolution of epidemics, as in the case of the COVID-19 pandemic [155]. In fact, the phylogenomic analysis of SARS-CoV-2 evolution has benefited from the surge of sequencing data that are first collected and then made publicly available on portals such as GISAID [138], Nextstrain [155] or Cov-Lineages.org [248]. In addition, in the last months, the number of deep sequencing data made available on public databases (e.g., NCBI [160]) has increased, highlighting the benefits provided by those data.

We remark the difference between Sanger sequencing and deep sequencing data. The former returns consensus sequences including the most abundant nucleotide in each genome position. The latter can provide information about the frequency of each variant present in each genome position (i.e., variant frequency – VF). For instance, as one can see in figure 3.1, we can consider as clonal mutations the ones that are common to all of the virions and which are usually included in a consensus sequence (VF $> 50\%$), while as intra-host minor mutations the others. In fact, in each host, there is a mixture of heterogeneous virus subpopulations known as *viral quasispecies* [214], which are supposed to underlie most of the adaptive potential to the immune system response and anti-viral therapies [238].

Interestingly, sequencing data of viral genomes, collected during an epidemic, share similarities with cancer data. In fact, (*i*) we can call mutations by comparing each viral genome with the ancestral one; (*ii*) viral samples are obtained daily, so many of them share the same set of fixed mutations, similarly to a clonal population; (*iii*) the mutational profile reflects the quasispecies heterogeneity that emerged in the hosts. However, the main difference between virus diffusion and cancer progression is that an infected host does not entirely inherit the viral genome after contagion. In particular, the frequency of intra-host minor variants could dramatically change after a few consecutive transmission events, due to founder effects or bottlenecks, as already observed for other viruses (e.g., hepatitis C. [95]). In general, the chance of transmitting a mutation is roughly proportional to its frequency.

For these reasons, clonal and intra-hosts mutations need to be considered differently. In paper P#6, we propose a comprehensive statistical framework, which fully exploits the whole frequency spectrum of viral mutations to infer a robust phylogeny tree and then quantify the intra-host genomic diversity of viral samples. The workflow is called VERSO (Viral Evolution ReconStructiOn), and is divided into two steps, explained below.

VERSO step #1

The first step is based only on clonal (i.e., fixed) mutations (or consensus sequences). Clonal mutations are most likely transmitted in the infection chain, so they accumulate during the pandemic and can be used to infer the phylogenetic model of the virus. Phylo-

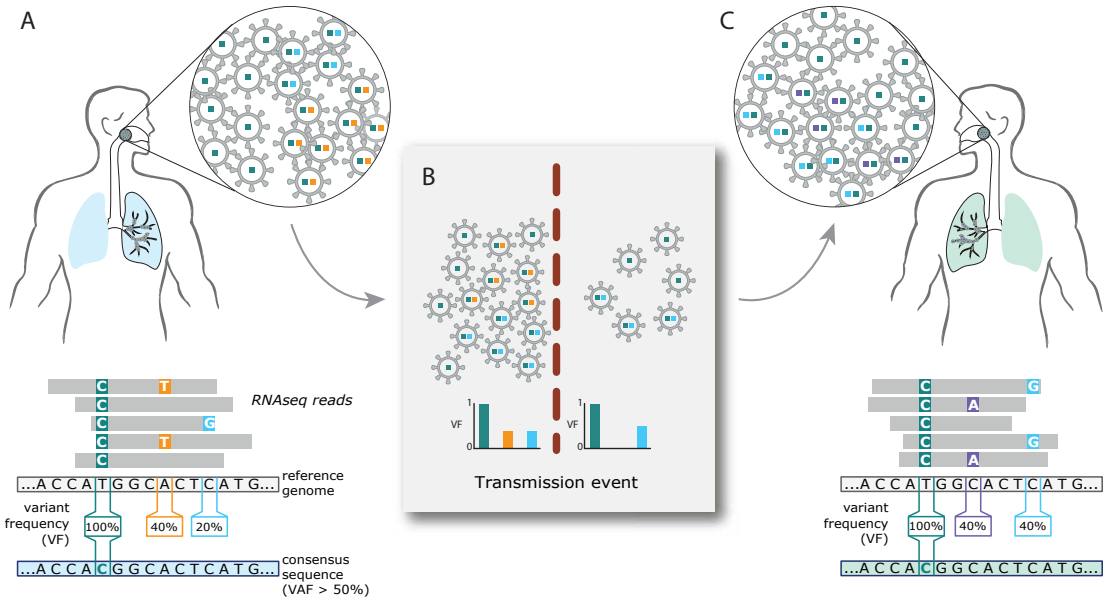


Figure 3.1: Representation of the main assumptions in a viral infection scenario. A) Infected patient carries a mixture of virions with common (green) and specific (orange or light blue) mutations. A deep RNA-seq experiment enables the detection of either clonal and intra-host minor variants. B) Original variant allele frequencies are changed due to founder effects and bottlenecks selection (red line). C) During the infection of a new host, further mutations are generated and selected.

genetic analyses that exploit viral *consensus sequences* are useful for example to provide insights into the environment's selective pressure and they can be helpful to evaluate the diffusion of the principal viral strain among countries, or to pinpoint dangerous variants [155, 249]. However, these methods might produce unreliable results when dealing with noisy data due to sequencing errors or sampling limitations [16]. Moreover, they struggle also in the presence of high number of samples with the same consensus sequence (i.e., polytomies) [29], which are expected at the beginning of the epidemic spread or in limited geographical areas

For these reasons, in the first step of our framework, we applied strategies borrowed from cancer evolution to return more robust results. In particular, we applied a noise-tolerant framework akin to paper P#5, to process binarized clonal variant profiles. Our approach employs perfect phylogeny constraint, assuming the ISA, to correct false-positive and false-negative variants, and to handle missing observations (e.g., due to low coverage). It is also designed to group samples with identical clonal genotype in polytomies, avoiding ungrounded random orderings.

We proved the performance of this inference step by generating an extensive array of simulations, and compared it with two state-of-the-art methods for phylogenetic reconstruction such as IQ-TREE [101] and BEAST 2 [77], which are outperformed in all the experimental scenarios. We also assessed the reliability of the results by applying it on two independent datasets of SARS-CoV-2 samples, obtaining similar phylogenetic trees.

VERSO step #2

In the second step, VERSO leverages the full variant frequency (VF) profile to characterize and visualize intra-host genomic similarity of samples with identical (corrected) clonal genotypes. The idea behind this choice lies in the fact that the transmission of low-frequency variants also involves the transfer of clonal mutations, but the opposite might not be true due e.g., to bottleneck events. Thus, patterns of co-occurrence of minor variants detected in hosts may underlie links of possible infections. Samples were compared after the application of techniques for dimensionality reduction and clustering strategies commonly used in single-cell analysis. At the end of the analysis, samples are projected and connected into a lower dimensional space (e.g., UMAP [159]) where connections may suggest a probable infection chain.

To assess the reliability of our assumption, we exploited real contact tracing information. The results indicate that the samples known to have infected each other were positioned closer in the space generated by our method.

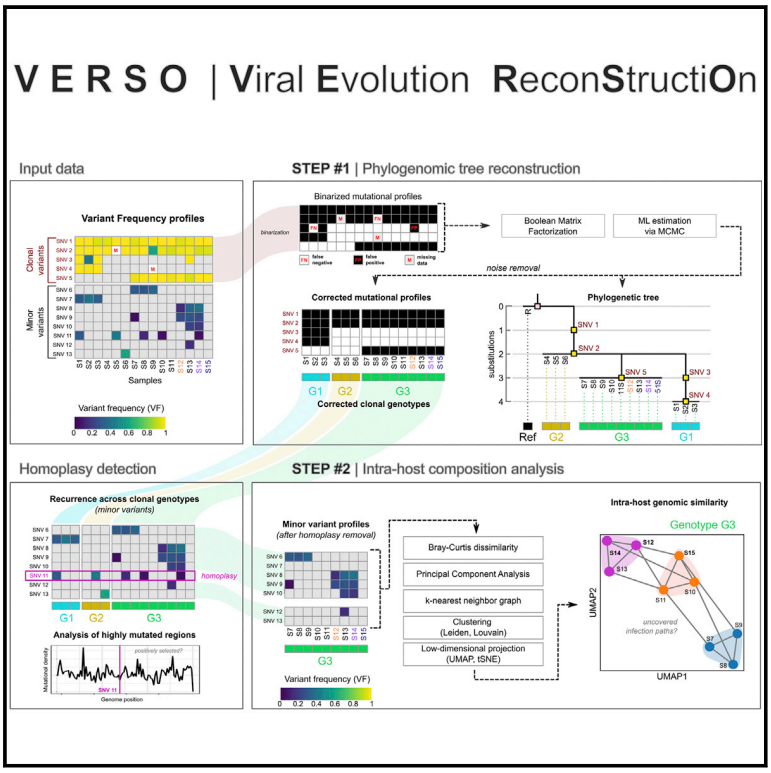
With VERSO we have therefore demonstrated how the use of only consensus sequences is limiting. Instead, deep sequencing data can be analyzed to return more robust results, and information on probable chains of infection, particularly in early phase of the outbreak. Finally, analyses provided by VERSO could be useful to assess the impact of precautionary measures (e.g., lockdowns) or to enhance contact tracing among individuals.

Article

Patterns

VERSO: A comprehensive framework for the inference of robust phylogenies and the quantification of intra-host genomic diversity of viral samples

Graphical abstract



Highlights

- The analysis of raw sequencing data improves the reconstruction of viral evolution
- Our method reconstructs robust phylogenies with noisy data and sampling limitations
- The dissection of intra-host genomic diversity reveals undetected infection chains
- The identification of positively selected variants may drive experimental research

Authors

Daniele Ramazzotti, Fabrizio Angaroni, Davide Maspero, Carlo Gambacorti-Passerini, Marco Antoniotti, Alex Graudenzi, Rocco Piazza

Correspondence

alex.graudenzi@ibfm.cnr.it (A.G.), rocco.piazza@unimib.it (R.P.)

In Brief

The generation of reliable phylogenomic models describing the evolution of SARS-CoV-2 is essential to explain its diffusion and to possibly predict the next evolutionary steps. We introduce a data-science framework that is an improvement on existing methods, by accounting for noise and sampling limitations in sequencing data and by dissecting the intra-host diversity of single samples. The application to large-scale datasets demonstrates that our approach can improve the estimation of SARS-CoV-2 evolution, refine contact tracing, and pinpoint possibly hazardous mutations.

Ramazzotti et al., 2021, Patterns 2, 100212
March 12, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.patter.2021.100212

CellPress

Patterns

CellPress
OPEN ACCESS

Article

VERSO: A comprehensive framework for the inference of robust phylogenies and the quantification of intra-host genomic diversity of viral samples

Daniele Ramazzotti,[1] Fabrizio Angaroni,[2] Davide Maspero,[2,3] Carlo Gambacorti-Passerini,[1] Marco Antoniotti,[2,4] Alex Graudenzi,[3,4,5,6,*] and Rocco Piazza[1,5,*]

[1]Department of Medicine and Surgery, Università degli Studi di Milano-Bicocca, Monza, Italy
[2]Department of Informatics, Systems and Communication, Università degli Studi di Milano-Bicocca, Milan, Italy
[3]Inst. of Molecular Bioimaging and Physiology, Consiglio Nazionale delle Ricerche (IBFM-CNR), Segrate, Milan, Italy
[4]Bicocca Bioinformatics, Biostatistics and Bioimaging Centre – B4, Milan, Italy
[5]Senior author
[6]Lead contact
*Correspondence: alex.graudenzi@ibfm.cnr.it (A.G.), rocco.piazza@unimib.it (R.P.)
https://doi.org/10.1016/j.patter.2021.100212

THE BIGGER PICTURE The gravity of the COVID-19 pandemic has fostered a surge of works analyzing SARS-CoV-2 consensus sequences to reconstruct phylogenomic models of its evolution and diffusion. Yet, such approaches do not account for intra-host genomic diversity and may deliver inaccurate predictions in conditions of noisy data and sampling limitations.

We propose VERSO, a data-science framework for the characterization of viral evolution from sequencing data. By accounting for uncertainty in the data, VERSO delivers robust phylogenies also in conditions of limited sampling and noisy observations. Additionally, the in-depth characterization of the intra-host genomic diversity of samples allows one to identify undetected infection chains and clusters and to intercept variants possibly undergoing positive selection. Accordingly, the joint application of our method and data-driven epidemiological models may deliver a high-precision platform for contact tracing and pathogen surveillance and characterization.

Proof-of-**Concept:** Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

We introduce VERSO, a two-step framework for the characterization of viral evolution from sequencing data of viral genomes, which is an improvement on phylogenomic approaches for consensus sequences. VERSO exploits an efficient algorithmic strategy to return robust phylogenies from clonal variant profiles, also in conditions of sampling limitations. It then leverages variant frequency patterns to characterize the intra-host genomic diversity of samples, revealing undetected infection chains and pinpointing variants likely involved in homoplasies. On simulations, VERSO outperforms state-of-the-art tools for phylogenetic inference. Notably, the application to 6,726 amplicon and RNA sequencing samples refines the estimation of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) evolution, while co-occurrence patterns of minor variants unveil undetected infection paths, which are validated with contact tracing data. Finally, the analysis of SARS-CoV-2 mutational landscape uncovers a temporal increase of overall genomic diversity and highlights variants transiting from minor to clonal state and homoplastic variants, some of which fall on the spike gene. Available at: https://github.com/BIMIB-DISCo/VERSO.

INTRODUCTION

The outbreak of coronavirus disease 2019 (COVID-19), which started in late 2019 in Wuhan (China)[1,2] and was declared a pandemic by the World Health Organization, is fueling the publication of an increasing number of studies aimed at exploiting the information provided by the viral genome of severe acute respiratory syndrome-coronavirus 2 (SARS-CoV-2) virus to identify its

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

proximal origin, characterize the mode and timing of its evolution, as well as to define descriptive and predictive models of geographical spread and evaluate the related clinical impact.[3–5] As a matter of fact, the mutations that rapidly accumulate in the viral genome[6] can be used to track the evolution of the virus and, accordingly, unravel the viral infection network.[7,8]

At the time of this writing, numerous independent laboratories around the world are isolating and sequencing SARS-CoV-2 samples and depositing them on public databases (e.g., GISAID[9]) whose data are accessible via the Nextstrain portal.[10] Such data can be employed to estimate models from genomic epidemiology and may serve, for instance, to estimate the proportion of undetected infected people by uncovering cryptic transmissions, as well as to predict likely trends in the number of infected, hospitalized, dead, and recovered people.[11–13]

More in detail, most studies employ phylogenomic approaches that process consensus sequences, which represent the dominant virus lineage within each infected host. A growing plethora of methods for phylogenomic reconstruction is available to this end, all relying on different algorithmic frameworks, including distance-matrix, maximum parsimony, maximum likelihood, or Bayesian inference, with various substitution models and distinct evolutionary assumptions (see, e.g., Refs.[10,14–22]). However, while such methods have repeatedly proven effective in unraveling the main patterns of evolution of viral genomes with respect to many different diseases, including SARS-CoV-2,[10,23–25] at least two issues can be raised.

First, most phylogenomics methods might produce unreliable results when dealing with noisy data, for instance due to sequencing issues, or with data collected with significant sampling limitations,[14,26,27] as witnessed for most countries during the epidemics.[28,29]

Second, most methods do not consider the key information on intra-host minor variants (also referred to as minority variants or intra-host single nucleotide variants), which can be retrieved from whole-genome deep sequencing raw data and might be essential to improve the characterization of the infection dynamics and to pinpoint positively selected variants.[30–32] Due to the high replication, mutation, and recombination rates of RNA viruses, subpopulations of mutant viruses, also known as viral quasispecies,[30] typically emerge and coexist within single hosts, and are supposed to underlie most of the adaptive potential to the immune system response and to anti-viral therapies.[31,33,34] In this regard, many recent studies highlighted the noteworthy amount of intra-host genomic diversity in SARS-CoV-2 samples,[35–43] similarly to what has already been observed in many distinct infectious diseases.[8,32,44–48]

Here, we introduce VERSO (viral evolution reconstruction), a new comprehensive framework for the inference of high-resolution models of viral evolution from raw sequencing data of viral genomes (see Figure 1). VERSO includes two consecutive algorithmic steps.

Step #1: robust phylogenomic inference from clonal variant profiles

VERSO first employs a probabilistic noise-tolerant framework to process binarized clonal variant profiles (or, alternatively, consensus sequences), to return a robust phylogenetic model also in conditions of sampling limitations and sequencing issues. By adapting algorithmic strategies widely employed in cancer evolution analysis,[49–52] VERSO is able to correct false-positive and false-negative variants, can manage missing observations due to low coverage, and is designed to group samples with identical (corrected) clonal genotype in polytomies, avoiding ungrounded arbitrary orderings. As a result, the accurate and robust phylogenomic models produced by VERSO may be used to improve the parameter estimation of epidemiological models, which typically rely on limited and inhomogeneous data.[11,29] Notice that this step can be executed independently from step #2; for instance, in case raw sequencing data are not available.

Homoplasy detection (clonal variants)

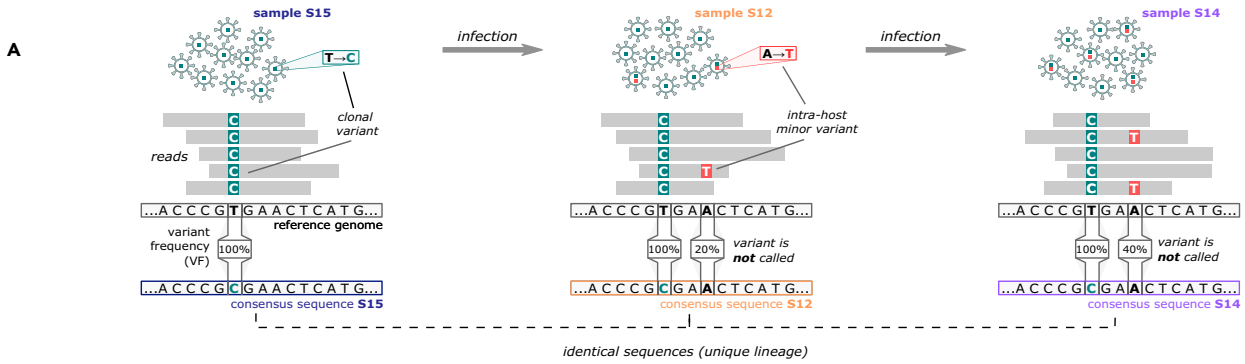
The first step of VERSO allows one to identify clonal mutations that might be involved in reticulation events[53,54] and, in particular, in homoplasies, possibly due to positive selection in a scenario of convergent/parallel evolution,[55] founder effects,[31] or mutational hotspots.[56] Such information might be useful to drive the design of opportune treatments and vaccines; for instance, by blacklisting positively selected genomic regions.

Step #2: characterization of intra-host genomic diversity

In the second step, VERSO exploits the information on variant frequency (VF) profiles obtained from raw sequencing data (if available), to characterize and visualize the intra-host genomic similarity of hosts with identical (corrected) clonal genotype. In fact, even though the extent and modes of transmission of quasispecies from a host to another during infections are still elusive,[31,57] patterns of co-occurrence of minor variants detected in hosts with identical clonal genotype may provide an indication on the presence of undetected infection paths.[8,58] For this reason, the second step of VERSO is designed to characterize and visualize the genomic similarity of samples by exploiting dimensionality reduction and clustering strategies typically employed in single-cell analyses.[59] Alternative approaches for the analysis of quasispecies, yet with different goals and algorithmic assumptions, have been proposed, for instance in Refs.[60–63] and recently reviewed in Knyazev et al.[64] As specified above, VERSO step #2 is executed on groups of samples with identical clonal genotype: the rationale is that the transmission of minor variants implicates the concurrent transfer of clonal variants, excluding the rare cases in which the VF of a clonal variant significantly decreases in a given host; for instance due to mutation losses (e.g., via recombination-associated deletions or via multiple mutations hitting an already mutated genome location[34]) or to complex horizontal evolution phenomena (e.g., super-infections[65,66]). Conversely, the transmission of clonal variants does not necessarily implicate the transfer of all minor variants, which are affected by complex recombination and transmission effects, such as bottlenecks.[31,57] As a final result, VERSO allows one to visualize the genomic similarity of samples on a low-dimensional space (e.g., UMAP [uniform manifold approximation and projection][67] or tSNE [t-distributed stochastic neighbor embedding][68]) representing the intra-host genomic diversity, and to characterize high-resolution infection chains, thus overcoming the limitations of methods relying on consensus sequences.

Homoplasy detection (minor variants)

Importantly, minor variants observed in hosts with distinct clonal genotypes (identified via VERSO step #1) may indicate



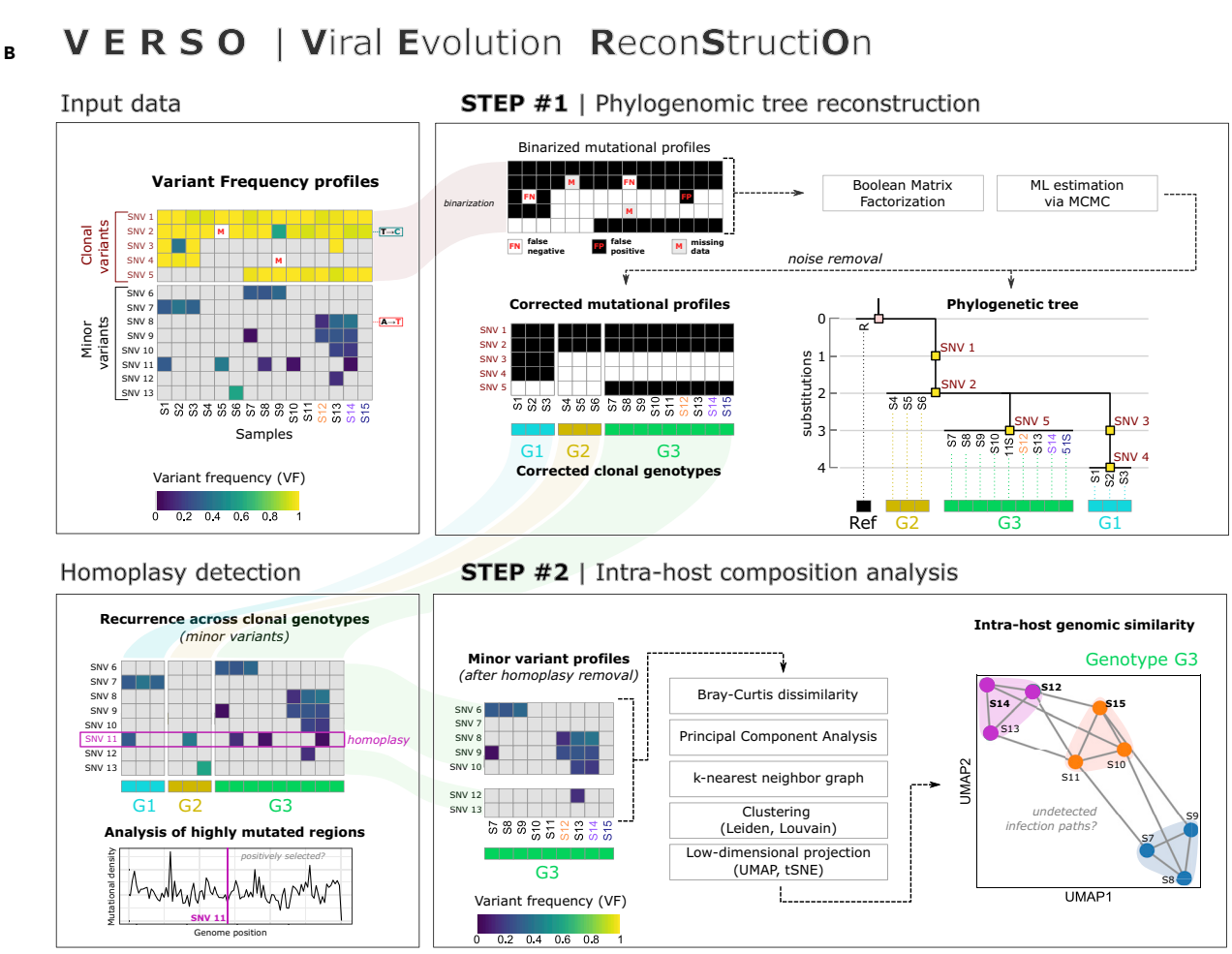


Figure 1. VERSO framework for viral evolution inference and intra-host genomic diversity quantification

(A) In this example, three hosts infected by the same viral lineage are sequenced. All hosts share the same clonal mutation (T>C, green), but two of them (#2 and #3) are characterized by a distinct minor mutation (A>T, red), which randomly emerged in host #2 and was transferred to host #3 during the infection. Standard sequencing experiments return an identical consensus sequence for all samples, by employing a threshold on VF and by selecting mutations characterizing the dominant lineage.

(B) VERSO takes as input the VF profiles of samples, generated from raw sequencing data. In step #1, VERSO processes the binarized profiles of clonal variants and solves a Boolean matrix factorization problem by maximizing a likelihood function via MCMC, in order to correct false-positives/-negatives and missing data. As output, it returns both the corrected mutational profiles of samples and the phylogenetic tree, in which samples with identical corrected clonal genotypes are grouped in polytomies. Corrected clonal genotypes are then employed to identify homoplasies of minor variants, which are further investigated to pinpoint positively selected mutations. The VF profile of minor variants (excluding homoplasies) is processed by step #2 of VERSO, which computes a refined genomic distance among hosts (via Bray-Curtis dissimilarity, after PCA) and performs clustering and dimensionality reduction, in order to project and visualize samples on a 2D space, representing the intra-host genomic diversity and the distance among hosts. This allows one to identify undetected transmission paths among samples with identical clonal genotype.

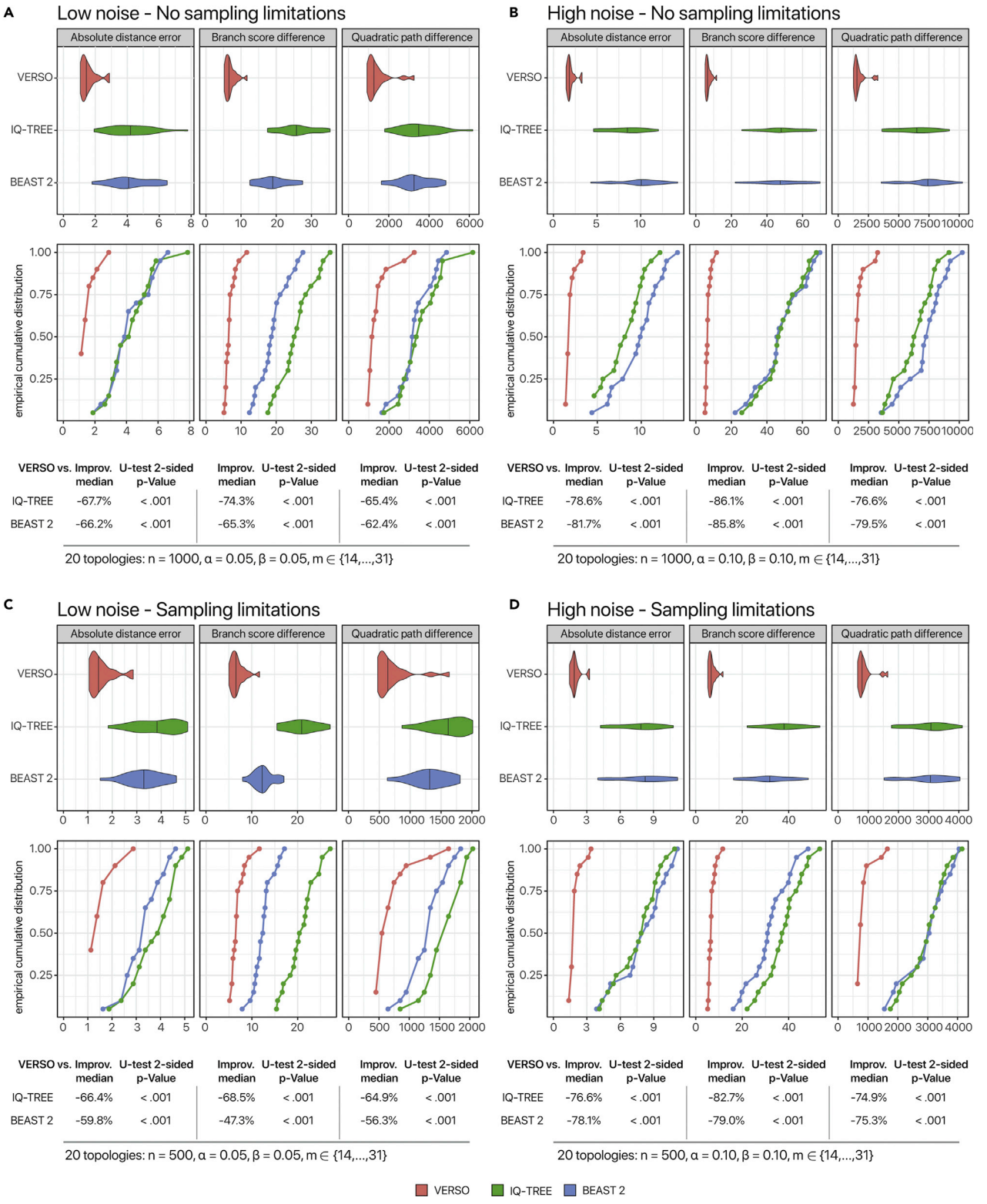


Figure 2. Comparative assessment on simulated data

(A–D) Synthetic datasets were generated via the widely used coalescent model simulator msprime[70] (see the Supplementary Material and Table S1 for the parameter settings). Twenty distinct topologies with 1,000 samples were generated, including a number of distinguishable variants in the range (14, 31). For each topology, four synthetic datasets were generated, with different sample sizes (n = 1000, 500), and different combinations of false-positives and false-negatives

(legend continued on next page)

homoplasies, due to mutational hotspots, phantom mutations, or to positive selection.[56] VERSO pinpoints such variants for further investigations and allows one to exclude them from the computation of the VF-based genomic similarity prior to VERSO step #2, to reduce the possible confounding effects.

To summarize, VERSO (1) returns accurate and robust phylogenies of viral samples, by removing noise from clonal variant profiles; (2) detects reticulation events due to homoplasies of clonal variants; (3) exploits minor variant profiles to characterize and visualize the intra-host genomic similarity of samples with identical (corrected) clonal genotype, thus pinpointing undetected infection paths; (4) allows one to identify and characterize homoplastic minor variants, which might be due to positive selection or mutational hotspots.

To assess the accuracy and robustness of the results produced by VERSO, we performed an extensive array of simulations, and compared with two state-of-the-art methods for phylogenetic reconstruction; i.e., IQ-TREE[10] and BEAST 2.[22] As a major result, VERSO outperforms competing methods in all settings and also in condition of high noise and sampling limitations.

Furthermore, we applied VERSO to two large-scale datasets, generated via amplicon and RNA-seq Illumina sequencing protocols, including 3,960 and 2,766 samples, respectively. The robust phylogenomic models delivered via VERSO step #1 allow us to refine the current estimation on SARS-CoV-2 evolution and spread. Besides, thanks to the in-depth analysis of the mutational landscape of both clonal and minor variants, we could identify a number of variants undergoing transition to clonality, as well as several homoplasies, including variants likely undergoing positive selection processes.

Remarkably, the infection chains identified via VERSO step #2, by assessing the intra-host genomic similarity of samples with the same clonal genotype, were validated by employing contact tracing data from Rockett et al.[69] This important result, which could not be achieved by analyzing consensus sequences, proves the effectiveness of employing raw sequencing data to improve the characterization of the transmission dynamics, in particular during the early phase of the outbreak, in which a relatively low diversity of SARS-CoV-2 has been observed at the consensus level.

VERSO is released as free open source tool at this link: https://github.com/BIMIB-DISCo/VERSO.

RESULTS

Comparative assessment on simulations

In order to assess the performance of VERSO and compare it with competing approaches, we executed extensive tests on simulated datasets, generated with the coalescent model simulator msprime.[70] Simulations allow one to compute a number of metrics with respect to the ground truth, which in this case is the phylogeny of samples resulting from a backwards-in-time coalescent simulation.[71] Accordingly, this allows one to evaluate the accuracy and robustness of the results produced by competing methods in a variety of in-silico scenarios.

In detail, we selected 20 simulation scenarios with n = 1,000 samples in which a number of clonal variants (with distinguishable profiles) between 14 and 31 was observed. We then inflated the datasets with false-positives with rate α and false-negatives with rate β, in order to mimic sequencing and coverage issues. Moreover, additional datasets were generated via random subsampling of the original datasets, to model possible sampling limitations and sampling biases. As a result, we investigated four simulations settings: (A) low noise, no subsampling; (B) high noise, no subsampling; (C) low noise, subsampling; and (D) high noise, subsampling (see Experimental procedures and the Supplemental experimental procedures for further details; the complete parameter settings of the simulations are provided in Table S1).

VERSO step #1 was compared with two state-of-the-art phylogenetic methods from consensus sequences: IQ-TREE,[10] the algorithmic strategy included in the Nextstrain-Augur pipeline,[72] and BEAST 2.[22] Consensus sequences to be provided as input to such methods were generated from simulation data by employing the reference genome SARS-CoV-2-ANC (see below).

The performance of methods was assessed by comparing the reconstructed phylogeny with the simulated ground truth, in terms of (1) absolute error evolutionary distance, (2) branch score difference,[73] and (3) quadratic path difference[74] (please refer to the Supplemental experimental procedures for a detailed description of all metrics).

Figure 2 shows the performance distribution of all methods with respect to all simulation settings. Notably, VERSO step #1 outperforms competing methods in all scenarios (Mann-Whitney U test, $p < 0.001$ in all cases), with noteworthy percentage improvements, also in conditions of high noise and sampling limitations. This important result shows that the probabilistic framework that underlies VERSO step #1 can produce more robust and reliable results when processing noisy data, as typically observed in real-world scenarios.

Reference genome

Different reference genomes have been employed in the analysis of SARS-CoV-2 origin and evolution. Two genome sequences from human samples, in particular, were used in early phylogenomic studies, namely sequence EPI_ISL_405839 (ref. #1 in the following) used, e.g., in Bastola et al.[75] and sequence EPI_ISL_402125 (ref. #2) used, e.g., in Andersen et al.[3] Excluding the polyA tails, the two sequences are identical for 29,865 of 29,870 genome positions (99.98%) and differ for only five SNPs at locations 8,782, 9,561, 15,607, 28,144, and 29,095, for which ref. #1 has haplotype TTCCT and ref. #2 has haplotype CCTTC.

In order to define a likely common ancestor for both sequences, we analyzed the Bat-CoV-RaTG13 genome (sequence

([$\alpha = 0.05$, $\beta = 0.05$], [$\alpha = 0.10$, $\beta = 0.10$]), for a total of four configurations (A, B, C, and D) and 80 independent datasets. VERSO step #1 was compared with IQ-TREE[10] and BEAST 2,[22] on (1) absolute error evolutionary distance, (2) branch score difference[73] and (3) quadratic path difference[74] with respect to the ground-truth sample phylogeny provided by msprime (see the Supplemental experimental procedures for the description of the metrics). In the upper panels, distributions are shown as violin plots, whereas lower panels include the empirical cumulative distribution functions. The percentage improvement of VERSO with respect to competing methods is shown on all metrics (computed on median values), in addition to the p value of the two-sided Mann-Whitney U test on distributions, for all settings.

EPI_ISL_402131)[1] and the Pangolin-CoV genome (sequence EPI_ISL_410721),[3,4] which were identified as closely related genomes to SARS-CoV-2.[76] In particular, it was hypothesized that SARS-CoV-2 might be a recombinant of an ancestor of Pangolin-CoV and Bat-CoV-RaTG13,[4,77] whereas more recent findings would suggest that the SARS-CoV-2 lineage is the consequence of a direct or indirect zoonotic jump from bats.[76] Whatever the case, both Bat-CoV-RaTG13 and Pangolin-CoV display haplotype TCTCT at locations 8,782, 9,561, 15,607, 28,144 and 29,095 and, therefore, one can hypothesize that such a haplotype was present in the unknown common ancestor of ref. #1 and #2.

For this reason, we generated an artificial reference genome, named SARS-CoV-2-ANC, which is identical to both ref. #1 and #2 on 29,865 (over 29,870) genome locations, includes the polyA tail of ref. #2 (33 bases), and has haplotype TCTCT at locations 8,782, 9,561, 15,607, 28,144, and 29,095 (see Figure S2 for a depiction of the artificial genome generation). SARS-CoV-2-ANC is a likely common ancestor of both genomes and was used for variant calling in downstream analyses (SARS-CoV-2-ANC is released in FASTA format as Data S1). Notice that VERSO pipeline is flexible and can employ any reference genome.

Application of VERSO to 3,960 samples from amplicon sequencing data (dataset #1)

We retrieved raw Illumina Amplicon sequencing data of 3,960 SARS-CoV-2 samples of dataset #1 and applied VERSO to the mutational profiles of 2,906 samples selected after quality check (mutational profiles were generated by executing variant calling via standard practices; see Experimental procedures for further details). Notice that the analysis of this dataset was performed independently from that of dataset #2 in order to exclude possible sequencing-related artifacts or idiosyncrasies.

VERSO step #1: robust phylogenomic inference from clonal variant profiles

We first applied VERSO step #1 to the mutational profile of the 29 variants detected as clonal (VF > 90%) in at least 3% of the samples, in order to reconstruct a robust phylogenomic tree. The VERSO phylogenetic model is displayed in Figure 3A and highlights the presence of 25 clonal genotypes, obtained by removing noise from data, and that define polytomies including different numbers of samples (see Experimental procedures for further details). The mapping between clonal genotype labels and the lineage dynamic nomenclature proposed by Rambaut et al.[78] was obtained via pangolin 2.0[79] and is provided in Data S3.

More in detail, variant g.29095T>C (*N*, synonymous) is the earliest evolutionary event from reference genome SARS-CoV-2-ANC and is detected in 2,454 samples of the dataset. The related clonal genotype G1, which is characterized by no further mutations, identifies a polytomy including 57 Australian, 15 Chinese, 12 American, and one South-African samples.

Three clades originate from G1: a first clade includes clonal genotypes G2 (six samples) and G3 (103), while a second clade includes clonal genotype G4 (86). Clonal genotypes G1ffiG4 are characterized by the absence of single nucleotide variants (SNVs) g.8782T>C (*ORF1ab*, synonymous) and g.28144C>T (*ORF8*, p.84S>L) and correspond to previously identified type A[24] (also type S[82]), which was hypothesized to be an early SARS-CoV-2 type.

The third clade originating from clonal genotype G1 includes all remaining clonal genotypes (G5-G25) and is characterized by the presence of both SNVs g.8782T>C and g.28144C>T. This specific haplotype corresponds to type B[24] (also type L[82]) and an increase of its prevalence has progressively recorded in the population, as one can see in Figure 3, as opposed to type A (S), which was rarely observed in late samples. In this regard, we note that there are currently insufficient elements to support any epidemiological claim on virulence and pathogenicity of such SARS-CoV-2 types, even if recent evidences would suggest the existence of a low correlation.[83]

Variant g.23403A>G (*S*, p.614D>G) is of particular interest, as proven by the increasing number of related studies.[84–87] Such a variant identifies a large clade including 11 clonal genotypes: G15 (493 samples), G16 (1), G17 (512), G18 (25), G19 (118), G20 (94), G21 (648), G22 (90),G23 (127), G24 (4), and G25 (86), for a total of 2,198 total samples, distributed especially in Australia (971), the United States (841), South Africa (257), and Israel (125). Importantly, a constant increase of the prevalence of the haplotype corresponding to such variant is observed in time (see Figure 3), which might hint at ongoing positive selection processes; e.g., due to increased viral transmission. However, this hypothesis is highly debated[88] and, in order to investigate the possible functional effect of such variant and the related clinical implications, *in vivo* and *in vitro* studies are needed.[87]

By looking more in detail at the geo-temporal localization of samples depicted via Microreact[81] (Figure 3B), one can see that the different clonal genotypes are distributed across the world in distinct complex patterns, suggesting that most countries might have suffered from multiple introductions, especially in the early phases of the epidemics. In particular, samples are distributed in 11 countries, with Australia (1,523 samples), United States (910), South Africa (260), Israel (133), and China (45) representing around 99% of the dataset.

The country displaying the largest number of samples is Australia, with 1,523 samples, distributed in 22 different clonal genotypes. The presence of a number of early clonal genotypes (i.e., G1, G2, G3, G4, and G6) supports the hypothesis of multiple introductions of SARS-CoV-2 in Australia. Interestingly, we note that, from the 16th week on, the composition of the Australian sample group tends to be polarized toward clonal genotypes G17 (108/311 ≈ 35%) and G25 (82/311 ≈ 26%).

910 samples from the United States are included in the dataset, distributed in 17 different clonal genotypes, with G21 being the most abundant in the population (376/910 ≈ 41%). Also in this case, samples collected in the initial weeks belong to the ancestral clades, supporting the hypothesis of multiple introductions. Notably, after the 17th week, all American samples display the haplotype g.23403A>G (S,p.614D>G) and we notice an overall decrease in genomic diversity, since the observed clonal genotypes pass from 16 (week interval 9–16, 2020) to 8 (week interval 17–29, 2020). Notice that only 49.1% of the American samples have a collection date.

Two-hundred and sixty samples from South Africa are included in the dataset, which are partitioned in six different clonal genotypes, four of which (G1, G8, G14, and G16) include

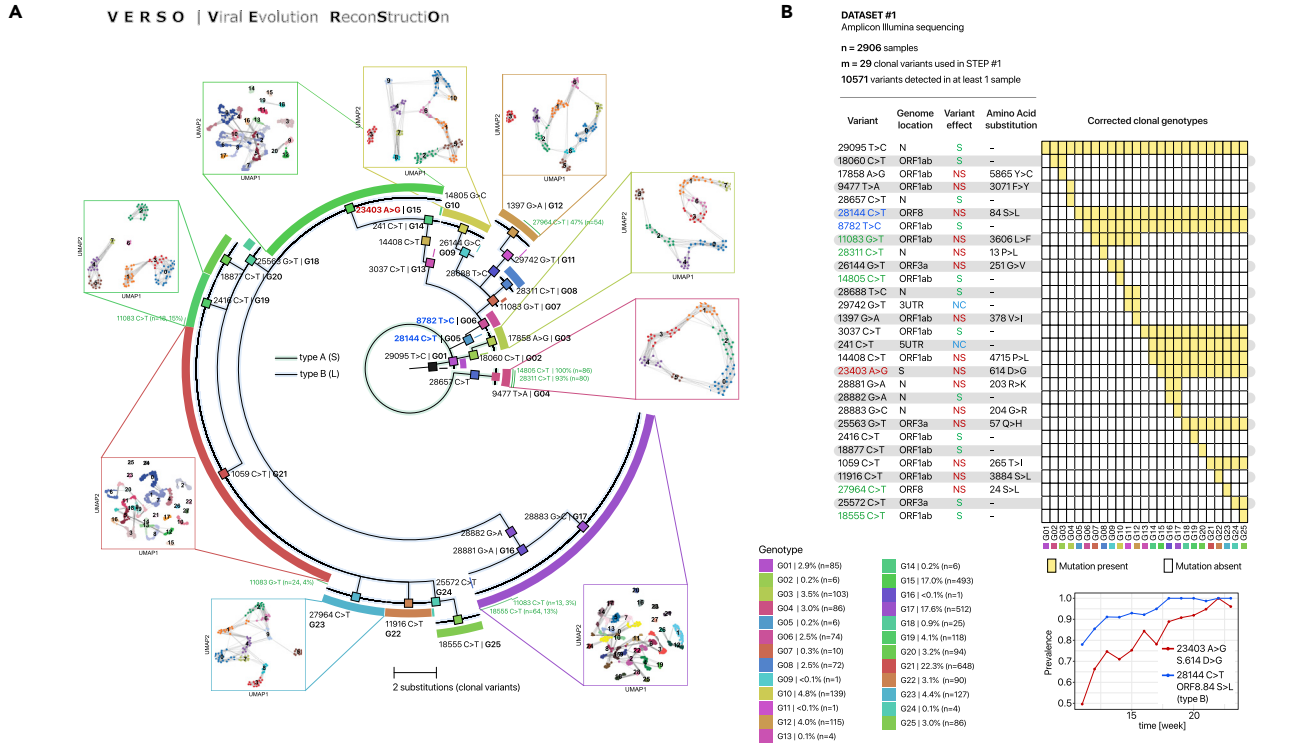


Figure 3. Viral evolution and intra-host genomic characterization of 2906 SARS-CoV-2 samples of via VERSO (dataset #1)

(A) The phylogenetic model returned by VERSO step #1 from the mutational profile of 2,906 samples selected after the quality check, on 29 clonal variants (VF > 90%) detected in at least 3% of the samples of dataset #1 (reference genome: SARS-CoV-2-ANC). Colors mark the 25 distinct clonal genotypes identified by VERSO (the mapping with the lineage nomenclature proposed in Rambaut et al.[78] and generated via pangolin 2.0[79] is provided in File S3). Samples with identical corrected clonal genotypes are grouped in polytomies and the black sample represents the SARS-CoV-2-ANC genome (visualization via FigTree[80]). The green curves juxtaposed to certain polytomies report the number and fraction of samples in which the five homoplastic mutations are observed (only if the mutation is detected in at least 10 samples with the same corrected clonal genotype; see Data S2 for a summary on the samples exhibiting homoplastic clonal variants). The projection of the intra-host genomic diversity computed by VERSO step #2 from VF profiles is shown on the UMAP low-dimensional space for the clonal genotypes including $\geq$100 samples. Samples are clustered via Leiden algorithm on the kNN graph (k = 10), computed on the Bray-Curtis dissimilarity on VF profiles, after PCA. Solid lines represent the edges of the k-NNG.

(B) The composition of the corrected clonal genotypes returned by VERSO step #1 is shown. Clonal SNVs are annotated with mapping on ORFs, synonymous (S), nonsynonymous (NS), and non-coding (NC) states, and related amino acid substitutions. Variants g.8782T>C (*ORF1ab*, synonymous) and g.28144C>T (*ORF8*, p.84S>L) are colored in blue, whereas variant g.23403 A¿G (*S*, p.614 D¿G) is colored in red. The prevalence variation in time of the relative haplotypes (i.e., the fraction of samples displaying such mutations) is also shown. The five homoplastic variants are colored in green.

(C and D) (C) The geo-temporal localization of the clonal genotypes via Microreact[81] and (D) the prevalence variation in time are displayed.

a single sample, whereas 98.46% of the samples exhibit the haplotype g.23403A>G (S,p.614D>G) and, specifically, are included in clonal genotypes G15 and G17. Finally, all Chinese samples were collected in the early phase (January–February, 2020) and are characterized by six different clonal genotypes (i.e., G1, G6, G7, G9, G11, and G12).

Homoplasy detection (clonal variants)

Five clonal variants included in our model show apparent violations of the accumulation hypothesis, namely g.11083G>T (*ORF1ab*, p.3606 L>F), g.14805C>T (*ORF1ab*, synonymous), g.18555C>T (*ORF1ab*, synonymous), g.27964C>T (*ORF8*, p.24S>L), and g.28311C>T (*N*,p.13P>L), suggesting that they

might be involved in homoplasies. In Figure 3 the samples in which the five homoplastic variants are detected are highlighted (if the mutation is detected in $\geq$10 samples with the same corrected clonal genotype), whereas in Figure S3 one can find the expanded clonal variant tree, in which the reticulation related to such variants is explicitly depicted.

Some of such variants have been exhaustively studied (e.g., g.11083G>T in van Dorp et al.[89]), specifically to verify possible scenarios of convergent evolution, which may unveil the fingerprint of adaptation of SARS-CoV-2 to human hosts. To this end, particular attention should be devoted to the three non-synonymous substitutions; i.e., g.11083G>T (present in 460 samples, $\approx$16% of the dataset), g.27964C>T (182 samples, $\approx$6%) and g.28311C>T (153 samples, $\approx$5%). As a first result, we note the prevalence dynamics of the haplotypes defined by such variants does not show any apparent growth trend in the population (see Figure S5).

To further investigate if such variants fall in a region prone to mutations of the SARS-CoV-2 genome, we evaluated the mutational density employing a sliding window approach similarly to Soares et al.[90] (see Supplemental experimental procedures for additional details). As shown in Figure S4, the mutational density, computed by considering synonymous minor variants, exhibits a median value of = 0.083 [syn.mutations][nucleotides]$^{-1}$. Interestingly, the three nonsynonymous SNVs (g.11083G>T, g.27964C>T and g.28311C>T) are located within windows with a higher mutational density than the median value: 0.085, 0.124, and 0.1 $\frac{\text{syn.mutations}}{\text{nucleotides}}$, respectively (see Table S5), and this would suggest that they might have originally emerged due to the presence of natural mutational hotspots or phantom mutations.

However, this analysis is not conclusive and further investigations are needed to characterize the functional effect of such mutations and the possible impact in the evolutionary and diffusion process of SARS-CoV-2.

Stability analysis

The choice of an appropriate VF threshold to identify clonal variants and, accordingly, to generate consensus sequences from raw sequencing data might affect the stability of the results of any downstream phylogenomic analysis. On the one hand, loose thresholds might increase the risk of including non-clonal variants in consensus sequences. On the other hand, too strict thresholds might increase the rate of false-negatives, especially with noisy sequencing data.

For this reason, we assessed the robustness of the results produced by VERSO step #1 on dataset #1 when different thresholds in the set $\delta \in \{0.5,0.6,0.7,0.8\}$ are employed to identify clonal variants, with those obtained with default threshold (δ = 0.9), in terms of tree accuracy (see the Supplemental experimental procedures for further details). As one can see in Figure S7, the tree accuracy varies between 0.97 and 0.98 in all settings, proved the results produced by VERSO step #1 are robust with regard to the choice of the VF threshold for clonal variant identification.

VERSO step #2: Characterization of intra-host genomic diversity

We then applied VERSO step #2 to the complete VF profiles of the samples with the same clonal genotype and projected their intra-host genomic diversity on the UMAP low-dimensional space. This was done excluding (1) the clonal variants employed in the phylogenetic inference via VERSO step #1, (2) all minor variants (VF $\leq$ 90%) observed in more than one clonal genotype (i.e., homoplasies) and that are likely emerged independently within the hosts, due to mutational hotspots, phantom mutations, or positive selection (see Experimental procedures and the next subsections). Even though, as expected, the VF profiles of minor variants are noisy, a complex intra-host genomic architecture is observed in several individuals. Moreover, patterns of co-occurrence of minor variants across samples support the hypothesis of transmission from one host to another.

In Figure 3 we display the UMAP plots for the clonal genotypes including more than 100 samples, plus clonal genotype G4 (n = 86 samples), which was used for contact tracing analyses. Such maps describe likely transmission paths among hosts characterized by the same (corrected) clonal genotype and, in most cases, suggests the existence of several distinct infection clusters with different size and density. This result was achieved by exploiting the different properties of clonal and minor variants via the two-step procedure of VERSO.

Contact tracing

To corroborate our findings, we employed the contact tracing data from Rockett et al.,[69] in which 65 samples from dataset #1 are characterized with respect to household, work location, or other direct contacts. Four distinct contact groups, including 36, 15, 12, and 2 samples, respectively, are associated directly or indirectly to three different New South Wales institutions (i.e., institutions #1, #2, and #3) and to the same household environment (household #1).

As a first result, all samples belonging to a specific contact group are characterized by the same clonal genotype, determined via VERSO step #1, a result that confirms recent findings.[42,69] More importantly, the analysis of the intra-host genomic diversity via VERSO step #2 allows one to highly refine this analysis.

In Figure 4 one can find the UMAP plot of clonal genotypes G4, G12, and G21, which include 36 (over 86), 14 (over 115), and 14 (over 648) samples with contact information. Strikingly, the distribution of the pairwise intra-host genomic distance among samples from the same institution/household (computed on the K-nearest neighbor graph [k-NNG] via Bray-Curtis dissimilarity, after principal component analysis [PCA]; see Experimental procedures) is significantly lower with respect to the distance of all samples with the same clonal genotype (p value of the Mann-Whitney U test <0.001 in all cases). Furthermore, all samples belonging to the same contact group are connected in the k-NNG, while a noteworthy proportion of samples without contact information in genotypes G12 and G21 are placed in disconnected graphs (24.9% and 76.4%, respectively).

This major result suggests that patterns of co-occurrence of minor variants can indeed provide useful indication on contact tracing dynamics, which would be masked when employing consensus sequencing data. Accordingly, the algorithmic strategy employed by VERSO step #2 and, especially, the identification of the k-NNG on intra-host genomic similarity provides an effective tool to dissect the complexity of viral evolution and transmission, which might in turn improve the reliability of currently available contact tracing tools.

Homoplasy detection (minor variants)

Several minor variants are found in samples with distinct clonal genotypes and might indicate the presence of homoplasies. In

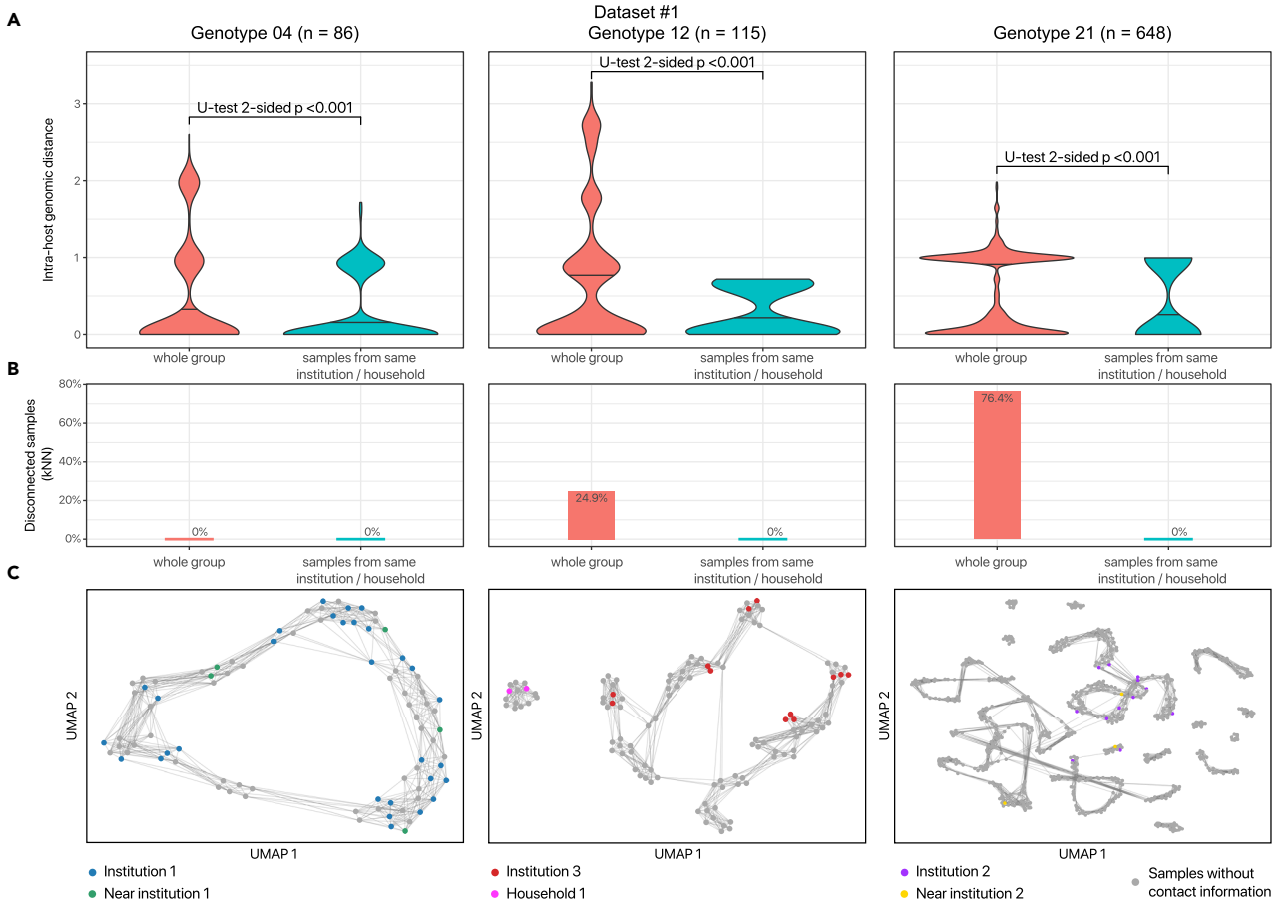


Figure 4. Infection dynamics revealed via characterization of intra-host genomic similarity (dataset #1)

(A) The distribution of the pairwise intra-host genomic distance (computed via Bray-Curtis dissimilarity on the kNN graph, with $k = 10$, after PCA; see Experimental procedures) for the samples belonging to the same household or institution (including samples marked as near), versus the pairwise distance of all samples belonging to clonal genotypes G4, G12, and G21. The p values of the Mann-Whitney U test two-sided are also shown.

(B) The proportion of samples that are disconnected in the kNN graph, with respect to the samples belonging to the same household or institution (including samples marked as near) and with respect to all samples.

(C) The UMAP projection of the intra-host genomic diversity of the samples belonging to clonal genotypes G4, G12, and G21, returned by VERSO step #2.

this respect, the heatmap in Figure 5F returns the distribution of minor SNVs with respect to (1) the number of distinct clonal genotypes in which they are detected, and (2) the mutational density of the region in which they are located (see the Supplemental experimental procedures for details on the mutational density analysis).

The intuition is that the variants detected in single clonal genotypes (left region of the heatmap) are likely spontaneously emerged private mutations, or the result of infection events between hosts with same clonal genotype (see above). Conversely, SNVs found in multiple clonal genotypes (right region of the heatmap) may have emerged due to positive selection in a parallel/convergent evolution scenario, or to mutational hotspots or phantom mutations. To this end, the mutational density analysis provides useful information to pinpoint mutation-prone regions of the genome.

Interestingly, a significant number of minor variants are observed in multiple clonal genotypes and fall in scarcely mutated regions of the genome (see Figure S4). This would suggest that some of these variants might have been positively selected, due to some possible functional advantage or to transmission-related founder effects. In this respect, we further focused our investigation on a list of 80 candidate minor variants that (1) are detected in more than one clonal genotype, (2) are present in at least 10 samples, (3) are nonsynonymous, and (4) fall in a region of the genome with mutational density lower than the median value (see Table S6 for details on such variants). In the following, we focus on a subset of such variants falling on the spike gene of the SARS-CoV-2 genome.

Considerations on homoplasies falling on the spike gene

The spike protein of SARS-CoV-2 plays a critical role in the recognition of the ACE2 receptor and in the ensuing cell membrane fusion process.[91] We prioritized three candidate homoplastic minor variants occurring on the SARS-CoV-2 spike gene (S) (see Table S6). Interestingly, two out of three, namely g.24552T>C (p.997I>T) and g.24557G>T (p.999G>C), detected in 57 samples in total (10 and 47 samples, respectively), clustered in the so-called connector region (CR), bridging between the two heptad repeat regions (HR1 and HR2) of the S2 subunit of the spike protein.

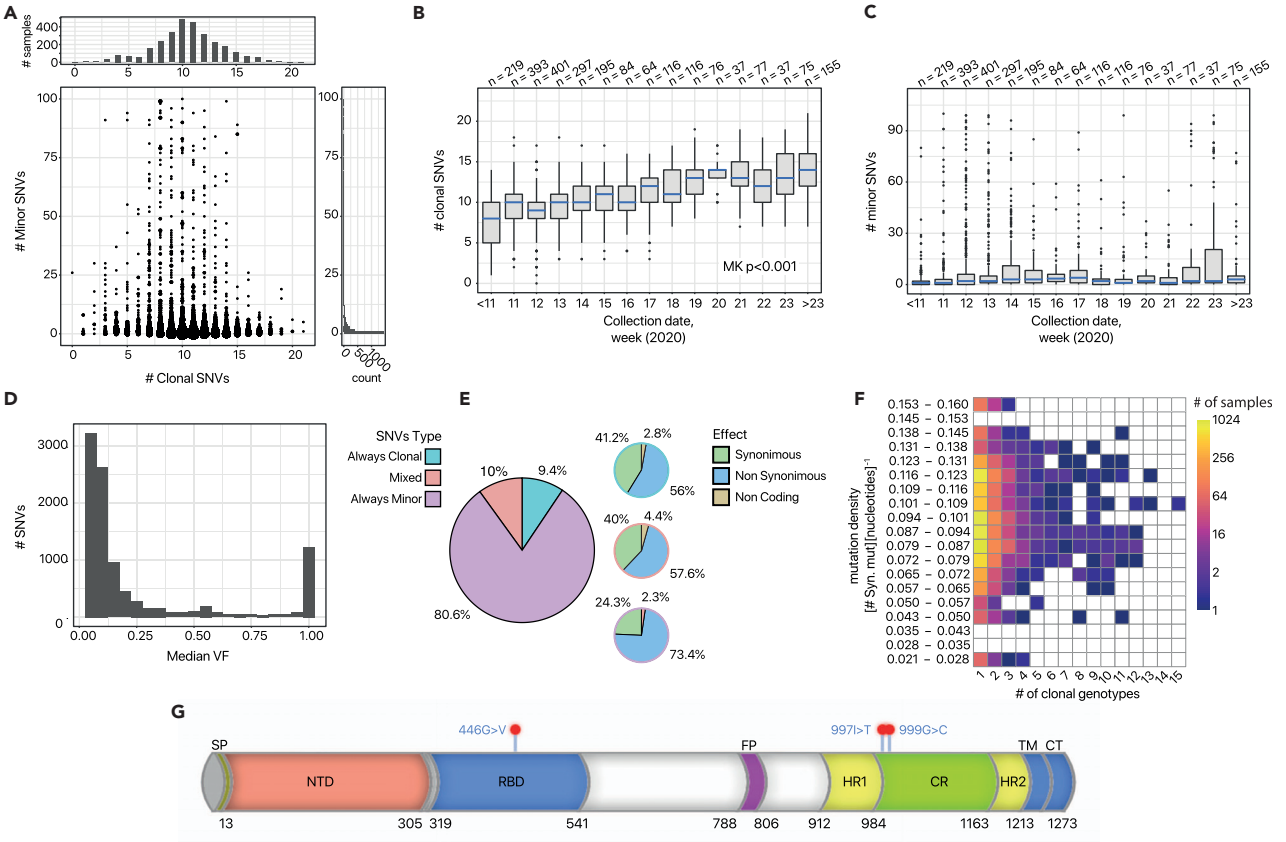


Figure 5. Mutational landscape of 2906SARS-CoV-2 samples (dataset #1)

(A) Scatterplot displaying, for each sample, the number of clonal (VF > 90%) and minor variants (VF ≤ 90%, node size proportional to the number of samples).

(B and C) Boxplots returning the distribution of the number of clonal (B) and minor variants (C), obtained by grouping samples according to collection date (weeks, 2020). The p value of the Mann-Kendall (MK) trend test on clonal variants is highly significant.

(D) Distribution of the median VF for all SNVs detected in the viral populations.

(E) Pie charts returning (left) the proportion of SNVs detected as always clonal, always minor, or mixed; (right) for each category, the proportion of synonymous, nonsynonymous, and non-coding variants (check the pie-chart border color for a visual clue).

(F) Heatmap returning the distribution of always minor SNVs with respect to (x axis) the number of clonal genotype of the phylogenomic model in Figure 3 in which each variant is observed, (y axis) the mutational density of the genome region in which it is located (see the Supplemental experimental procedures).

(G) Mapping of the candidate homoplastic minor variants located on the spike gene of the SARS-CoV-2 virus.

When the receptor binding domain (RBD) binds to ACE2 receptor on the target cell, it causes a conformational change responsible for the insertion of the fusion peptide (FP) into the target cell membrane. This, in turn, triggers further conformational changes, eventually promoting a direct interaction between HR1 trimer and HR2, which occurs upon bending of the flexible CR, in order to form a six-helical HR1-HR2 complex known as the fusion core region (FCR) in close proximity to the target cell plasma membrane, ultimately leading to viral fusion and cell entry.[92]

Peptides derived from the HR2 heptad region of enveloped viruses and able to efficiently bind to the viral HR1 region inhibit the formation of the FCR and completely suppress viral infection.[93] Therefore, the formation of the FCR is considered to be vital to mediate virus entry in the target cells, promoting viral infectivity. Of note, the CR is highly conserved across the Gammacoronavirus genus, supporting the notion that this region may play a very important but still unclear functional role (Figure 5G). Although structural and *in vitro* models will be required in order to extensively characterize the functional effect of these variants, the evidence that two of our three minor variants detected in the spike protein falls in a small domain comprising less than 14% of the entire spike protein length is intriguing, as it suggests a potential functional role for these mutations. It will be important to track the prevalence of these mutations, as well as of all other candidate convergent variants falling on different region of the SARS-CoV-2, to highlight possible transitions to clonality (see below). We also remark that, being a data-science computational approach, VERSO can struggle in dissecting complex mutational cases, since all the experimental hypotheses that can be generated are clearly data dependent. For this reason, and given the heterogeneity and limitations of currently available SARS-CoV-2 datasets, any hypothesis delivered by VERSO requires additional independent investigations and *ad hoc* experimental validations.

Mutational landscape

We analyzed in depth the mutational landscape of the samples of dataset #1. First, the comparison of the number of clonal

(VF > 90%) and minor variants detected in each host (Figure 5A) reveals a bimodal distribution of clonal variants (with first mode at 4 and s mode at 10), whereas minor variants display a more dispersed long-tailed distribution with median equal to 2 and average ≈23. From the plot, it is also clear that individuals characterized by the same clonal genotype may display a significantly different number of minor variants, with distinct distributions observed across clonal genotypes.

The comparison of the distribution of the number of variants obtained by grouping the samples with respect to collection week (Figures 5B and 5C) allows us to highlight a highly statistically significant increasing trend for clonal variants (Mann-Kendall trend test on median number of clonal variants, $p < 0.001$). This result would strongly support both the hypothesis of accumulation of clonal variants in the population and that of a concurrent increase of overall genomic diversity of SARS-CoV-2,[36,94] whereas the relevance of this phenomenon on minor variants is unclear.

We then focused on the properties of the SNVs detected in the population. Surprisingly, the distribution of the median VF for each detected variant (Figure 5D) reveals a bimodal distribution, with the large majority of variants showing either a very low or a very high VF, with only a small proportion of variants showing a median VF within the range 10%–90%. This behavior is typical of systems where the prevalence of some subpopulations is driven by positive Darwinian selection while others are purified.[95]

In order to analyze the two components of this distribution, we further categorized the variants as always clonal (i.e., SNVs detected with VF >90% in all samples), always minor (i.e., SNVs detected with VF 5% and ≤90% in all samples), and mixed (i.e., SNVs detected as clonal in at least one sample and as minor in at least another sample). As one can see in Figure 5E, 9.4%, 80.6%, and 10% all SNVs are respectively detected as always clonal, always minor, and mixed in our dataset. Moreover, 56%, 73.4%, and 57.6% of always clonal, always minor, and mixed variants, respectively, are nonsynonymous, whereas the large majority of remaining variants are synonymous.

These results would suggest that, in most cases, randomly emerging SARS-CoV-2 minor variants tend to remain at a low frequency in the population, whereas, in some circumstances, certain variants can undergo frequency increases and even become clonal, due to undetected mixed transmission events or to selection shifts, as it was observed by Poon et al.[8] for the cases of H3N2 and H1N1/2009 influenza. Interestingly, 15 variants identified as possibly convergent (see above) fall into this category and deserve further investigations (see Table S6 for additional details).

Transmission bottleneck analysis

The estimation of transmission bottlenecks might be of specific interest during the current pandemics. Despite most available methods requiring data collected on donor-host couples (see, e.g., Sobel Leonard et al.[96] and Ghafari et al.[97]), here we employed a strategy akin to Monsion et al.[98] and Lequime et al.[99] that is roughly based on the analysis of the variation of the VF variance of a number of candidate neutral mutations. The intuition is that variance shrinking indicates significant transmission bottlenecks, which, accordingly, would result in lower viral diversity transferred from a host to another and, possibly, in purification of certain variants in the population. As the analysis ideally requires the comparison of groups in which infection events have occurred, here we considered groups of samples with distinct clonal genotypes, separately. We then selected a number of variants as neutral markers. The rationale is that transmission phenomena such as bottlenecks are expected to significantly affect the VF variance of neutral markers (please see Supplemental experimental procedures for further details).

More in detail, we first split the samples of each clonal genotype for which a collection date is available into non-overlapping groups corresponding to two consecutive time windows; i.e., before and after the 14th week, 2020. Accordingly, three SNVs were selected as candidate-neutral or quasineutral markers, namely variants g.634T>C, g.14523A>G, and g.15168G>A. In Figure S6, one can find the distribution of the VF of the selected markers with respect to the time windows, which highlights moderate variations of the variance for all markers (see also Table S7. All in all, this result would suggest the presence of mild bottleneck effects, consistent with recent studies involving donor-host data.[43]

Application of VERSO to 2,766 samples from RNA-sequencing data (dataset #2)

We retrieved the raw Illumina RNA-sequencing data of 2,766 samples included in dataset #2 and applied VERSO to the mutational profiles of 1,438 samples selected after quality check. Twenty-three clonal variants were employed in the analysis, according to the filters described later.

The resulting phylogenetic model is consistent with the one obtained for dataset #1, despite minor differences (Figure S8A). Specifically, 18 distinct clonal genotypes are identified by VERSO step #1, 11 of which are identical to those found in the analysis of dataset #1 (in such cases the same genotype label was maintained; see Data S3 for the mapping with the lineage nomenclature proposed by Rambaut et al.[78]). Five further clonal genotypes are evolutionarily consistent and represent independent branches detected due to the non-overlapping composition of the dataset, and are labeled with progressive letters from the closest genotype (i.e., G13a, G21a, G22a, G22b, G23a), while the two samples of genotype G13a* might be safely assigned to genotype G13a, since the absence of mutation g.3037C>T is likely due to low coverage.

By excluding the remaining clonal genotype GH, which presents inconsistencies due to the presence of the candidate homoplastic variant g.11083G>T (*ORF1ab*, p.3606L>F, see above), all clonal genotypes display the same ordering in both datasets (also see the expanded clonal variant tree in Figure S9). This proves the robustness of the results delivered by VERSO step #1 even when dealing with data generated from distinct sequencing platforms.

By looking at the geo-temporal localization of samples obtained via Microreact[81] (Figure S8B), one can see that that dataset #2 includes samples with a significantly different geographical distribution with respect to dataset #1. This dataset contains sample from 10 countries, with the large majority collected in the United States (96.8%). More in detail, the samples of such countries are mostly characterized by clonal genotype G21. We further notice that, also for dataset #2, mutation g.23403A>G (*S*,p.614D>G) becomes prevalent in the population at late collection dates. Moreover, only samples belonging to previously defined type B are detected in this dataset.

The analysis of the intra-host genomic diversity was also performed for dataset #2 via VERSO step #2, which would suggest the existence of undetected infection events and of several infection clusters with distinct properties, even though no contact tracing is available in this case. Overall, this proves the general applicability of the VERSO framework, which can produce meaningful results when applied to data produced with any sequencing platforms. However, in order to minimize the possible impact of data- and platform-specific biases, our suggestion is to perform the VERSO analysis on datasets generated from different protocols separately.

Scalability

We finally assessed the computational time required by VERSO in a variety of simulated scenarios. The results are shown in the Supplemental experimental procedures (Figure S10) and demonstrate the scalability of VERSO also when processing large-scale datasets.

DISCUSSION

We introduced VERSO, a comprehensive framework for the high-resolution characterization of viral evolution from sequencing data, which is an improvement on currently available methods for the analysis of consensus sequences. VERSO exploits the distinct properties of clonal and minor variants to dissect the complex interplay of genomic evolution within hosts and transmission among hosts.

On the one hand, the probabilistic framework underlying VERSO step #1 delivers highly accurate and robust phylogenetic models from clonal variants, also in conditions of noisy observations and sampling limitations, as proved by extensive simulations and by the application to two large-scale SARS-CoV-2 datasets generated from distinct sequencing platforms. On the other hand, the characterization of intra-host genomic diversity provided by VERSO step #2 allows one to identify undetected infection paths, which were in our case validated with contact tracing data, as well as to intercept variants involved in homoplasies.

This may represent a major advancement in the analysis of viral evolution and spread and should be quickly implemented in combination with data-driven epidemiological models to deliver a high-precision platform for pathogen detection and surveillance.[12,100] This might be particularly relevant for countries that suffered outbreaks of exceptional proportions and for which the limitations and inhomogeneity of diagnostic tests have proved insufficient to define reliable descriptive/predictive models of disease diffusion. For instance, it was hypothesized that the rapid diffusion of COVID-19 might be likely due to the extremely high number of untested asymptomatic hosts.[101]

More accurate and robust phylogenetic models may allow one to improve the assessment of molecular clocks and, accordingly, the estimation of the parameters of epidemiological models such as susceptible-infected-recovered (SIR) and susceptible-infected-susceptible (SIS),[11,102] as well as to unravel the cryptic transmission paths.[8,12,13,103] Furthermore, the finer grain of the analysis on intra-host genomic similarity from sequencing data might be employed to enhance the active surveillance; for instance, by facilitating the identification of infection clusters and super-spreaders.[104] Finally, the characterization of variants possibly involved in positive selection processes might be used to drive the experimental research on treatments and vaccines.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Alex Graudenzi, Institute of Molecular Bioimaging and Physiology, Consiglio Nazionale delle Ricerche (IBFM-CNR), via F.lli Cervi, 93, 20,090 Segrate, Milan, Italy. alex.graudenzi@ibfm.cnr.it.

Materials availability

This study did not generate new unique reagents.

Data and code availability

VERSO is freely available at this link: https://github.com/BIMIB-DISCo/VERSO. VERSO step #1 is provided as an open source standalone R tool, whereas step #2 is provided as a Python script. The source code to replicate all the analyses presented in the manuscript, both on simulated and real-world datasets, is available at this link: https://github.com/BIMIB-DISCo/VERSO-UTILITIES.

SCANPY[59] is available at this link: https://scanpy.readthedocs.io/en/stable/. The Web-based tool for the geo-temporal visualization of samples, Microreact,[81] is available at this link: https://microreact.org/showcase. The tool employed to plot the phylogenomic model returned by VERSO step #1 (in Newick file format) is FigTree[80] and is available at this link: http://tree.bio.ed.ac.uk/software/figtree/. The tool used for the mapping between clonal genotype labels and the dynamic nomenclature proposed by Rambaut et al.[78] is pangolin 2.0[79] and is available at this link: https://github.com/cov-lineages/pangolin.

VERSO step #1: robust phylogenomic inference from clonal variant profiles

VERSO is a novel framework for the reconstruction of viral evolution models from raw sequencing data of viral genomes. It includes a two-step procedure, which we describe in the following.

The first step of VERSO employs a probabilistic maximum-likelihood framework for the reconstruction of robust phylogenetic trees from binarized mutational profiles of clonal variants (or, alternatively, from consensus sequences). This step relies on an evolved version of the algorithmic framework introduced by Ramazzotti et al.[105] for the inference of cancer evolution models from single-cell sequencing data, and can be executed independently from step #2, in case raw sequencing data are not available.

Inputs

The method takes as input a n (samples) $\times m$ (variants) binary mutational profile matrix, as defined on the basis of clonal SNVs only. In this case, an entry in a given sample is equal to 1 (present) if the VF is larger than a certain threshold (in our analyses, equal to 90%), it is equal to 0 if lower than a distinct threshold (in our analyses, equal to 5%), and is considered as missing (NA) in the other cases, thus modeling possible uncertainty in sequencing data or low coverage.

Notice that consensus sequences can be processed by VERSO step #1 by generating a consistent binarized mutational profile matrix. We also note that, given the intrinsic challenges associated with a reliable identification of low VF indels, the analysis focuses only on SNVs. Further details on the variant calling pipeline employed in this study are provided next.

The algorithmic framework

VERSO step #1 is a probabilistic framework that solves a Boolean matrix factorization problem with perfect phylogeny constraints and relying on the infinite sites assumption (ISA).[106,107] The ISA subsumes a consistent process of accumulation of clonal variants characterizing the evolutionary history of the virus and does not allow for losses of mutations or convergent variants (i.e., mutations observed in distinct clades).

In this regard, we recall that that the variant accumulation hypothesis holds only when considering clonal mutations. In fact, clonal mutations (e.g., A, B, C, D, and E) are present, by definition, in the large majority of the quasispecies of a given sample, depending on the chosen VF threshold (in our case, equal to 90%; see above). Since such variants are rarely lost, they are most likely transmitted from one host to another during infections. In addition, the origination of

new clonal mutations in single samples leads to the definition of new clonal genotypes, following a standard branching process (e.g., A, AB, ABC, ABD, ABDE). As a result, clonal mutations typically accumulate during the evolutionary history of a virus, excluding complex scenarios involving reticulation events,[53] whereas clonal genotypes can clearly become extinct. Conversely, variants with lower frequency do not necessarily accumulate, due to the high recombination rates, as well as to bottlenecks, founder effects, and stochasticity,[31] and this is the reason why they were considered separately in the analysis, via VERSO step #2 (see below).

More in detail, VERSO step #1 accounts for uncertainty in the data, by employing a maximum-likelihood approach (via Markov chain Monte Carlo [MCMC] search) that allows for the presence of false-positives, false-negatives, and missing data points in clonal variant profiles. As shown by Ramazzotti et al.[105] in a different experimental context, our algorithmic framework ensures robustness and scalability also in case of high rates of errors and missing data, due, for instance, to sampling limitations. Furthermore, it is robust to mild violations of the ISA (e.g., due to reticulation events, such as convergent variants) or mutation losses, which can be characterized after the inference, if present (see the specific features on homoplasy detection discussed next). Please refer to the Supplemental experimental procedures for further details on the algorithmic framework and its assumptions, including the probabilistic graphical model depicted in Figure S1 and the summary of notation in Table S4.

Outputs

The inference returns a set of maximum-likelihood variants trees (minimum 1) as sampled during the MCMC search, representing the ordering of accumulation of clonal variants, and a set of maximum-likelihood attachments of samples to variants. Given the variants tree and the maximum-likelihood attachments of samples to variants, VERSO outputs (1) a phylogenetic model where each leaf correspond to a sample, whereas internal nodes correspond to accumulating clonal variants; (2) the corrected clonal genotype of each sample (i.e., the binary mutational profile on clonal variants obtained after removing false-positives, false-negatives, and missing data).

The model naturally includes polytomies, which group samples with the same corrected clonal genotype. The length of the branches in the model represents the number of clonal substitutions (which can be normalized with respect to genome length), as in standard phylogenomic models, and the clades of the model correspond to viral lineages. The VERSO phylogenetic model is provided as output in Newick file format and can be processed and visualized in standard tools for phylogenetic analysis, such as FigTree[80] or Dendroscope.[108] Furthermore, VERSO allows one to visualize the geo-temporal localization of clonal genotypes via Microreact.[81]

Additional feature: homoplasy detection on clonal variants

Violations of the ISA are possible and can be due to reticulation events[53] such as homoplasies (i.e., identical variants detected in samples belonging to different clades) or to rare occurrences involving mutation losses (e.g., due to recombination-related deletions or to multiple mutations hitting an already mutated genome location[34]), as well as to infrequent transmission phenomena, such as super-infections[65,66] (a discussion on the general limitations of approaches based on phylogenetic trees when dealing with reticulation events is available elsewhere[109–111]).

In this regard, VERSO allows one to identify clonal mutations likely involved in homoplasies, in a similar fashion to the plethora of works on mitochondrial evolution and phylogenetic networks (discussed elsewhere[53,54,56,112–114]). In detail, given the maximum-likelihood phylogenetic tree, VERSO can estimate the variants that are theoretically expected in each sample. By comparing the theoretical observations with the input data, VERSO can estimate the rate of false-positives (i.e., the variants that are observed in the data but are not predicted by VERSO), and false-negatives (i.e., variants that are not observed but predicted). Variants that show particularly high estimated error rates represent candidate homoplasies and are flagged. First, this allows one to pinpoint samples exhibiting homoplastic mutations (see Figures 3 and S8) and, second, to reconstruct an expanded clonal variants tree, in which candidate homoplastic mutations are duplicated after the inference, so to allow the visualization of reticulation events, as proposed by Skála and Zrzavý[112] (see, e.g., Figures S3 and S9).

Furthermore, once this procedure has been completed, the list of flagged variants can include (1) mutations falling in highly mutated regions due to mutational hotspots, (2) phantom mutations (i.e., systematic artifacts generated during sequencing processes[56]), or (3) mutations that have been positively selected in the population (e.g., due to a particular functional advantage).

Since one might be interested in identifying positively selected mutations, VERSO allows one to perform a consecutive analysis that aims at highlighting the mutation-prone regions of the genome and that might be due to mutational hotspots or phantom mutations (see the Supplemental experimental procedures for further details). We finally note that the detection of homoplasies for minor variants requires a different algorithmic procedure, which is detailed in the following.

VERSO step #2: characterization of intra-host genomic diversity

In the second step, VERSO takes into account the VF profiles of groups of samples with the same corrected clonal genotype (identified via VERSO step #1), in order to characterize their intra-host genomic diversity and visualize it on a low-dimensional space. This allows one to highlight patterns of co-occurrence of minor variants, possibly underlying undetected infection events, as well as homoplasies involving; e.g., positively selected variants. Notice that this step requires raw sequencing data and the prior execution of step #1.

Inputs

VERSO step #2 takes as input a n (samples)$\times m$(variants) VF profile matrix, in which each entry includes the VF $\in (0, 1)$ of a given mutation in a certain sample, after filtering out (1) the clonal variants employed in step #1 and (2) the minor variants possibly involved in homoplasies (see below). The variant calling pipeline employed in this work is detailed next.

The algorithmic framework

While it is sound to binarize clonal variant profiles to reconstruct a phylogenetic tree, it is opportune to consider the VF profiles when analyzing intra-host variants, for several reasons. First, VF profiles describe the intra-host genomic diversity of any given host, and this information would be lost during binarization. Second, minor variant profiles might be noisy, due to the relatively low abundance and to the technical limitations of sequencing experiments. Accordingly, such data may possibly include artifacts, which can be partially mitigated during the quality-check phase and by including in the analysis only highly confident variants. However, binarization with arbitrary thresholds might increase the false-positive rate, compromising the accuracy of any downstream analysis. Third, as specified above, the extent of transmission of minor variants among individuals is still partially obscure. The VF of minor variants is, in fact, highly affected by recombination processes, as well as by complex transmission phenomena, involving stochastic fluctuations, bottlenecks, and founder effects, which may lead certain variants changing their VF, not being transmitted, or even becoming clonal in the infected host.[57] The latter issue also suggests that the hypothesis of accumulation of minor variants during infections may not hold and should be relaxed.

For these reasons, VERSO step #2 defines a pairwise genomic distance, computed on the VF profiles, to be used in downstream analyses. The intuition is that samples displaying similar patterns of co-occurrence of minor variants might have a similar quasispecies architecture, thus being at a small evolutionary distance. Accordingly, this might indicate a direct or indirect infection event. In particular, in this work we employed the Bray-Curtis dissimilarity, which is defined as follows: given the ordered VF vectors of two samples (i.e. $v_i = \{VF_1^i, \ldots, VF_r^i, \}$ and $v_j = \{VF_1^j, \ldots, VF_r^j, \}$), the pairwise Bray-Curtis dissimilarity $d(i,j)$ is given by:

$$d(v_i, v_j) = \frac{\sum_{l=1}^{r} \left| VF_l^i - VF_l^j \right|}{\sum_{l=1}^{r} \left| VF_l^i + VF_l^j \right|}. \qquad \text{(Equation 1)}$$

Since this measure weights the pairwise VF dissimilarity on each variant with respect to the sum of the VF of all variants detected in both samples, it can be effectively used to compare the intra-host genomic diversity of samples, as proposed, for instance, by Srinivas et al.[115] However, VERSO allows one to employ different distance metrics on VF profiles, such as correlation or Euclidean distance.

As a design choice, in VERSO, the genomic distance is computed among all samples associated to any given corrected clonal genotype, as inferred in step #1. The rationale is that, in a statistical inference framework modeling a complex interplay involving heterogeneous dynamical processes, it is crucial to stratify samples into homogeneous groups, to reduce the impact of possible

confounding effects.[116] Furthermore, as specified above, due to the distinct properties of clonal and minor variants during transmission, it is reasonable to assume that the event in which certain minor variants and no clonal variants are transmitted from a host to another during the infection is extremely unlikely. Accordingly, the clonal variants employed for the reconstruction of the phylogenetic tree in step #1 are excluded from the computation of the intra-host genomic distance among samples.

In order to produce useful knowledge from the genomic distance discussed above and since, in real-world scenarios, this is a typically complex high-dimensional problem, it is sound to employ state-of-the-art strategies for dimensionality reduction and (sample) clustering, as typically done in single-cell analyses.[117] In this regard, the workflow employed in VERSO ensures high scalability with large datasets, also making it possible to take advantage of effective analysis and visualization features. In detail, the workflow includes three steps: (1) the computation of k-NNG, which can be executed on the original VF matrix, or after applying PCA, to possibly reduce the effect of noisy observations (when the number of samples and variants is sufficiently high); (2) the clustering of samples via either Louvain or Leiden algorithms for community detection;[118] (3) the projection of samples on a low-dimensional space via standard tSNE[68] or UMAP[67] plots.

Outputs

As output, VERSO step #2 delivers both the partitioning of samples in homogeneous clusters and the visualization in a low-dimensional space, also allowing samples to be labeled according to other covariates, such as collection date or geographical location. In the map in Figure 3, for instance, the intra-host genomic diversity of each sample and the genomic distance among samples are projected on the first two UMAP components, whereas samples that are connected by k-NNG edges display similar patterns of co-occurrence of variants. Accordingly, the map shows clusters of samples likely affected by infection events, in which (a fraction of) quasispecies might have been transmitted from one host to another. This represents a major novelty introduced by VERSO and also allows one to effectively visualize the space of VF profiles.

To facilitate the usage, VERSO step #2 is provided as a Python script which employs the SCANPY suite of tools,[59] which is typically used in single-cell analyses and includes a number of highly effective analysis and visualization features.

Additional feature: homoplasy detection on minor variants

Also in the case of minor variants, it is important to pinpoint possible homoplasies that might be due to mutational hotspots, phantom mutations, and convergent variants. Given the phylogenetic model retrieved via step #1, VERSO allows one to flag the variants that are detected in a number of clonal genotypes exceeding a user-defined threshold. In our case, the threshold is equal to 1, meaning that all minor variants found in more than one clonal genotype are flagged.

Such variants are then excluded from the computation of the intra-host genomic distance, prior to the execution of step #2. Furthermore, the list of flagged variants can be investigated as proposed for step #1 (see above), in order to possibly identify mutations involved in positive selection scenarios.

Datasets description

Dataset #1 (Illumina Amplicon sequencing)

We analyzed 3,960 samples from distinct individuals obtained from 22 NCBI BioProjects, which, at the time of writing, are all the publicly available datasets including raw Illumina Amplicon sequencing data. In detail, we selected the following NCBI BioProjects: (1) PRJNA613958, (2) PRJNA614546, (3) PRJNA616147, (4) PRJNA622817, (5) PRJNA623683, (6) PRJNA625551, (7) PRJNA627229, (8) PRJNA627662, (9) PRJNA629891, (10) PRJNA631042, (11) PRJNA633948, (12) PRJNA634119, (13) PRJNA636446, (14) PRJNA636748, (15) PRJNA639066, (16) PRJNA643575, (17) PRJNA645906, (18) PRJNA647448, (19) PRJNA647529, (20) PRJNA650037, (21) PRJNA656534, and (22) PRJNA656695.

Dataset #2 (Illumina RNA sequencing)

We analyzed 2,766 samples from distinct individuals obtained from 22 NCBI BioProjects, which, at the time of writing, are all the publicly available datasets including raw Illumina RNA-sequencing data. In detail, we selected the following NCBI BioProjects: (1) PRJNA601736, (2) PRJNA603194, (3) PRJNA605983, (4) PRJNA607948, (5) PRJNA608651, (6) PRJNA610428, (7) PRJNA615319, (8) PRJNA616446, (9) PRJNA623895, (10) PRJNA624792, (11) PRJNA626526, (12) PRJNA631061, (13) PRJNA636446, (14) PRJNA637892, (15) PRJNA639591, (16) PRJNA639864, (17) PRJNA650134, (18) PRJNA650245, (19) PRJNA655577, (20) PRJNA657938, (21) PRJNA657985, and (22) PRJNA658211.

Contact tracing data

Contact tracing data were obtained from the study presented by Rockett et al.[69] In detail, for 65 samples included in dataset #1 (NCBI BioProject: PRJNA633948), information on households, work institutions, and epidemiological linkages are provided. Thus, it is possible to identify three different contact groups based on institutions regularly frequented by patients and one-household couples. Contact information was employed to assess the relation between the intra-host genomic similarity and the contact dynamics. The results are provided in the main text.

Parameter settings

Parameter settings of variant calling (datasets #1 and #2)

We converted all the samples to FASTQ files using the Sequence Read Archive (SRA) toolkit. Following Bastola et al.,[75] we used Trimmomatic (version 0.39) to remove the nucleotides with low quality score from the RNA sequences with the following settings: LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:40. We then used bwa mem (version 0.7.17) to map reads to SARS-CoV-2-ANC reference genome (Data S1; see Results). We generated sorted BAM files from bwa mem results using SAMtools (version 1.10) and removed duplicates with Picard (version 2.22.2). Variant calling was performed generating mpileup files using SAMtools and then running VarScan (min-var-freq parameter set to 0.01).[119]

We note that it was recently reported that some currently available SARS-CoV-2 datasets exhibit quality issues.[13,120] Accordingly, one should be extremely careful when performing quality check and, especially, when considering low-frequency variants, which might possibly result from sequencing artifacts even in case of high-coverage experiments. In this regard, many effective approaches can be employed to reduce false variants. For instance, the Broad Institute recently updated an effective variant calling pipeline for viral genome data,[121] while new methods for error correction of viral sequencing have been proposed at a widely used website (https://virological.org), which also includes a number of useful up-to-date guidelines and best practices for viral evolution analyses.

In our case, we here employed the following significance filters on variants. In particular, we kept only the mutations (1) showing a VarScan significance p value <0.01 (Fisher's exact test on the read counts supporting reference and variant alleles) and more than 25 reads of support in at least 75% of the samples, (2) displaying a VF >5%. As a result, we selected a list of 15,892 (over 55,280 overall SNVs) highly confident SNVs for dataset #1 and 7,389 (over 53,354) for dataset #2.

High-quality variants were then mapped on SARS-CoV-2 coding sequences (CDSs) via a custom R script, also by highlighting synonymous/nonsynonymous states and amino acid substitutions for the related open reading frame (ORF) product. In particular, we translated reference and mutated CDSs with the seqinr R package to obtain the relative amino acid sequences, which we compared to assess the effect of each nucleotide variation in terms of amino acid substitution.

We finally note that availability of the cycle threshold (Ct) values generated by qPCR and the related quantification of the amounts of viral transcripts would be very useful to characterize samples with high viral load, yet this information is not available for the considered datasets.

Quality check (datasets #1 and #2)

In order to select high-quality samples, we selected only those exhibiting high coverage and in particular those with at least 25 reads in more than 90% of the SARS-CoV-2-ANC genome. In addition, we filtered out all samples exhibiting more than 100 minor variants (VF $\leq$ 90%).

We finally excluded samples SRR11597146 and SRR11476447 from dataset #1, as the first sample displays zero SNVs and the second one reports an unfeasible collection date (i.e., 30th Jan. 2019).

After the quality-check filters, 2,906 samples of dataset #1 are left for downstream analyses, in which 10,571 distinct high-quality SNVs are observed, and 1,438 samples are left for dataset #2, with 6,143 high-quality SNVs.

Parameter settings of VERSO *(datasets #1 and #2)*

The phylogenomic analysis via VERSO step #1 was performed on datasets #1 and #2 by considering only clonal variants (VF > 90%) detected in at least 3% of the samples. A grid search comprising 16 different error rates was employed (see Table S3). Samples with the same corrected clonal genotype were grouped in polytomies in the final phylogenetic models.

The analysis of the intra-host genomic diversity via VERSO step #2 was performed by considering the VF profiles of all samples, by excluding (1) the clonal variants employed in the phylogenomic reconstruction via VERSO step #1, (2) the minor variants involved in homoplasies (i.e., observed in more than one clonal genotype returned by VERSO step #1). Missing values (NA) were imputed to 0 for downstream analysis. A number of principal components equals to 10 was employed in PCA step, prior to the computation of the k-NNG (k = 10) on the Bray-Curtis dissimilarity of VF profiles. Leiden algorithm was applied with resolution = 1 (see Table S3 for the parameter settings of VERSO employed in the case studies).

Parameter settings of simulations

In order to compare the performance of VERSO step #1 with competing phylogenomic tools (i.e., IQ-TREE[10] and BEAST 2[22]), we performed extensive simulations via msprime,[70] which simulates a backwards-in-time coalescent model.

In particular, we simulated 20 distinct evolutionary processes, with the following parameters: n = 1,000 total samples, effective population size N_e = 0.5 (i.e., haploid population), mutational rate $M = 2 \times 10^{-6}$ mutations per site per generation, and a genome of length L = 29,903 bases. Such parameters were chosen to roughly approximate the mutational rate currently estimated for SARS-CoV-2 (i.e., $M \approx 10^{-3}$ mutations per site per year and $\approx 10^{-3}\frac{\text{generation}}{\text{year}}122$) and to obtain a number of clonal mutations (in the range 15–30) that is comparable with the one observed in the real-word scenarios (see the case studies). As output, msprime returns a phylogenetic tree representing the genealogy between the samples, the genotype of all samples (i.e., the leaves of the tree), and the location of all mutations.

The genotypes of the samples were then inflated with different levels of noise, with false-positive rate α and false-negative rate β (see the parameter settings in Table S1), in order to assess the performance of the methods in conditions of noisy observations and possible sequencing issues. Finally, we subsampled all datasets to obtain two distinct samples sizes (500 and 1,000 samples), in order to test the robustness of methods in conditions of sampling limitations.

The parameters of the phylogenetic methods employed in the comparative assessment are reported in the Supplemental experimental procedures (Table S2).

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.patter.2021.100212.

ACKNOWLEDGMENTS

This work was partially supported by the Elixir Italian Chapter and the SysBioNet project, a Ministero dell'Istruzione, dell'UniversitÃ e della Ricerca initiative for the Italian Roadmap of European Strategy Forum on Research Infrastructures, and by AIRC-IG grant 22082. Partial support was also provided by the CRUK/AECC/AIRC Accelerator Award #22790, Single-cell Cancer Evolution in the Clinic. We thank Giulio Caravagna, Chiara Damiani, Lucrezia Patruno, and Francesco Craighero for helpful discussions. We also thank David Posada for interesting suggestions on the preliminary version of the manuscript.

AUTHOR CONTRIBUTIONS

D.R., F.A., D.M., A.G., and R.P. designed the approach. D.R., F.A., D.M., and A.G. defined, implemented, and executed the computational methods. D.R., F.A., D.M., and A.G. performed the simulations. D.R., F.A., D.M., C.G.-P., M.A., A.G., and R.P. analyzed the data and interpreted the results. R.P. supervised the experimental data analysis. A.G. and D.R. supervised the computational analysis. A.G. and R.P. drafted the manuscript, which all authors discussed, reviewed, and approved.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 12, 2020
Revised: November 30, 2020
Accepted: January 22, 2021
Published: January 28, 2021

REFERENCES

1. Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature *579*, 270–273, https://doi.org/10.1038/s41586-020-2012-7.
2. Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., et al. (2020). A new coronavirus associated with human respiratory disease in China. Nature *579*, 265–269, https://doi.org/10.1038/s41586-020-2008-3.
3. Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., and Garry, R.F. (2020). The proximal origin of SARS-CoV-2. Nat. Med. *26*, 450–452, https://doi.org/10.1038/s41591-020-0820-9.
4. Xiao, K., Zhai, J., Feng, Y., Zhou, N., Zhang, X., Zou, J.J., Li, N., Guo, Y., Li, X., Shen, X., et al. (2020). Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature *583*, 286–289, https://doi.org/10.1038/s41586-020-2313-x.
5. Deng, X., Gu, W., Federman, S., du Plessis, L., Pybus, O.G., Faria, N.R., Wang, C., Yu, G., Bushnell, B., Pan, C.Y., et al. (2020). Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. Science *369*, 582–587, https://doi.org/10.1126/science.abb9263.
6. Grubaugh, N.D., Petrone, M.E., and Holmes, E.C. (2020). We shouldn't worry when a virus mutates during disease outbreaks. Nat. Microbiol. *5*, 529–530, https://doi.org/10.1038/s41564-020-0690-4.
7. Grenfell, B.T., Pybus, O.G., Gog, J.R., Wood, J.L., Daly, J.M., Mumford, J.A., and Holmes, E.C. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. Science *303*, 327–332, https://doi.org/10.1126/science.1090727.
8. Poon, L.L., Song, T., Rosenfeld, R., Lin, X., Rogers, M.B., Zhou, B., Sebra, R., Halpin, R.A., Guan, Y., Twaddle, A., et al. (2016). Quantifying influenza virus diversity and transmission in humans. Nat. Genet. *48*, 195, https://doi.org/10.1038/ng.3479.
9. Shu, Y., and McCauley, J. (2017). GISAID: global initiative on sharing all influenza data–from vision to reality. Euro Surveill. *22*, https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.
10. Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274, https://doi.org/10.1093/molbev/msu300.
11. Volz, E.M., Koelle, K., and Bedford, T. (2013). Viral phylodynamics. PLoS Comput. Biol. *9*, https://doi.org/10.1371/journal.pcbi.1002947.
12. Faria, N.R., Quick, J., Claro, I., Theze, J., de Jesus, J.G., Giovanetti, M., Kraemer, M.U., Hill, S.C., Black, A., da Costa, A.C., et al. (2017). Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature *546*, 406–410, https://doi.org/10.1038/nature22401.
13. Worobey, M., Pekar, J., Larsen, B.B., Nelson, M.I., Hill, V., Joy, J.B., Rambaut, A., Suchard, M.A., Wertheim, J.O., and Lemey, P. (2020). The emergence of SARS-CoV-2 in Europe and North America. Science *370*, 564–570, https://doi.org/10.1126/science.abc8169.
14. Bandelt, H.J., Forster, P., and Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. *16*, 37–48, https://doi.org/10.1093/oxfordjournals.molbev.a026036.
15. O'Meara, B.C. (2012). Evolutionary inferences from phylogenies: a review of methods. Annu. Rev. Ecol. Evol. Syst. *43*, 267–285, https://doi.org/10.1146/annurev-ecolsys-110411-160331.

16. Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. *61*, 539–542, https://doi.org/10.1093/sysbio/sys029.

17. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313, https://doi.org/10.1093/bioinformatics/btu033.

18. Didelot, X., Fraser, C., Gardy, J., and Colijn, C. (2017). Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. Mol. Biol. Evol. *34*, 997–1007, https://doi.org/10.1093/molbev/msw275.

19. Skums, P., Zelikovsky, A., Singh, R., Gussler, W., Dimitrova, Z., Knyazev, S., Mandric, I., Ramachandran, S., Campo, D., Jha, D., et al. (2018). Quentin: reconstruction of disease transmissions from viral quasispecies genomic data. Bioinformatics *34*, 163–170, https://doi.org/10.1093/bioinformatics/btx402.

20. De Maio, N., Worby, C.J., Wilson, D.J., and Stoesser, N. (2018). Bayesian reconstruction of transmission within outbreaks using genomic variants. PLoS Comput. Biol. *14*, e1006117, https://doi.org/10.1093/sysbio/sys029.

21. Kosakovsky Pond, S.L., Weaver, S., Leigh Brown, A.J., and Wertheim, J.O. (2018). HIV-TRACE (TRAnsmission Cluster Engine): a tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens. Mol. Biol. Evol. *35*, 1812–1819, https://doi.org/10.1093/molbev/msy016.

22. Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., et al. (2019). Beast 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. *15*, e1006650, https://doi.org/10.1371/journal.pcbi.1006650.

23. Lai, A., Bergna, A., Acciarri, C., Galli, M., and Zehender, G. (2020). Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2. J. Med. Virol. *92*, 675–679, https://doi.org/10.1002/jmv.25723.

24. Forster, P., Forster, L., Renfrew, C., and Forster, M. (2020). Phylogenetic network analysis of SARS-CoV-2 genomes. Proc. Natl. Acad. Sci. U S A *117*, 9241–9243, https://doi.org/10.1073/pnas.2004999117.

25. Dong, R., Pei, S., Yin, C., He, R.L., and Yau, S.S.T. (2020). Analysis of the hosts and transmission paths of SARS-CoV-2 in the COVID-19 outbreak. Genes *11*, 637, https://doi.org/10.3390/genes11060637.

26. Nakhleh, L. (2009). A metric on the space of reduced phylogenetic networks. IEEE/ACM Trans. Comput. Biol. Bioinform. *7*, 218–222, https://doi.org/10.1126/10.1109/TCBB.2009.2.

27. Yuan, K., Sakoparnig, T., Markowetz, F., and Beerenwinkel, N. (2015). BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. Genome Biol. *16*, 36, https://doi.org/10.1186/s13059-015-0592-6.

28. Villabona-Arenas, C.J., Hanage, W.P., and Tully, D.C. (2020). Phylogenetic interpretation during outbreaks requires caution. Nat. Microbiol. *5*, 876–877, https://doi.org/10.1038/s41564-020-0738-5.

29. Mavian, C., Marini, S., Manes, C., Capua, I., Prosperi, M., and Salemi, M. (2020). Regaining perspective on SARS-CoV-2 molecular tracing and its implications. medRxiv. https://doi.org/10.1101/2020.03.16.20034470.

30. Domingo, E., Martínez-Salas, E., Sobrino, F., de la Torre, J.C., Portela, A., Ortín, J., López-Galindez, C., Pérez-Breña, P., Villanueva, N., Nájera, R., et al. (1985). The quasispecies (extremely heterogeneous) nature of viral RNA genome populations: biological relevance–a review. Gene *40*, 1–8, https://doi.org/10.1016/0378-1119(85)90017-4.

31. Domingo, E., Sheldon, J., and Perales, C. (2012). Viral quasispecies evolution. Microbiol. Mol. Biol. Rev. *76*, 159–216, https://doi.org/10.1128/MMBR.05023-11.

32. Acevedo, A., Brodsky, L., and Andino, R. (2014). Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature *505*, 686–690, https://doi.org/10.1038/nature12861.

33. Novella, I.S., Domingo, E., and Holland, J.J. (1995). Rapid viral quasispecies evolution: implications for vaccine and drug strategies. Mol. Med. Today *1*, 248–253, https://doi.org/10.1016/s1357-4310(95)91551-6.

34. Simon-Loriere, E., and Holmes, E.C. (2011). Why do RNA viruses recombine? Nat. Rev. Microbiol. *9*, 617–626, https://doi.org/10.1038/nrmicro2614.

35. Graudenzi, A., Maspero, D., Angaroni, F., Piazza, R., and Ramazzotti, D. (2020). Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity. iScience *24*, 102116.

36. Shen, Z., Xiao, Y., Kang, L., Ma, W., Shi, L., Zhang, L., Zhou, Z., Yang, J., Zhong, J., Yang, D., et al. (2020). Genomic diversity of SARS-CoV-2 in coronavirus disease 2019 patients. Clin. Infect. Dis. *71*, https://doi.org/10.1093/cid/ciaa203.

37. Wölfel, R., Corman, V.M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M.A., Niemeyer, D., Jones, T.C., Vollmar, P., Rothe, C., et al. (2020). Virological assessment of hospitalized patients with COVID-2019. Nature *581*, 465–469, https://doi.org/10.1038/s41586-020-2196-x.

38. Capobianchi, M.R., Rueca, M., Messina, F., Giombini, E., Carletti, F., Colavita, F., Castilletti, C., Lalle, E., Bordi, L., Vairo, F., et al. (2020). Molecular characterization of SARS-CoV-2 from the first case of COVID-19 in Italy. Clin. Microbiol. Infect. *26*, 954–956, https://doi.org/10.1016/j.cmi.2020.03.025.

39. Rose, R., Nolan, D.J., Moot, S., Feehan, A., Cross, S., Garcia-Diaz, J., and Lamers, S.L. (2020). Intra-host site-specific polymorphisms of SARS-CoV-2 is consistent across multiple samples and methodologies. medRxiv. https://doi.org/10.1101/2020.04.24.20078691.

40. Lu, J., du Plessis, L., Liu, Z., Hill, V., Kang, M., Lin, H., Sun, J., François, S., Kraemer, M.U., Faria, N.R., et al. (2020). Genomic epidemiology of SARS-CoV-2 in Guangdong province, China. Cell *181*, 997–1003.e9, https://doi.org/10.1016/j.cell.2020.04.023.

41. Lythgoe, K.A., Hall, M.D., Ferretti, L., de Cesare, M., MacIntyre-Cockett, G., Trebes, A., Andersson, M., Otecko, N., Wise, E.L., Moore, N., et al. (2020). Shared SARS-CoV-2 diversity suggests localised transmission of minority variants. bioRxiv. https://doi.org/10.1101/2020.05.28.118992.

42. Seemann, T., Lane, C.R., Sherry, N.L., Duchene, S., Gonçalves da Silva, A., Caly, L., Sait, M., Ballard, S.A., Horan, K., Schultz, M.B., et al. (2020). Tracking the COVID-19 pandemic in Australia using genomics. Nat. Commun. *11*, 4376, https://doi.org/10.1038/s41467-020-18314-x.

43. Popa, A., Genger, J.W., Nicholson, M.D., Penz, T., Schmid, D., Aberle, S.W., Agerer, B., Lercher, A., Endler, L., Colaço, H., et al. (2020). Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. Sci. Transl. Med. *12*, https://doi.org/10.1126/scitranslmed.abe2555.

44. Miralles, R., Gerrish, P.J., Moya, A., and Elena, S.F. (1999). Clonal interference and the evolution of RNA viruses. Science *285*, 1745–1747, https://doi.org/10.1126/science.285.5434.1745.

45. Xu, D., Zhang, Z., and Wang, F.S. (2004). SARS-associated coronavirus quasispecies in individual patients. N. Engl. J. Med. *350*, 1366–1367, https://doi.org/10.1056/NEJMc032421.

46. Wright, C.F., Morelli, M.J., Thébaud, G., Knowles, N.J., Herzyk, P., Paton, D.J., Haydon, D.T., and King, D.P. (2011). Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. J. Virol. *85*, 2266–2275, https://doi.org/10.1128/JVI.01396-10.

47. Park, D., Huh, H.J., Kim, Y.J., Son, D.S., Jeon, H.J., Im, E.H., Kim, J.W., Lee, N.Y., Kang, E.S., Kang, C.I., et al. (2016). Analysis of intrapatient heterogeneity uncovers the microevolution of middle east respiratory syndrome coronavirus. Mol. Case Stud. *2*, a001214, https://doi.org/10.1101/mcs.a001214.

48. Ni, M., Chen, C., Qian, J., Xiao, H.X., Shi, W.F., Luo, Y., Wang, H.Y., Li, Z., Wu, J., Xu, P.S., et al. (2016). Intra-host dynamics of Ebola virus during

2014. Nat. Microbiol. *1*, 16151, https://doi.org/10.1038/nmicrobiol.2016.151.

49. Ramazzotti, D., Caravagna, G., Olde Loohuis, L., Graudenzi, A., Korsunsky, I., Mauri, G., Antoniotti, M., and Mishra, B. (2015). CAPRI: efficient inference of cancer progression models from cross-sectional data. Bioinformatics *31*, 3016–3026, https://doi.org/10.1093/bioinformatics/btv296.

50. Beerenwinkel, N., Schwarz, R.F., Gerstung, M., and Markowetz, F. (2014). Cancer evolution: mathematical models and computational inference. Syst. Biol. *64*, e1–e25, https://doi.org/10.1093/sysbio/syu081.

51. Caravagna, G., Graudenzi, A., Ramazzotti, D., Sanz-Pamplona, R., De Sano, L., Mauri, G., Moreno, V., Antoniotti, M., and Mishra, B. (2016). Algorithmic methods to infer the evolutionary trajectories in cancer progression. Proc. Natl. Acad. Sci. U S A *113*, E4025–E4034, https://doi.org/10.1073/pnas.1520213113.

52. Schwartz, R., and Schäffer, A.A. (2017). The evolution of tumour phylogenetics: principles and practice. Nat. Rev. Genet. *18*, 213–229, https://doi.org/10.1038/nrg.2016.170.

53. Posada, D., and Crandall, K.A. (2001). Intraspecific gene genealogies: trees grafting into networks. Trends Ecol. Evol. *16*, 37–45, https://doi.org/10.1016/S0169-5347(00)02026-7.

54. Boc, A., Diallo, A.B., and Makarenkov, V. (2012). T-rex: a web server for inferring, validating and visualizing phylogenetic trees and networks. Nucleic Acids Res. *40* (W1), W573–W579, https://doi.org/10.1093/nar/gks485.

55. Bull, J., Badgett, M., Wichman, H.A., Huelsenbeck, J.P., Hillis, D.M., Gulati, A., Ho, C., and Molineux, I. (1997). Exceptional convergent evolution in a virus. Genetics *147*, 1497–1507. https://www.genetics.org/content/147/4/1497.

56. Bandelt, H.J., Quintana-Murci, L., Salas, A., and Macaulay, V. (2002). The fingerprint of phantom mutations in mitochondrial DNA data. Am. J. Hum. Genet. *71*, 1150–1160, https://doi.org/10.1086/344397.

57. Gutierrez, S., Yvon, M., Pirolles, E., Garzo, E., Fereres, A., Michalakis, Y., and Blanc, S. (2012). Circulating virus load determines the size of bottlenecks in viral populations progressing within a host. PLoS Pathog. *8*, 1–10, https://doi.org/10.1371/journal.ppat.1003009.

58. Firestone, S.M., Hayama, Y., Bradhurst, R., Yamamoto, T., Tsutsui, T., and Stevenson, M.A. (2019). Reconstructing foot-and-mouth disease outbreaks: a methods comparison of transmission network models. Sci. Rep. *9*, 4809, https://doi.org/10.1038/s41598-019-41103-6.

59. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15, https://doi.org/10.1186/s13059-017-1382-0.

60. Prosperi, M.C., and Salemi, M. (2012). Qure: software for viral quasispecies reconstruction from next-generation sequencing data. Bioinformatics *28*, 132–133, https://doi.org/10.1093/bioinformatics/btr627.

61. Giallonardo, F.D., Töpfer, A., Rey, M., Prabhakaran, S., Duport, Y., Leemann, C., Schmutz, S., Campbell, N.K., Joos, B., Lecca, M.R., et al. (2014). Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. Nucleic Acids Res. *42*, e115, https://doi.org/10.1093/nar/gku537.

62. Töpfer, A., Marschall, T., Bull, R.A., Luciani, F., Schönhuth, A., and Beerenwinkel, N. (2014). Viral quasispecies assembly via maximal clique enumeration. PLoS Comput. Biol. *10*, 1–10, https://doi.org/10.1371/journal.pcbi.1003515.

63. Barik, S., Das, S., and Vikalo, H. (2018). QSdpR: viral quasispecies reconstruction via correlation clustering. Genomics *110*, 375–381, https://doi.org/10.1016/j.ygeno.2017.12.007.

64. Knyazev, S., Hughes, L., Skums, P., and Zelikovsky, A. (2020). Epidemiological data analysis of viral quasispecies in the next-generation sequencing era. Brief. Bioinformatics, bbaa101, https://doi.org/10.1093/bib/bbaa101.

65. Alizon, S. (2013). Co-infection and super-infection models in evolutionary epidemiology. Interface Focus *3*, 20130031, https://doi.org/10.1098/rsfs.2013.0031.

66. Garcia-Vidal, C., Sanjuan, G., Moreno-García, E., Puerta-Alcalde, P., Garcia-Pouton, N., Chumbita, M., Fernandez-Pittol, M., Pitart, C., Inciarte, A., Bodro, M., et al. (2020). Incidence of co-infections and super-infections in hospitalized patients with COVID-19: a retrospective cohort study. Clin. Microbiol. Infect. *27*, 83–88, https://doi.org/10.1016/j.cmi.2020.07.041.

67. McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: uniform manifold approximation and projection. J. Open Source Softw. *3*, 861, https://doi.org/10.21105/joss.00861.

68. van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. J. Machine Learn. Res. *9*, 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html.

69. Rockett, R.J., Arnott, A., Lam, C., Sadsad, R., Timms, V., Gray, K.A., Eden, J.S., Chang, S., Gall, M., Draper, J., et al. (2020). Revealing COVID-19 transmission in Australia by SARS-CoV-2 genome sequencing and agent-based modeling. Nat. Med. *26*, 1398–1404, https://doi.org/10.1038/s41591-020-1000-7.

70. Kelleher, J., Etheridge, A.M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput. Biol. *12*, 1–22, https://doi.org/10.1371/journal.pcbi.1004842.

71. Wakeley, J. (2009). Coalescent Theory: An Introduction (Roberts and Company Publishers).

72. Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R.A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics *34*, 4121–4123, https://doi.org/10.1093/bioinformatics/bty407.

73. Kuhner, M.K., and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. *11*, 459–468, https://doi.org/10.1093/oxfordjournals.molbev.a040126.

74. Steel, M.A., and Penny, D. (1993). Distributions of tree comparison metrics–some new results. Syst. Biol. *42*, 126–141, https://doi.org/10.1093/sysbio/42.2.126.

75. Bastola, A., Sah, R., Rodriguez-Morales, A.J., Lal, B.K., Jha, R., Ojha, H.C., Shrestha, B., Chu, D.K., Poon, L.L., Costello, A., et al. (2020). The first 2019 novel coronavirus case in Nepal. Lancet Infect. Dis. *20*, 279–280, https://doi.org/10.1016/S1473-3099(20)30067-0.

76. Boni, M.F., Lemey, P., Jiang, X., Lam, T.T.Y., Perry, B.W., Castoe, T.A., Rambaut, A., and Robertson, D.L. (2020). Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the covid-19 pandemic. Nat. Microbiol. *5*, 1408–1417, https://doi.org/10.1038/s41564-020-0771-4.

77. Li, X., Giorgi, E.E., Marichannegowda, M.H., Foley, B., Xiao, C., Kong, X.P., Chen, Y., Gnanakaran, S., Korber, B., and Gao, F. (2020). Emergence of SARS-CoV-2 through recombination and strong purifying selection. Sci. Adv. *6*, eabb9153, https://doi.org/10.1126/sciadv.abb9153.

78. Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., and Pybus, O.G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. *5*, 1403–1407, https://doi.org/10.1038/s41564-020-0770-5.

79. O'Toole, A., McCrone, J., and Scher, E. (2020). pangolin 2.0. https://github.com/cov-lineages/pangolin.

80. Rambaut, A. (2009). Figtree v1. 3.1. http://tree.bio.ed.ac.uk/software/figtree/.

81. Argimón, S., Abudahab, K., Goater, R.J., Fedosejev, A., Bhai, J., Glasner, C., Feil, E.J., Holden, M.T., Yeats, C.A., Grundmann, H., et al. (2016). Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microb. Genomics *2*, e000093, https://doi.org/10.1099/mgen.0.000093.

82. Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., Duan, Y., Zhang, H., Wang, Y., Qian, Z., et al. (2020). On the origin and continuing evolution of SARS-CoV-2. Natl. Sci. Rev. *7*, 1012–1023, https://doi.org/10.1093/nsr/nwaa036.

83. Zhang, X., Tan, Y., Ling, Y., Lu, G., Liu, F., Yi, Z., Jia, X., Wu, M., Shi, B., Xu, S., et al. (2020). Viral and host factors related to the clinical outcome of covid-19. Nature *583*, 437–440, https://doi.org/10.1038/s41586-020-2355-0.

84. Volz, E.M., Hill, V., McCrone, J.T., Price, A., Jorgensen, D., O'Toole, A., Southgate, J.A., Johnson, R., Jackson, B., Nascimento, F.F., et al. (2020). Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. Cell. https://doi.org/10.1016/j.cell.2020.11.020.

85. Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E.E., Bhattacharya, T., Foley, B., et al. (2020). Tracking changes in SARS-CoV-2 spike: evidence that d614g increases infectivity of the covid-19 virus. Cell *182*, 812–827, https://doi.org/10.1016/j.cell.2020.06.043.

86. Yurkovetskiy, L., Wang, X., Pascal, K.E., Tomkins-Tinch, C., Nyalile, T., Wang, Y., Baum, A., Diehl, W.E., Dauphin, A., Carbone, C., et al. (2020). Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. Cell *183*, 739–751, https://doi.org/10.1016/j.cell.2020.09.032.

87. Grubaugh, N.D., Hanage, W.P., and Rasmussen, A.L. (2020). Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear. Cell *182*, 794–795, https://doi.org/10.1016/j.cell.2020.06.040.

88. van Dorp, L., Richard, D., Tan, C.C., Shaw, L.P., Acman, M., and Balloux, F. (2020). No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. bioRxiv. https://doi.org/10.1101/2020.05.21.108506.

89. van Dorp, L., Acman, M., Richard, D., Shaw, L.P., Ford, C.E., Ormond, L., Owen, C.J., Pang, J., Tan, C.C., Boshier, F.A., et al. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect. Genet. Evol. *83*, 104351, https://doi.org/10.1016/j.meegid.2020.104351.

90. Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Röhl, A., Salas, A., Oppenheimer, S., Macaulay, V., and Richards, M.B. (2009). Correcting for purifying selection: an improved human mitochondrial molecular clock. Am. J. Hum. Genet. *84*, 740–759, https://doi.org/10.1016/j.ajhg.2009.05.001.

91. Letko, M., Marzi, A., and Munster, V. (2020). Functional assessment of cell entry and receptor usage for RNA and other lineage b betacoronaviruses. Nat. Microbiol. *5*, 562–569, https://doi.org/10.1038/s41564-020-0688-y.

92. Xia, S., Xu, W., Wang, Q., Wang, C., Hua, C., Li, W., Lu, L., and Jiang, S. (2018). Peptide-based membrane fusion inhibitors targeting hcov-229e spike protein HR1 and HR2 domains. Int. J. Mol. Sci. *19*, 487, https://doi.org/10.3390/ijms19020487.

93. Xia, S., Yan, L., Xu, W., Agrawal, A.S., Algaissi, A., Tseng, C.T.K., Wang, Q., Du, L., Tan, W., Wilson, I.A., et al. (2019). A pan-coronavirus fusion inhibitor targeting the hr1 domain of human coronavirus spike. Sci. Adv. *5*, https://doi.org/10.1126/sciadv.aav4580.

94. Li, X., Wang, W., Zhao, X., Zai, J., Zhao, Q., Li, Y., and Chaillon, A. (2020). Transmission dynamics and evolutionary history of 2019-nCoV. J. Med. Virol. *92*, 501–511, https://doi.org/10.1002/jmv.25701.

95. Fay, J.C., Wyckoff, G.J., and Wu, C.I. (2001). Positive and negative selection on the human genome. Genetics *158*, 1227–1234. https://www.genetics.org/content/158/3/1227.

96. Sobel Leonard, A., Weissman, D.B., Greenbaum, B., Ghedin, E., and Koelle, K. (2017). Transmission bottleneck size estimation from pathogen deep-sequencing data, with an application to human influenza a virus. J. Virol. *91*, https://doi.org/10.1128/JVI.00171-17.

97. Ghafari, M., Lumby, C.K., Weissman, D.B., and Illingworth, C.J. (2020). Inferring transmission bottleneck size from viral sequence data using a novel haplotype reconstruction method. J. Virol. https://doi.org/10.1128/JVI.00014-20.

98. Monsion, B., Froissart, R., Michalakis, Y., and Blanc, S. (2008). Large bottleneck size in cauliflower mosaic virus populations during host plant colonization. PLoS Pathog. *4*, 1–7, https://doi.org/10.1371/journal.ppat.1000174.

99. Lequime, S., Fontaine, A., Ar Gouilh, M., Moltini-Conclois, I., and Lambrechts, L. (2016). Genetic drift, purifying selection and vector genotype shape dengue virus intra-host genetic diversity in mosquitoes. PLoS Genet. *12*, 1–24, https://doi.org/10.1371/journal.pgen.1006111.

100. Gardy, J.L., and Loman, N.J. (2018). Towards a genomics-informed, real-time, global pathogen surveillance system. Nat. Rev. Genet. *19*, 9, https://doi.org/10.1038/nrg.2017.88.

101. Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., and Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science *368*, 489–493, https://doi.org/10.1126/science.abb3221.

102. Volz, E.M., Pond, S.L.K., Ward, M.J., Brown, A.J.L., and Frost, S.D. (2009). Phylodynamics of infectious disease epidemics. Genetics *183*, 1421–1430, https://doi.org/10.1534/genetics.109.106021.

103. Bedford, T., Greninger, A.L., Roychoudhury, P., Starita, L.M., Famulare, M., Huang, M.L., Nalla, A., Pepper, G., Reinhardt, A., Xie, H., et al. (2020). Cryptic transmission of SARS-CoV-2 in Washington state. Science *370*, 571–575, https://doi.org/10.1126/science.abc0523.

104. Gomez-Carballa, A., Bello, X., Pardo-Seco, J., Martinon-Torres, F., and Salas, A. (2020). Mapping genome variation of SARS-CoV-2 worldwide highlights the impact of COVID-19 super-spreaders. Genome Res. http://www.genome.org/cgi/doi/10.1101/gr.266221.120.

105. Ramazzotti, D., Angaroni, F., Maspero, D., Ascolani, G., Castiglioni, I., Piazza, R., Antoniotti, M., and Graudenzi, A. (2020). Longitudinal cancer evolution from single cells. bioRxiv. https://doi.org/10.1101/2020.01.14.906453.

106. Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics *61*, 893–903. https://www.genetics.org/content/61/4/893.

107. Gusfield, D. (1991). Efficient algorithms for inferring evolutionary trees. Networks *21*, 19–28, https://doi.org/10.1002/net.3230210104.

108. Huson, D.H., and Scornavacca, C. (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst. Biol. *61*, 1061–1067, https://doi.org/10.1093/sysbio/sys062.

109. Mindell, D.P. (1993). Merger of taxa and the definition of monophyly (reply to Jan Zrzavý and Zdeněk Skála). Biosystems *31*, 130–133, https://doi.org/10.1016/0303-2647(93)90041-A.

110. Zrzavý, J., and Skála, Z. (1993). Holobionts, hybrids, and cladistic classification (reply to David P. Mindell). Biosystems *31*, 127–130, https://doi.org/10.1016/0303-2647(93)90040-J.

111. Chan, J.M., Carlsson, G., and Rabadan, R. (2013). Topology of viral evolution. Proc. Natl. Acad. Sci. U S A *110*, 18566–18571, https://doi.org/10.1073/pnas.1313480110.

112. Skála, Z., and Zrzavý, J. (1994). Phylogenetic reticulations and cladistics: discussion of methodological concepts. Cladistics *10*, 305–313, https://doi.org/10.1111/j.1096-0031.1994.tb00180.x.

113. Brandstätter, A., Sänger, T., Lutz-Bonengel, S., Parson, W., Béraud-Colomb, E., Wen, B., Kong, Q.P., Bravi, C.M., and Bandelt, H.J. (2005). Phantom mutation hotspots in human mitochondrial DNA. Electrophoresis *26*, 3414–3429, https://doi.org/10.1002/elps.200500307.

114. Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.J., Kronenberg, F., Salas, A., and Schönherr, S. (2016). HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. *44* (W1), W58–W63, https://doi.org/10.1093/nar/gkw233.

115. Srinivas, G., Möller, S., Wang, J., Künzel, S., Zillikens, D., Baines, J.F., and Ibrahim, S.M. (2013). Genome-wide mapping of gene–microbiota

interactions in susceptibility to autoimmune skin blistering. Nat. Commun. *4*, 2462, https://doi.org/10.1038/ncomms3462.

116. Pearl, J. (2009). Causality (Cambridge University Press).

117. Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. *15*, https://doi.org/10.15252/msb.20188746.

118. Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. *9*, 1–12, https://doi.org/10.1038/s41598-019-41695-z.

119. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). Varscan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. *22*, 568–576, https://doi.org/10.1101/gr.129684.111.

120. De Maio, N., Walker, C., Borge, R., Weilguny, L., Slodkowick, G., and Goldmand, N. (2020). Issues with SARS-CoV-2 sequencing data. https://virological.org/.

121. Park, D., Tomkins-Tinch, C., Ye, S., Jungreis, I., Shlyakhter, I., Metsky, H., Hanna, Lin, M., Le, V., Lin, A., et al. (2019). broadinstitute/viral-ngs: v1.25.0. https://doi.org/10.5281/zenodo.3509008.

122. Bar-On, Y.M., Flamholz, A., Phillips, R., and Milo, R. (2020). Science forum: SARS-CoV-2 (COVID-19) by the numbers. eLife *9*, e57309, https://doi.org/10.7554/eLife.57309.

3.2.1.1 Improving the evolution inference with COB-tree

Background. Another tranche of research on phylogenomic inference from mutational profiles regarded the in-depth analysis of the solution space.

In brief, the approximate estimate of the computational complexity of the statistical frameworks proposed in LACE (paper P#5) and VERSO (paper P#6) is $\mathcal{O}(nm^3 \log(m))$ to reach MCMC convergence, where n is the number of cells/samples (observations), and m is the number of mutations (variables), as initially discussed in [97]. Thus, it is evident that the complexity mainly depends on the number of mutations included in the model. Clearly, as this number increases it may becomes unfeasible to reach convergence. In addition, in many cases, the algorithm may return equivalent solutions, which share the same likelihood value, but with different topologies.

For these reasons, we investigated the possibility of summarising the collection of tree solutions explored during the MCMC inference, so to return a unique *Consensus Optimum Branching* tree (COB tree), instead of the Maximum Likelihood one (ML tree). More in detail, the goal is to design an algorithm that takes as input a collection of trees sampled during the MCMC of an algorithm for phylogenetic inference and exploits the regularities of the solution space to return a unique COB tree.

Note that similar approaches have already been employed in classical phylogenetic studies. For example, BEAST 2 [77] uses the Maximum Clade Credibility method, instead in [161] the authors proposed the Majority Rule. Such approaches could not be directly employed in our analyses due to the intrinsic differences between phylogenetic trees and clonal trees. In particular, the former are binary trees, and thus the number of edges is fixed and depends on the number of samples. This is not valid in clonal trees where any node could have an arbitrary number of outgoing edges.

We also point out that the opportunity of computing a consensus tree in cancer phylogeny is debated. For example, authors in [169] argue that summary methods returning only a single tree may not accurately represent the topological features of the solution space. By assuming that the solution space is rugged and includes different local minima related to clonal trees displaying distant topologies, the authors suggest to cluster the tree solutions and, successively, apply a summary method for each cluster. Such approach is not computationally feasible for our goal, because the clustering step is limited to a small number of (small) trees, which is orders of magnitudes lower than the trees sampled during an MCMC. We also note that, in their conclusion, the authors state that a proper characterization of the solution space under an error model of single-cell mutation profiles has not been presented yet.

For these many reasons, in this preliminary work, we first characterized the search space of the clonal trees inferred via SCITE [113] from synthetic datasets generated from a number of distinct topologies. Our synthetic experiments show that the trees sampled during each inference describe a solution space with an apparent unique global minimum.

Thus, in this case, applying a consensus approach that does not depend on a clustering step seems a reasonable choice. Notice that the evaluation of the solution space should also be performed considering real data, to prove the reliability of the simulated ones, and so the effectiveness of the consensus approach proposed here.

COB-tree: a new algorithm for phylogenetic inference. On this basis, we conceived a new algorithm for the inference of phylogenetic trees from binary mutational profiles.

The general idea is to reconstruct a unique consensus tree, obtained by exploiting the solutions explored during the MCMC of a generic algorithm for phylogenetic reconstruction, such as LACE, VERSO or SCITE. In detail, given an edge-weighted digraph in which any weight is the number of times that a given parental relation is returned during the MCMC, we identify a unique consensus tree via optimum branching. The outcome COB tree will include all nodes connected with a set of edges that maximises the weight sum. To do this, in our case we employed the efficient implementation of Tarjan [11] of the optimum branching tree method originally proposed by Chu and Liu, and Edmonds [3, 5]. This method is analog to the minimum spanning tree problem, but considering a directed graph. Note that our algorithm is: (*i*) deterministic; (*ii*) computationally efficient, to handle the vast number of trees sampled during an MCMC; (*iii*) independent of the order of the input tree list.

The algorithmic steps are detailed in the following:

- The COB-tree algorithm takes as input a binary data matrix D, with n rows representing samples (i.e., single cells or biological samples) and m columns representing genomic mutations. Each entry of D is equal to 1 if the mutation is present in a given sample, 0 if it is absent, NA if the information is missing (see paper P#5).

- In the first step, a generic algorithm for the reconstruction of clonal/mutational trees (e.g., LACE, VERSO or SCITE) is applied to input data D, recording all the solutions sampled during the MCMC. In the output tree T nodes represents mutations (clones) and edges represent parental relations, as in paper P#5.

- For each tree T^p sampled during the MCMC, we generate the corresponding adjacency matrix M^p with dimension $m \times m$, where $p \in [1, \dots r]$ and r is the number of MCMC iterations.

- We compute a weighted adjacency matrix W with dimensions $m \times m$, where each entry is defined as: $W_{i,j} = \frac{\sum_{p=1}^{r} M^p_{i,j}}{r}$. So, W_{ij} stores a weight that corresponds to the frequency by which the mutation i is found as parent of mutation j in all sampled trees. W represents a edge-weighted digraph.

- Finally, we apply the Tarjan algorithm to W in order to find the COB tree.

Synthetic data generation. In order to test the COB-tree algorithm, we generate a number of simulated datasets with following procedure. We first randomly generated a number of ground-truth topologies T_{gt}. In particular, given a specific number of nodes m (i.e., mutations or clones), 20 topologies are created. Starting from the root, we attached a random number (between 2 to 5) of children nodes. We then selected one of the children nodes and repeated the process until all the nodes were attached. Trees including $50, 100$ and 200 mutations were considered, thus, a total of 60 topologies are finally generated. An example of a tree with 50 mutations is reported in figure 3.4.

We sampled 1000 single cell to generate the ground-truth single-cell genotypes matrix D_{gt} (cells x mutations) from each topology T_{gt}. In particular, we populated each row of D_{gt} (i.e., cell genotype) by randomly selecting a mutation/clone (i.e., node). Then, the genotype of a cell (i.e., row of D_{gt}) is populated by assigning 1, if the mutation is included in the shortest path from the selected node to the root, or 0, otherwise. Notice that this path is unique because each node could have only one parent. Since it is unrealistic to observe a high number of clones in a cancer sample, we increased the probability of selecting any of the leaf nodes (i.e., most recent clones) with respect to selecting one of the internal nodes. The former probability is five times higher than the latter. As a result, D_{gt} is a binary matrix with 1000 rows and m columns (notice that each clone can be defined by the last mutation accumulated, so the number of clones equals the number of mutations).

In order to include noise in the simulated datasets, we defined *low*, *middle*, and *high* noise levels by setting the value of False Positive rate (α), False Negative rate (β), and Missing value rate (γ) rate as follows:

- Low noise level: $\alpha = 0.005$, $\beta = 0.05$, and $\gamma = 0$
- Middle noise level: $\alpha = 0.01$, $\beta = 0.1$, and $\gamma = 0.1$
- High noise level: $\alpha = 0.02$, $\beta = 0.2$, and $\gamma = 0.2$

From each D_{gt}, 3 different noise datasets D are generated, by randomly selecting α of the entries equal to 1, β of 0 entries and changing them into 0 and 1, respectively. Then a fraction γ of all the entries are replaced with missing values.

Simulation settings. The procedure described above yields a total of 180 noisy datasets. In this preliminary analyses, we employed SCITE [113] as inference framework, since it is one of the state-of-the-art approaches for single-cell phylogenetic inference.

In particular, each inference is performed multiple times for each dataset, with distinct values of MCMC iterations. In detail, we performed 10 independent SCITE runs (with 10 restarts), with the following MCMC iterations:

- for models with $m = 50 \rightarrow [1000, 2000$ *(short)*$, \ldots, 6000$ *(average)*$, \ldots, 10000$ *(long)*$]$ MCMC iterations,
- for models with $m = 100 \rightarrow [5000, 10000$ *(short)*$, \ldots, 30000$ *(average)*$, \ldots, 50000$ *(long)*$]$ MCMC iterations,
- for models with $m = 200 \rightarrow [50000, 100000$ *(short)*$, \ldots, 300000$ *(average)*$, \ldots, 500000$ *(long)*$]$ MCMC iterations.

COB tree reconstruction. Next, we proceeded with the consensus optimum branching tree reconstruction. To this end, we considered all the trees sampled during the MCMC. Since SCITE discards the first 25% trees, we only considered the remaining 75% and kept only the trees with a likelihood between L_{best} and $L_{best} \times 1.3$, so to focus on the final part of the MCMC. The Tarjan algorithm was finally applied to retrieve the unique COB tree. Notice that we also kept track of the ML tree (L_{best}).

Metrics We tested our approach by comparing each COB tree and the corresponding ML tree in terms of differences with the ground-truth topologies. To this aim, we computed two different metrics (i.e., Parent-Child distance PC and Clonal Genotype errors GC), which assess either the local or the global structure in terms of errors between the obtained COB and ML trees, and the ground-truth topologies.

- Parent-Child distance (PC). This metric is widely used to compare different trees, for example in [153, 169]. In brief, the parent-child distance between two trees enumerates the edges unique in either trees. Small values of this metric reflect a correct recovery of the relations between two consecutive nodes, but disregard their position in the topology, so it is considered a local measure. We compute the PC_{ML}, and PC_{COB} for evaluate the goodness of maximum likelihood and optimal branching tree respectively.
- Clonal Genotype errors (CG). As explained above, each clone can be associated with the node representing the last accumulated mutation. So, its genotype includes all the mutations in the path from such node to the root. Thus, clonal genotypes depend on the overall topology. For each inference, we transformed the ground-truth topology, the ML tree, and the COB tree into the corresponding clonal genotype matrices. Then, we computed the Hamming distance, i.e., the total number of errors, between the clonal genotype of the ground-truth and either the ML tree (CG_{ML}) or the COB trees (CG_{COB}).

Finally, we define $\Delta PC = PC_{ML} - PC_{COB}$ distance and $\Delta CG = CG_{ML} - CG_{COB}$. Positive values indicate an improvement of our approach with respect to the ML tree.

Results I): characterization of the search space. In order to plot the distribution of the sampled trees during the MCMC, we applied the Principal Coordinate Analysis [4, 6] on the distance matrix computed considering the parent-child distance. This approach returns a 2-dimensional representation of the tree space, where the relative distance among each point (i.e., sampled trees) is maintained. We added the value of the likelihood L of each tree as a third dimension, by computing the $-\ln L$. Notice that, after the transformation, the best likelihood values are the lowest. In figure 3.2 we reported the results of the Principal Coordinate Analysis applied on the trees sampled from the inference of one dataset, generated from a tree topology including $m = 100$ mutations. We show the solution considering three different MCMC lengths i.e., 10000 (short), 30000 (average), and 50000 (long) steps.

As one can see, the 10 independent chains tend to reach the same global minimum but, when the MCMC is short, the trees with better likelihood are far from each other. We also marked the COB trees with red dots, and they appear to be placed in an average position among the trees sampled late in the inference. Further tests in different scenarios, as well as the application to real datasets, are currently ongoing.

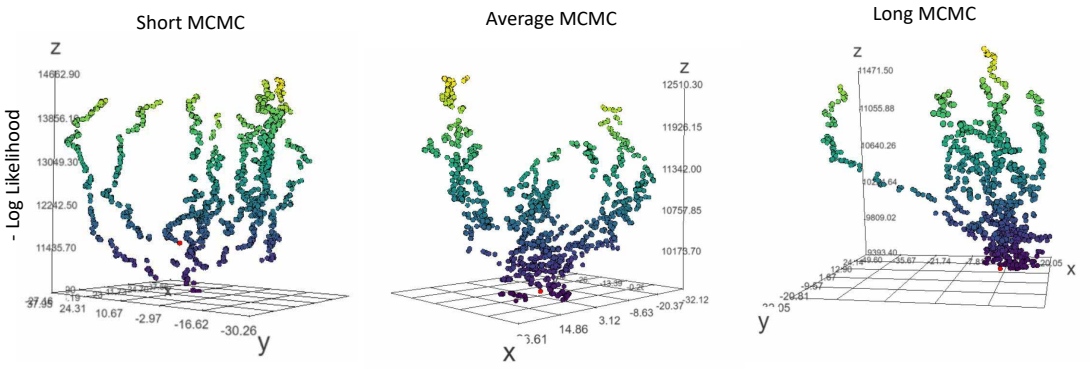


Figure 3.2: A visual representation of the explored solution space during the inference of a clonal tree from the same synthetic single-cell dataset (with 100 mutations and 1000 cells), in three independent MCMC runs with 10 restarts (via SCITE). From left to right, the total number of MCMC increases (10000, 30000, and 50000 steps). The solution space is defined by computing a Principal Coordinate Analysis on the distance matrix (using parent-child distance) of the trees sampled during each inference. Z-axis reports the corresponding likelihood value. Red dots indicate the position of the COB trees.

Results II: performance assessment of COB-tree. We considered three MCMC lengths to evaluate how this affects the performance the COB-tree method. Results are reported in figure 3.3. It is possible to observe how our approach improves the local

structure, by recovering a better ordering of the accumulation of mutations. Instead, the improvement is not that evident when considering the CG metric underlying the global structure. Even though the COB tree often improves this metric as well, it sometimes returns trees with a global structure very far from the ground-truth. Note

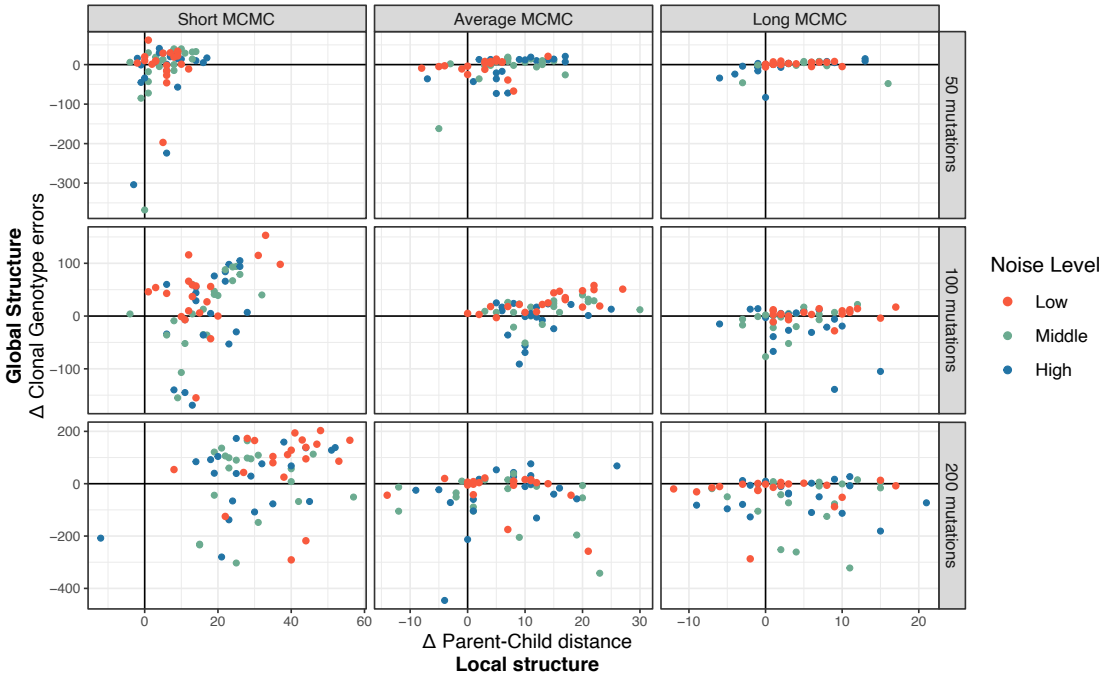


Figure 3.3: Differences between Maximum Likelihood and Consensus Optimum Branching trees are reported. Positive values of Δ Clonal Genotypes errors or Δ Parent-Child distance indicate an improvement for the global structure or the local structure of our COB trees over the ML ones. Colours indicate the level of noise (i.e., rate of FP events, FN events, and missing values) included in the simulated datasets.

that it is possible to have trees with high PC distance values and low CG metric values. A possible explanation is illustrated in the example depicted in figure 3.4. In the plot, it is possible to observe how the COB tree retrieves a better ordering of mutation pairs. Still, the few errors could drastically change the global topology of the tree by shifting an entire subtree.

Conclusion. To conclude, the preliminary analyses illustrated here show that in case one is interested in defining the ordering of mutational events, instead of characterizing clonal genotypes, the COB-tree algorithm may be a reliable solution. Instead, different approaches to compute the consensus tree should be conceived if one is more interested in the global structures.

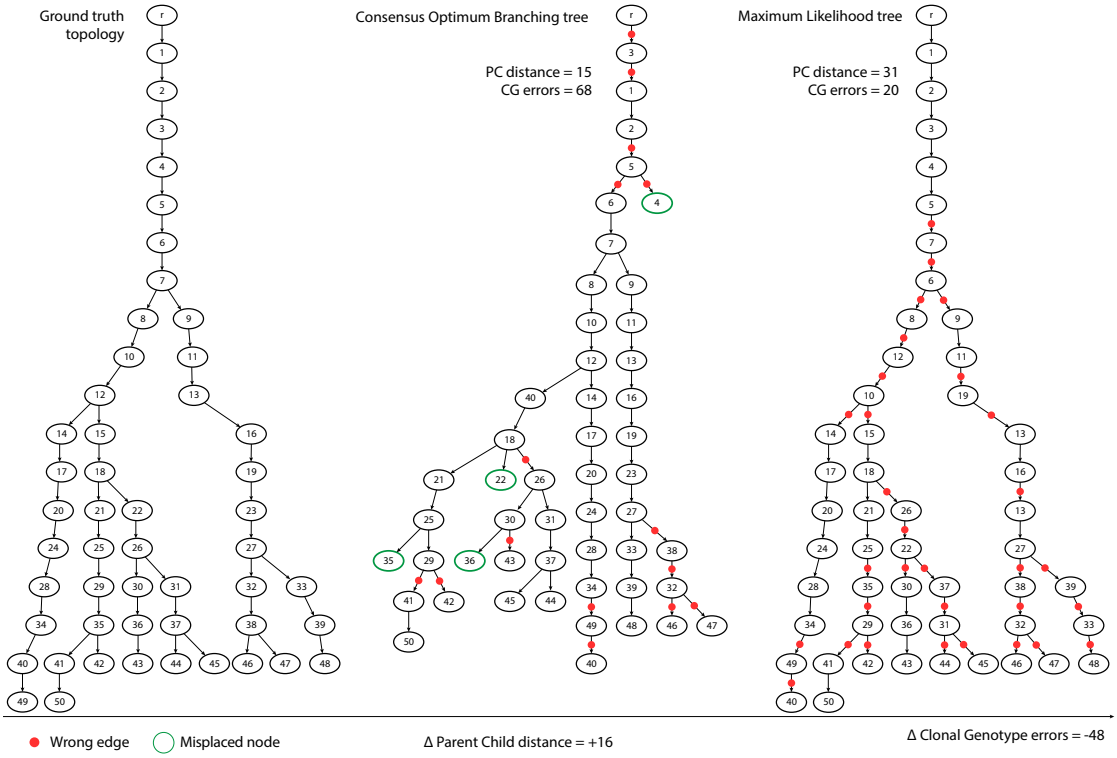


Figure 3.4: The 3 clonal tree comparison highlights the difference between local and global tree structures. The order of mutational events are improved in COB tree (wrong edges marked with red dots, $\Delta PC = +16$), while the global structure is worst ($\Delta CG = -48$) due to a error propagation of few nodes being misplaced (most relevant are highlighted with green circle). The numbers in the nodes indicate distinct mutations.

3.2.2 Decomposition of mutational profiles of viral samples into mutational signatures

Further efforts to exploit the mutational profiles of viral samples were devoted to the characterization of the intra-host mutational landscape and to that of the related mutational processes.

In humans, different mechanisms damage the genome, inducing point mutations. Nucleotides (i.e., A, T, C, or G) are often substituted with a specific frequency that is characteristic of the process. This phenomenon is reflected in the so-called *mutational signature*, which in the case of cancer have been quite thoroughly characterized [204]. For example, tobacco smoking mainly induces C to A transversions.

In our case, for each viral sample we can obtain the total count of each substitution type directly by analyzing deep sequencing data, as already shown in paper P#6. Statistical approaches can be applied to deconvolve the signal into distinct mutational signatures with host-specific intensities. This is typically done by applying a Non-negative Matrix Factorization approach to the mutational spectra of the samples of a given dataset, as proposed in a different context in [240].

Notice that, in our case, it is sound to execute the analysis on intra-host minor mutations only (low variant frequency – VF), because such mutations are likely not transmitted during infections, but may have emerged in the host. In such a way, we can associate to each signature a mechanism causing mutations in the viral genome, and which might be related to the interaction with the host.

Thanks to this approach in paper P#7, we unveiled three non-overlapping mutational signatures related to specific nucleotide substitutions, likely induced by Reactive Oxygen Species damage, but also by the action of two human enzymes, i.e., APOBEC and ADAR), which are known to be active against a broad range of viral infections [253].

Moreover, we showed that is possible to cluster samples based on their signature activities, with standard unsupervised machine learning methods. This exciting result highlights the heterogeneous host responses to SARS-CoV-2 infections, a phenomenon which should be investigated further. In fact, such groups should be correlated with clinical covariates (e.g., age, gender, disease progression and severity, etc.) to better understand the infection mechanisms and the immune system reactions. Unfortunately, to our knowledge, no publicly available datasets include such information.

We also evaluated the different selection pressure by designing a new corrected-for-signatures dN/dS analysis, which considers the non-uniform distribution of nucleotide substitution exhibited by the distinct signature-based clusters of samples. All in all, these analyses prove the relevance of distinguishing between transmitted mutations and intra-host ones in the investigation of viral evolution and heterogeneity.

The paper discussing such analyses is presented below. Note that the formal definition of several algorithmic procedures, such as (*i*) signature decomposition, (*ii*) identi-

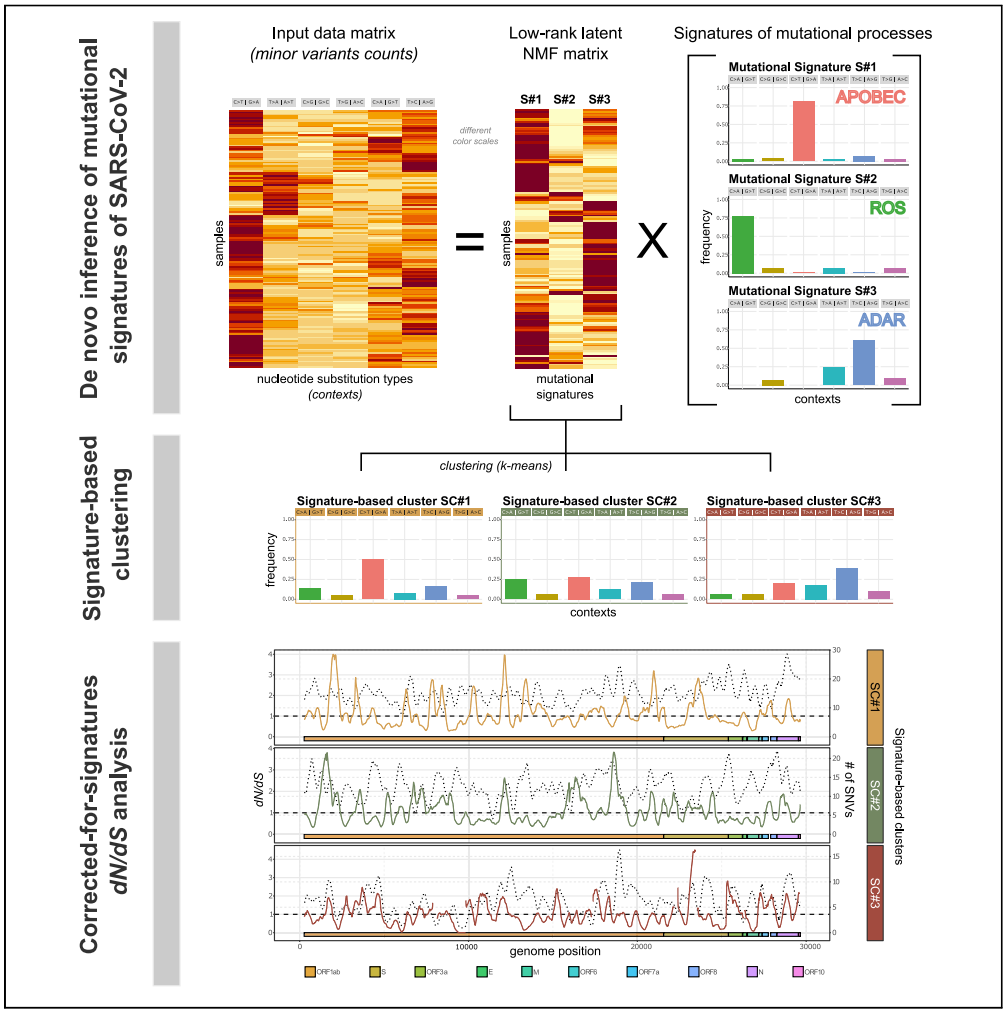
fication of signature-based clusters, (*iii*) assessment of signatures significance, (*iv*) the formal definition of the corrected-for-signatures dN/dS analysis, can be found on the Supplementary Information of the article (online), as well in the related protocol included in the Appendix P#A1.

iScience

CellPress
OPEN ACCESS

Article

Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity



Alex Graudenzi, Davide Maspero, Fabrizio Angaroni, Rocco Piazza, Daniele Ramazzotti

alex.graudenzi@ibfm.cnr.it (A.G.)
rocco.piazza@unimib.it (R.P.)
daniele.ramazzotti@unimib.it (D.R.)

HIGHLIGHTS

The intra-host genomic diversity of SARS-CoV-2 samples reveals host-related processes

Three non-overlapping mutational signatures are inferred from minor variant profiles

Most mutations are purified, yet many variants exhibit wide frequency spectra

The study of homoplasies shows that minor variants are transmitted across hosts

Graudenzi et al., iScience 24, 102116
February 19, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.isci.2021.102116

iScience

CellPress
OPEN ACCESS

Article

Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity

Alex Graudenzi,[1,2,5,7,*] Davide Maspero,[1,3,5] Fabrizio Angaroni,[3,5] Rocco Piazza,[4,6,*] and Daniele Ramazzotti[4,6,*]

SUMMARY

To dissect the mechanisms underlying the inflation of variants in the Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) genome, we present a large-scale analysis of intra-host genomic diversity, which reveals that most samples exhibit heterogeneous genomic architectures, due to the interplay between host-related mutational processes and transmission dynamics. The decomposition of minor variants profiles unveils three non-overlapping mutational signatures related to nucleotide substitutions and likely ruled by APOlipoprotein B Editing Complex (APOBEC), Reactive Oxygen Species (ROS), and Adenosine Deaminase Acting on RNA (ADAR), highlighting heterogeneous host responses to SARS-CoV-2 infections. A corrected-for-signatures *dN/dS* analysis demonstrates that such mutational processes are affected by purifying selection, with important exceptions. In fact, several mutations appear to transit toward clonality, defining new clonal genotypes that increase the overall genomic diversity. Furthermore, the phylogenomic analysis shows the presence of homoplasies and supports the hypothesis of transmission of minor variants. This study paves the way for the integrated analysis of intra-host genomic diversity and clinical outcomes of SARS-CoV-2 infections.

INTRODUCTION

The COronaVIrus Disease 2019 (COVID-19) pandemic has currently affected 216 countries and territories worldwide with ≈70 million people being infected, while the number of casualties has reached the impressive number of ≈1.6 million (World Health Organization (WHO) (2020), update 15th December 2020). The origin and the main features of Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) evolution have been investigated (Zhou et al., 2020b; Wu et al., 2020; Andersen et al., 2020; Xiao et al., 2020; Deng et al., 2020), also due to the impressive amount of consensus viral sequences included in public databases, such as Global Initiative on Sharing Avian Influenza Data (GISAID) (Shu and McCauley, 2017). However, only a few currently available data sets include raw sequencing data, which are necessary to quantify intra-host genomic variability.

Due to the combination of high error and replication rates of viral polymerase, subpopulations of viruses with distinct genotypes, also known as viral quasispecies (Domingo et al., 1985), usually coexist within single hosts. Such heterogeneous mixtures are supposed to underlie most of the adaptive potential of RNA viruses to internal and external selection phenomena, which are related, e.g., to the interaction with the host's immune system or to the response to antiviral agents. For instance, it was hypothesized that intra-host heterogeneity may be correlated with prognosis and clinical outcome (Novella et al., 1995; Domingo et al., 2012). Furthermore, even if the modes of transmission of intra-host variants in the population are still elusive, one may hypothesize that, in certain circumstances, infections allow such variants to spread, sometimes inducing significant changes in their frequency (Lythgoe et al., 2020).

In particular, several studies on SARS-CoV-2 support the presence of intra-host genomic diversity in clinical samples and primary isolates (Ramazzotti et al., 2021; Shen et al., 2020; Wölfel et al., 2020; Capobianchi et al., 2020; Rose et al., 2020; Lu et al., 2020; Lythgoe et al., 2020; Seemann et al., 2020; Popa et al., 2020), whereas similar results were obtained on Severe Acute Respiratory Syndrome CoronaVirus

[1]Inst. of Molecular Bioimaging and Physiology, Consiglio Nazionale delle Ricerche (IBFM-CNR), Segrate, Milan, Italy

[2]Bicocca Bioinformatics, Biostatistics and Bioimaging Centre – B4, Milan, Italy

[3]Department of Informatics, Systems and Communication, Univ. of Milan-Bicocca, Milan, Italy

[4]Department of Medicine and Surgery, Univ. of Milan-Bicocca, Monza, Italy

[5]These authors contributed equally

[6]Co-senior authors

[7]Lead contact

*Correspondence: alex.graudenzi@ibfm.cnr.it (A.G.), rocco.piazza@unimib.it (R.P.), daniele.ramazzotti@unimib.it (D.R.)

https://doi.org/10.1016/j.isci.2021.102116

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

(SARS-CoV) (Xu et al., 2004), Middle East Respiratory Syndrome (MERS) (Park et al., 2016), Ebola virus (Ni et al., 2016), and Hemagglutinin Type 1 and Neuraminidase Type 1 (H1N1) influenza (Poon et al., 2016). We here present one of the the largest up-to-date studies on intra-host genomic diversity of SARS-CoV-2, based on a large data set including 1133 high-quality samples for which raw sequencing data are available (NCBI BioProject: PRJNA645906). The results were validated on 4 independent data sets including a total of 953 samples (NCBI BioProject:PRJNA625551, PRJNA633948, PRJNA636748, and PRJNA647529; see the Validation section).

Our analysis shows that ≈15% of the SARS-CoV-2 genome has already mutated in at least one sample, including ≈1% of positions exhibiting multiple mutations. The large majority of samples shows a heterogeneous intra-host genomic composition, with 892 out of 1133 samples (≈79%) exhibiting at least one low frequency variant (Variant Frequency, VF >5% and ≤90%, named minor variants or iSNVs), 171 samples more than 5, and 101 samples more than 10. Importantly, several variants are observed as clonal (VF > 90%) in certain samples and at a low frequency in others, demonstrating that transition to clonality might be due not only to functional selection shifts but also to complex transmission dynamics involving bottlenecks and founder effects (Domingo et al., 2012).

Strikingly, our analysis allowed us to identify three non-overlapping "mutational signatures", i.e., specific distributions of nucleotide substitutions, in which are observed in distinct mixtures and with significantly different intensity in three well-separated clusters of samples, suggesting the presence of host-related mutational processes. One might hypothesize that such processes are related to the interaction of the virus with the host's immune system and might pave the way for a better understanding of the molecular mechanisms underlying different clinical outcomes.

In particular, the first signature is dominated by C>T:G>A substitution and it is likely related to APOlipoprotein B Editing Complex (APOBEC) activity, the second signature is mostly characterized by G> T:C> A substitution and it might be associated to Reactive Oxygen Species (ROS)-related processes, while a third signature is predominantly associated to A>G:T>C substitution, which is usually imputed to Adenosine Deaminase Acting on RNA (ADAR) activity.

A corrected-for-signatures version of the *dN*/*dS* analysis would suggest that, as expected, the three signatures are affected by mild purifying selection in the population, yet with some exceptions that would suggest the existence of positively selected genomic regions. Furthermore, a certain proportion of samples of two signature-based clusters mostly associated to APOBEC and ROS appear to be hypermutated (up to 87 minor variants detected in a single host), whereas this effect is mitigated for the remaining cluster, dominated by ADAR-related processes.

Finally, the analysis of the phylogenetic model, obtained from the profiles of clonal variants via the Viral Evolution ReconStructiOn (VERSO) framework (Ramazzotti et al., 2021), allowed us to assess how many minor variants are either detected in single samples, in multiple samples of the same clade, or in multiple samples of independent clades (i.e., homoplasies). Strikingly, an approximately monotonic decrease of the median VF is observed with respect to the number of clades in which minor variants are observed: minor variants detected in single clades exhibit the largest (median) VF, as opposed to variants shared in multiple clades, which display a progressively lower VF.

On the one hand, this result supports the hypothesis of transmission of minor variants during infections and of the concurrent existence of bottleneck effects (Gutierrez et al., 2012; Domingo et al., 2012). On the other hand, the significant number of minor variants observed at a low frequency in multiple clades would suggest the presence of mutational hotspots and of phantom mutations related to sequencing artifacts (Bandelt et al., 2002).

RESULTS

Mutational landscape of SARS-CoV-2 from variant frequency profiles of 1133 samples – data set #1

We performed variant calling from Amplicon raw sequencing data of 1188 samples from the NCBI BioProject: PRJNA645906 and by aligning sequences to reference genome SARS-CoV-2-ANC, which is a

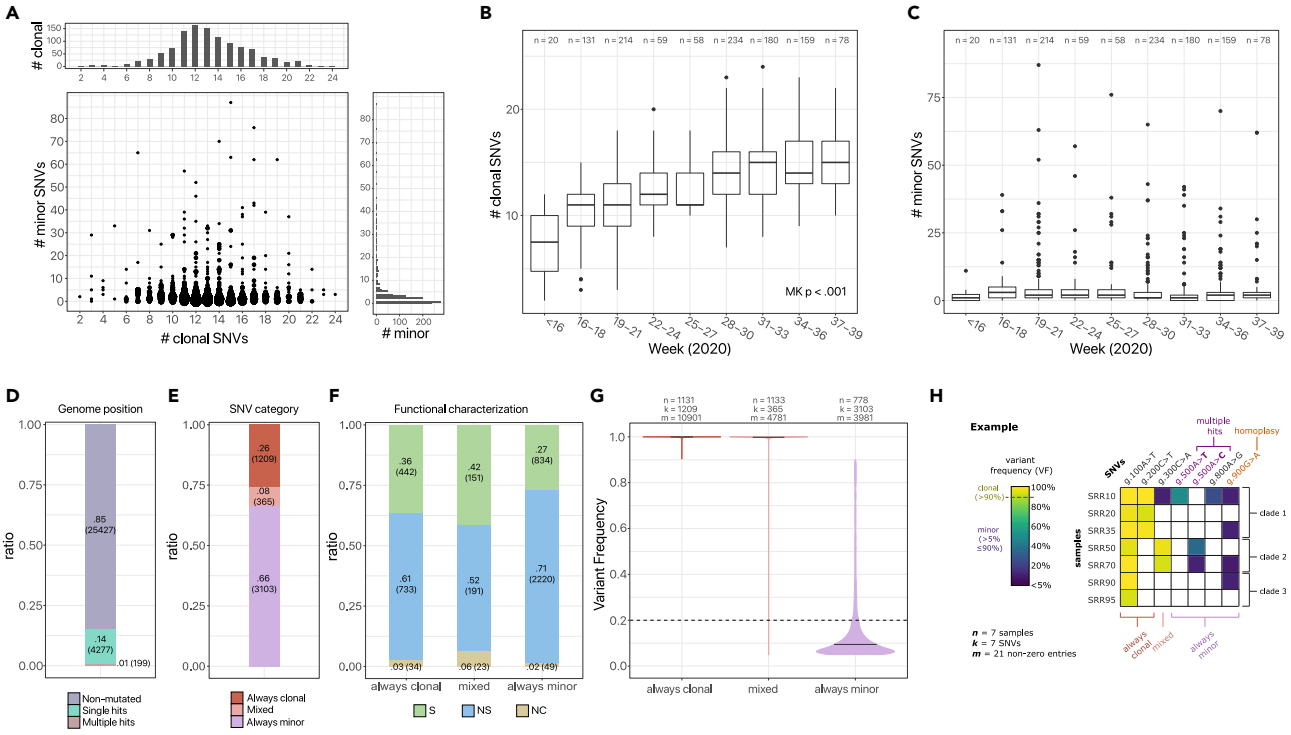


Figure 1. Mutational landscape of 1133 SARS-CoV-2 samples – data set #1 (NCBI BioProject: PRJNA645906)

(A) Scatter plot displaying the number of clonal (VF >90%) and minor (VF >5% and ≤90%) variants for 1133 samples of data set #1 (node size proportional to the number of samples).

(B) Box plots returning the distribution of the number of clonal and (C) minor variants, obtained by grouping samples according to collection date (weeks, 2020; Mann-Kendall trend test p value also shown). *n* returns the number of samples in each group.

(D) Bar plot returning the proportion of sites of the SARS-CoV-2 genome that are either non-mutated, mutated with a unique SNV, or mutated with multiple SNVs.

(E) Stacked bar plots returning the proportion of SNVs detected as always clonal, mixed, or always minor.

(F) The ratio of synonymous (S), non-synonymous (NS), and non-coding (NC) mutations, for each category.

(G) Violin plots returning the distribution of VF of all SNVs (*n* returns the number of samples, *k* the number of distinct SNVs, *m* the number of non-zero entries of the VF matrix).

(H) Graphical representation of an example data set.

likely ancestral SARS-CoV-2 genome (Ramazzotti et al., 2021). The mutational profiles of 1133 high-quality samples selected after quality check were analyzed in depth (see Methods and Figure S1).

In detail, 4677 distinct single-nucleotide variants (SNVs, identified by genome location and nucleotide substitution) were detected in the data set, for a total of 19663 non-zero entries of the VF matrix (see Methods for further details; the VF profiles of all samples are included in Data S1; see Figure 1H for a graphical representation of an example data set). In particular, in our analysis, we consider any SNV detected in any given sample as "clonal", if its VF is >90% and as "minor" if its VF is >5% and ≤90%.

The distribution of the number of minor and clonal variants observed in each sample (Figure 1A) unveils an approximately normal distribution of clonal variants (median = 13, mean = 13.2, and max = 24). Minor variants are detected in ≈78.7% of the samples and show a long-tail distribution (median = 2, mean = 4.16 and max = 87). One hundred nine samples (≈9.6% of the data set) display a number of minor variants ≥ 10, up to a maximum of 87.

Interestingly, we observe a statistically significant increase of genomic diversity on clonal variants with respect to collection week (Mann-Kendall test for trend on median number of clonal variants $p < 0.001$, Figure 1B), due to the accumulation of clonal variants in the population, and which confirms recent findings (Li

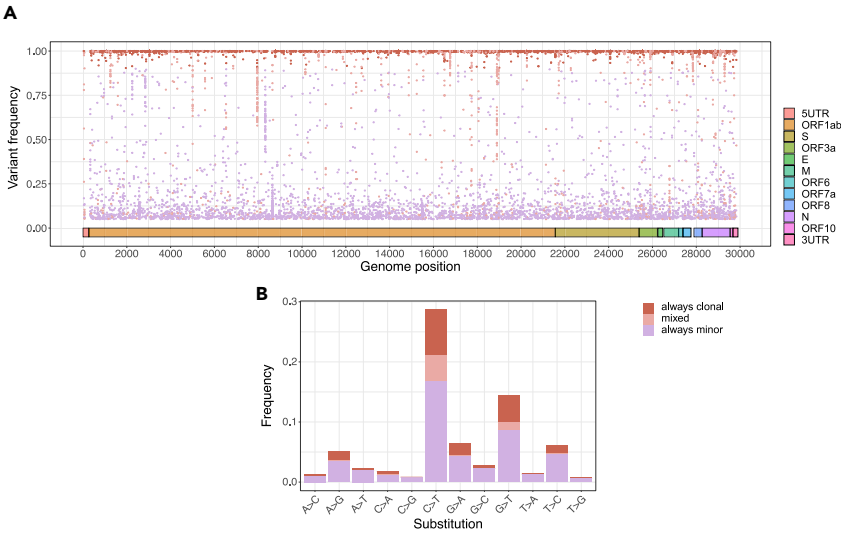


Figure 2. Characterization of SNVs detected on the SARS-CoV-2 genome

(A) Scatter plot returning the genome location and the VF of all SNVs detected in the data set, colored according to category.

(B) Stacked bar plots returning the normalized substitution proportion of all SNVs detected in at least one sample of the data set, with respect to all 12 possible nucleotide substitutions, grouped by variant type.

et al., 2020; Shen et al., 2020; Ramazzotti et al., 2021), whereas, as expected, this phenomenon is less evident for minor variants (Figure 1C). This aspect is further investigated in the following and hints at the interplay involving the evolutionary dynamics within hosts and the transmission among hosts, which differently affects clonal and minor variants (Chan et al., 2013).

Evidence of transition to clonality

We further categorize each detected SNV as follows: (*i*) "always clonal", if clonal in all samples in which it is detected, (*ii*) "always minor", if minor in all samples in which it is detected, (*iii*) "mixed", if observed as clonal in at least one sample and as minor in at least another sample.

Forty thousand six hundred seventy seven SNVs were detected on 4476 distinct genome sites (≈ 14.9% of the SARS-CoV-2 genome), of which 199 sites (≈0.6% of the genome) display multiple nucleotide substitutions (see Figure 1D). This suggests that the proportion of mutated genomic sites might be considerably higher in the overall population, especially if considering minor variants. Overall, 25.8%, 7.8%, and 66.3% SNVs are detected as always clonal, mixed, and always minor, respectively, and are mostly non-synonymous (see Figures 1E and 1F).

The analysis of the VF distribution (Figure 1G) unveils an impressive scarcity of variants showing VF in the middle range, i.e., between 20% and 90%, for all categories. This phenomenon is likely due to transmission bottlenecks, which tend to purify low-frequency variants in the population. Nonetheless, both mixed and always minor variants display broad VF spectra, an aspect that is particularly relevant for the former category. In this respect, 24.4% of all mixed variants (89 on 365) never display a VF ≤20%: one may hypothesize that such variants are indeed "transiting to clonality" in the population because either positively selected, as a result of the strong immunologic pressure within human hosts (Lucas et al., 2001), or because affected by transmission phenomena involving founder effects, bottlenecks, and stochastic fluctuations (Gutierrez et al., 2012; Domingo et al., 2012).

Conversely, one might hypothesize that most remaining mixed variants may result from random mutations hitting positions of SNVs that are already present as clonal in the population.

Furthermore, the distribution of SNVs with respect to each region of the genome in Figure 2A demonstrates that mutations are approximately uniformly distributed across the genome (also see Figure S2).

Overall, this analysis provides one of the first large-scale quantifications of transition to clonality in SARS-CoV-2 and might serve to intercept variants possibly involved in functional modifications, bottlenecks, or founder effects.

De novo inference of SARS-CoV-2 mutational signatures

In order to investigate the existence of mutational processes related to the interaction between the host and the SARS-CoV-2, we analyzed the distribution of nucleotide substitutions for all SNVs detected in the data set. In Figure 2B, one can see the proportion of SNVs for each of the 12 nucleotide substitution types (e.g., number of C>T's) over the total number of nucleotides present in the reference genome for each substitution type (e.g., number of C's).

Certain substitutions present a significantly higher normalized abundance, confirming recent findings on distinct cohorts (Simmonds, 2020; Di Giorgio et al., 2020; Popa et al., 2020). In particular, C> T substitutions are observed in ≈28% of all C nucleotides in the SARS-CoV-2 genome, G>T's in ≈15% of all G's, T> C's in ≈7% of all T's, and A>G's in ≈5% of all A's.

Although, traditionally, a 12-substitution pattern has been used in order to report mutations occurring in single-stranded genomes, we reasoned that, owing to the intrinsically double-stranded nature of the viral life cycle (i.e., a mutation occurring on a plus strand can be transferred on the minus strand by RdRP and vice versa), it is sound to consider a total of 6 substitution classes (obtained by merging equivalent substitutions in complementary strands) to investigate the possible presence of viral mutational signatures (Alexandrov et al., 2013). Clonal variants were not considered in the next analyses to focus on SNVs likely related to host-specific mutational processes and by excluding variants presumably transmitted during infection events.

In particular, in order to identify and characterize the mutational processes underlying the emergence of SARS-CoV-2 variants with a statistically grounded approach, we applied a Non-negative Matrix Factorization (NMF) approach (Brunet et al., 2004) and standard metrics to determine the optimal rank (see Methods). In particular, we analyzed the mutational profiles of 150 samples exhibiting at least 6 always minor variants (on 1133 total samples) to ensure a sufficient sampling of the distributions.

Strikingly, 3 distinct and non-overlapping mutational signatures are found and explain 96.5% of the variance in the data (Figures 3A and S4; cophenetic correlation coefficient = 0.998, cosine similarity between predictions and observations = 0.973, harmonic mean p value of the one-sided Mann-Whitney U test on bootstrap re-sampling <0.01 for all signatures, see Methods). In particular, signature *S*#1 is predominantly related to substitution C>T:G>A (81.2%), signature *S*#2 to substitution C>A:G>T (77.7%), while signature *S*#3 is dominated by substitutions T>C:A>G (*S*#3) and T>A:A>T (23.6%).

Characterization of mutational signatures of SARS-CoV-2

Signature *S*#1 is related to C>T:G>A substitution, which was often associated to APOBEC i.e., a cytidine deaminase involved in the inhibition of several viruses and retrotransposons (Sharma et al., 2015). An insurgence of APOBEC-related mutations was observed in other coronaviruses shortly after spillover (Woo et al., 2007), and it was recently hypothesized that APOBEC-like editing processes might have a role in the response of the host to SARS-CoV-2 (Simmonds, 2020).

As specified above, a mutational process occurring on single-stranded RNA with a given pattern, e.g., C> T, could occur as a C>T mutation on the plus reference strand but could similarly occur on the minus strand, again as a C>T substitution. However, C>T events originally occurring in the minus strand would be recorded as G>A owing to the mapping of the mutational event as a reverse complement on the plus reference genome. Starting from these considerations and hypothesizing that the C>T:G>A substitution is mediated by APOBEC, which operates on single-stranded RNA and is similarly active on both strands, the analysis of the C>T/G>A ratio (or, more generally, of a plus/minus substitution ratio) should give an accurate measurement of the molar ratio between the two viral strands inside the infected cells.

In our case, by comparing the proportion of substitutions of all minor variants detected in the data set, the ratio C>T/G>A is 5.1 $\left(\frac{1602/5491}{335/5863}\right)$.

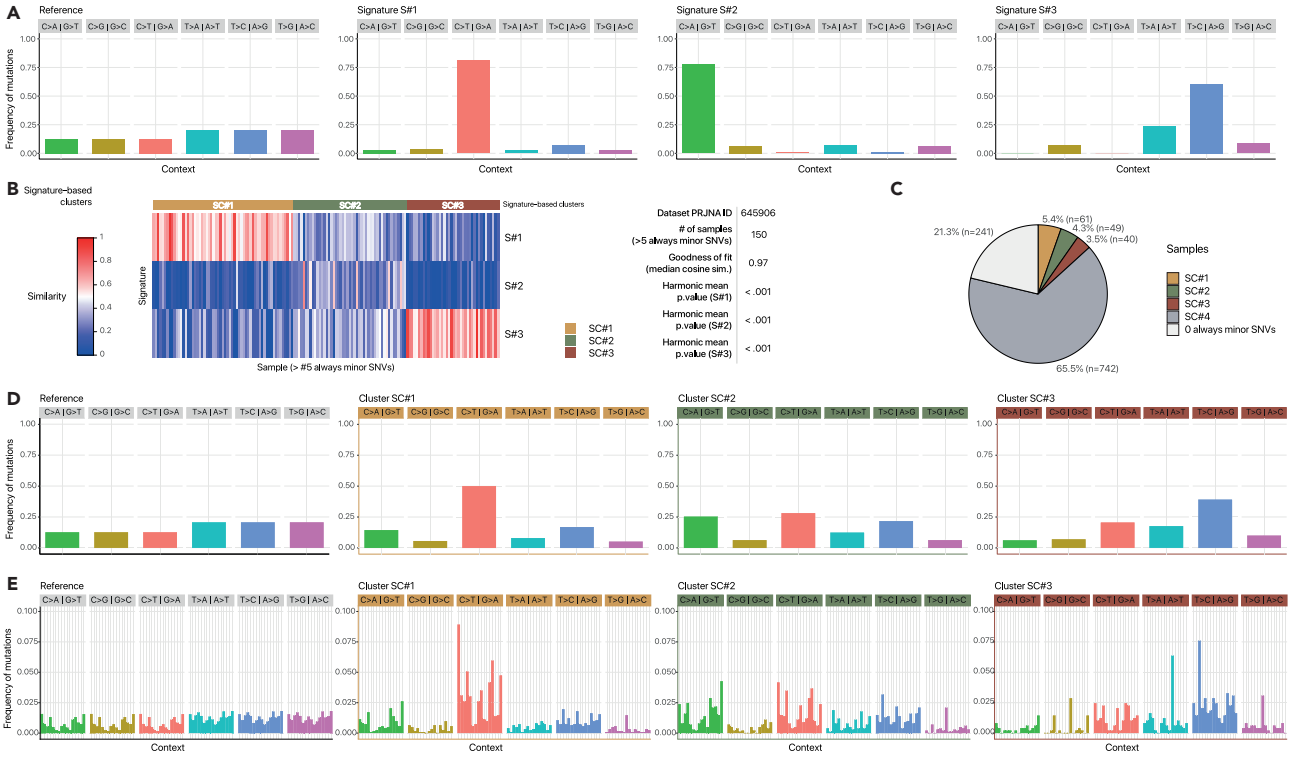


Figure 3. Mutational signatures of SARS-CoV-2

(A) The nucleotide class distribution in SARS-CoV-2-ANC reference genome (Ramazzotti et al., 2021) and for the 3 SARS-CoV-2 mutational signatures retrieved via NMF on 6 substitution classes is shown.

(B) Heatmap returning the clustering of 150 samples with ≥6 always minor variants (≈13% of the data set), computed via k-means on the low-rank latent NMF matrix. The goodness of fit in terms of median cosine similarity between observations and predictions and the harmonic mean p value of the one-sided Mann-Whitney U test on bootstrap re-sampling, are shown for all signatures, see Methods.

(C) Pie chart returning the proportion of samples in the three signature-based clusters, plus a fourth cluster *SC#4* including all samples with ≥ 1 and < 6 always minor variants and the group of samples with 0 always minor SNVs.

(D-E) Categorical normalized cumulative VF distribution of all SNVs detected in each signature-based cluster, with respect to (D) 6 substitution classes and to (E) 96 trinucleotide contexts, as compared to the theoretical distribution in SARS-CoV-2-ANC reference genome (left).

This result allows us to hypothesize that plus and minus viral strands of the SARS-CoV-2 genome are present in infected cells with a molar ratio in strong favor of the plus strand and are consistent with the expected activity of APOBEC on single-stranded RNA. Further experimental analyses will be required to confirm this hypothesis.

The second signature *SC#2* is predominantly characterized by substitution C>A:G>T, whose origin is however still obscure. To gain insight into the mechanisms responsible for its onset, also in this case, we analyzed the C>A and G>T substitution frequency, which revealed a strong disproportion in favor of the latter: the ratio G>T/C>A is 9.5 $\left(\frac{804/5863}{79/5491}\right)$. Overall, this result suggests that, in this case, the G> T substitution is the active mutational process.

In this respect, one might hypothesize a role for ROS as a mutagenic agent underlying this signature, as observed, for instance, in clonal cancer evolution (Alexandrov et al., 2020). ROSs are extremely reactive species formed by the partial reduction of oxygen. A large number of ROS-mediated DNA modifications have already been identified; in particular, however, guanine is extremely vulnerable to ROS because of its low redox potential (David et al., 2007). ROS activity on guanine causes its oxidation to 7,8-dihydro-8-oxo-2′-deoxyguanine (oxoguanine). Notably, (*i*) oxoguanine can pair with adenine, ultimately causing G>T transversions, and (*ii*) ROSs are able to operate on single-stranded RNA; therefore, their mutational process closely resembles the C>A:G> T pattern we see in signature *SC#2*. Thus, it is sound to hypothesize that the C>A:G>T substitution is generated by ROS, whose production is triggered upon infection, in line with several reports indicating that a strong ROS burst is often triggered during the early phases of several viral infections (Molteni et al., 2014; Reshi et al., 2014).

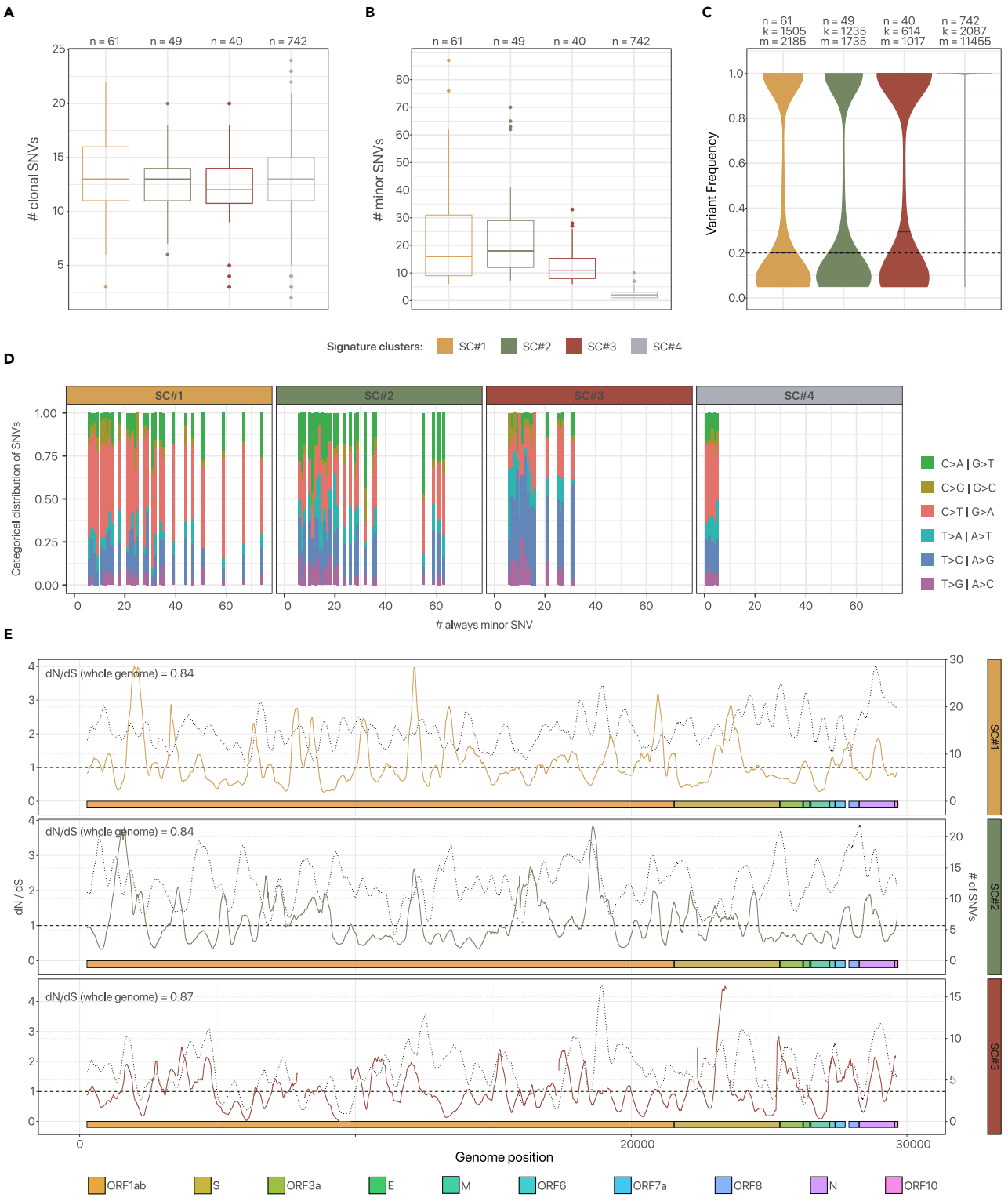


Figure 4. Characterization of signature-based clusters of SARS-CoV-2 samples

(A) Distribution of the number of clonal variants with respect to the 4 signature-based clusters described in the text.

(B) Distribution of the number of minor variants for the 4 signature-based clusters.

Figure 4. *Continued*

(C) Violin plots returning the VF distribution with respect to signature-based clusters (*n* returns the number of samples, *k* the number of distinct SNVs, *m* the number of non-zero entries of the VF matrix).

(D) (Average) proportion of substitution classes of always minor variants for all the samples included in the 4 signature-based clusters, grouped and sorted by the number of minor SNVs (e.g., at position 10 of the x axis one can find the average proportion of substitution classes for all samples with 10 minor SNVs).

(E) Corrected-for-signatures *dN*/*dS* ratio plot, as computed by normalizing the ratio on cluster substitution distribution, on a 300-base sliding window, with respect to signature-based clusters (see Methods). The superimposed dotted line returns the mutational density in each window (rightmost y axis).

Finally, signature *SC*#3 is primarily characterized by A>G:T>C substitution, which is typically imputed to the ADAR deaminase mutational process (Nishikura, 2010). ADAR targets adenosine nucleotides, causing deamination of the adenine to inosine, which is structurally similar to guanine, ultimately leading to an A> G substitution. Unlike APOBEC, ADAR targets double-stranded RNA; hence, it is active only on plus/minus RNA dimers. In line with this mechanism and in sharp contrast with APOBEC, A>G's and the equivalent T>C's show a similar prevalence: the ratio A>G/T>C is 0.81 $\left(\frac{384/8954}{508/9595}\right)$. This supports the notion that the A>G:T>C mutational process is exquisitely selective for double-stranded RNA, where it can similarly target adenines present on both strands.

Identification of signature-based clusters

We then clustered the 150 samples with at least 6 always minor mutations (on 1133 total samples) by applying k-means on the normalized low-rank latent NMF matrix and employing standard heuristics to determine the optimal number of clusters (see Methods). As a result, 3 signature-based clusters (*SC*#1, *SC*#2, and *SC*#3) are retrieved, including 61, 49, and 40 samples, respectively (see Figure 3B).

Remarkably, clusters *SC*#1 and *SC*#3 are characterized by distinctive signatures, *S*#1 (dominated by substitution C>T:G>A) and *SC*#3 (T>C:A>G and T>A:A>T), respectively, whereas cluster *SC*#2 is characterized by a mixtures of all three signatures. In particular, the samples of the distinct clusters display dissimilar categorical VF distributions (see Figure 3D), pointing at the existence of different host-related mutational processes.

We here recall that samples with a number of always minor variants between 1 and 5 (742 samples, 65.5%) cannot be reliably associated to signature-based clusters, due to the low number of SNVs. For this reason, such samples were considered separately in the analysis and were labeled as cluster *SC*#4 from now on (Figure 3C).

Importantly, by computing the categorical VF distribution of all minor SNVs with respect to all 96 trinucleotide contexts (i.e., by considering flanking bases), one can notice that clusters *SC*#1 and *SC*#2 display profiles that resemble that of the theoretical substitution distribution of the reference genome, thus suggesting that, in such cases, the host-related mutational processes are likely independent from flanking bases. Conversely, *SC*#3 displays a distribution of T substitutions with prevalent peaks in certain contexts and, especially, in G[T>A]G, A[T>C]G, and C[T>G]T.

We finally note that, due to the possible transmission of minor variants among hosts during infections (see above), signature-based clusters might include both samples with host-related mutational processes and samples with minor variants inherited from infecting hosts.

Characterization of signature-based clusters

We analyzed in depth the intra-host genomic diversity of the samples of the 4 different signature-based clusters. As a first noteworthy result, while the distributions of the number of clonal variants are significantly alike across clusters (Kolmogorov-Smirnov, KS test $p>0.20$ for 6/6 pairwise comparisons; see Figure 4A and Data S2), clusters *SC*#1 and *SC*#2 display a similar distribution of minor variants (KS test $p = 0.86$) but significantly different distributions from the remaining clusters (KS $p<0.05$ for all remaining pairwise comparisons; see Figure 4B and Data S2). The relative proportion of substitution types for the samples of each signature-based cluster can be found in Figure 4D.

In particular, clusters *SC*#1 and *SC*#2 are characterized by a significantly higher number of minor variants (median 16 and 18, mean 22.3 and 22.6, max 87 and 70, for *SC*#1 and *SC*#2, respectively). Accordingly, both clusters include a certain proportion of highly mutated samples (with $\geq$10 minor variants), 44 on 61

and 41 on 49 for *SC*#1 and *SC*#2, respectively. This result supports the existence of highly active mutational processes and is consistent with the hypothesis of processes related to APOBEC and ROS. Conversely, cluster SC displays a much lower number of minor variants (median 11, mean 13.2, max 33; 23 samples on 40 with $\geq$10 variants). This finding hints at the existence of milder spontaneous mutational processes related to ADAR.

Interestingly, the VF distribution for all SNVs highlights a remarkable similarity among signature clusters *SC*#1, *SC*#2, and *SC*#3, with the large majority of variants found either at a high or a low frequency, whereas, by construction, *SC*#4 is dominated by clonal variants (Figure 4C). Moreover, only minor differences are observed in the distribution of substitutions with respect to SARS-CoV-2 Open Reading Frames (ORFs) (see Figure S3).

Overall, these results reinforce the hypothesis of distinct mutational processes active in different hosts. When clinical data would be available in combination to sequencing data, this will allow us to assess the correlation with clinical outcomes.

Evidence of purifying selection against signature-related mutagenic processes

To investigate the evolutionary dynamics of SARS-CoV-2, we implemented a corrected-for-signatures version of the *dN*/*dS* ratio analysis, i.e., obtained by normalizing the NS/S rate with respect to the theoretical distribution of substitutions detected in each cluster, as suggested in a different context in Van den Eynden and Larsson (2017).

Interestingly, the corrected *dN*/*dS* ratio computed on the genome coding regions (i.e., =29133 basis) is equal to 0.84, 0.84, and 0.87 for the three signature-based clusters, respectively, and suggests the existence of purifying selection for all signature-related mutational processes.

We refined the analysis via a 300-base sliding window approach. On the one hand, the analysis of the mutational density confirms that the large majority of variants is indeed observed in purified regions of the genome. On the other hand, however, the variation of the corrected *dN*/*dS* ratio across the genome shows that some regions exhibit a ratio significantly larger than 1. This phenomenon, which is particularly evident in signature-based clusters *SC*#1 and *SC*#2, hints at possible positive selection processes affecting specific genomic regions and deserves further investigations.

Phylogenomic model of SARS-CoV-2 reveals transmission of minor variants and homoplasies

We employed VERSO (Ramazzotti et al., 2021) to reconstruct a robust phylogeny of samples from the binarized VF profiles of the 28 clonal variants (VF >90%) detected in at least 3% of the data set. In Figure 5A, one can see the output phylogenetic tree, which describes the existence of 23 clades and in which samples with identical corrected clonal genotype are grouped in polytomies (see Figure 5B and the Methods section for further details). The mapping between clonal genotype labels and the lineage dynamic nomenclature proposed in Rambaut et al. (2020) and generated via pangolin 2.0 (O'Toole et al., 2020) is included in Data S3, whereas the phylogenetic model returned via MrBayes (Ronquist et al., 2012) on data set #1 is displayed in Figure S5 (see Methods).

Interestingly, SNV g.29095T>C (mapped on ORF *N*, synonymous) appears to be the earliest evolutionary event from reference genome SARS-CoV-2-ANC (Ramazzotti et al., 2021). All downstream clades belong to type B type (Forster et al., 2020; Tang et al., 2020), as determined by presence of mutations g.8782T>C (*ORF1ab*, synonymous) and g.28144C>T (*ORF1ab*, p.84S>L). Importantly, we note that variant g.23403A>G (*S*, p.614D>G), whose correlation with viral transmissibility was investigated in depth (Lokman et al., 2020; Daniloski et al., 2020; Korber et al., 2020; Zhou et al., 2020a; Grubaugh et al., 2020; Plante et al., 2020), is found in 18 clades, which include 1113 samples of the data set.

In addition, the model unveils the presence of a number of homoplasies, as a few clonal variants are observed in independent clades and, especially, mutation g.11083G>T (*ORF1ab*, p.3606L> F), which was investigated in a number of recent studies on SARS-CoV-2 evolution (van Dorp et al., 2020; Ramazzotti et al., 2021), and is observed in 42 samples and 8 distinct clades. One might hypothesize that such SNVs have spontaneously emerged in unrelated samples and were selected either due to some functional advantage or alternatively to the combination of founder and stochastic effects involved in variant transmission during infections, which might lead certain minor SNVs transiting to clonality in the population (see above).

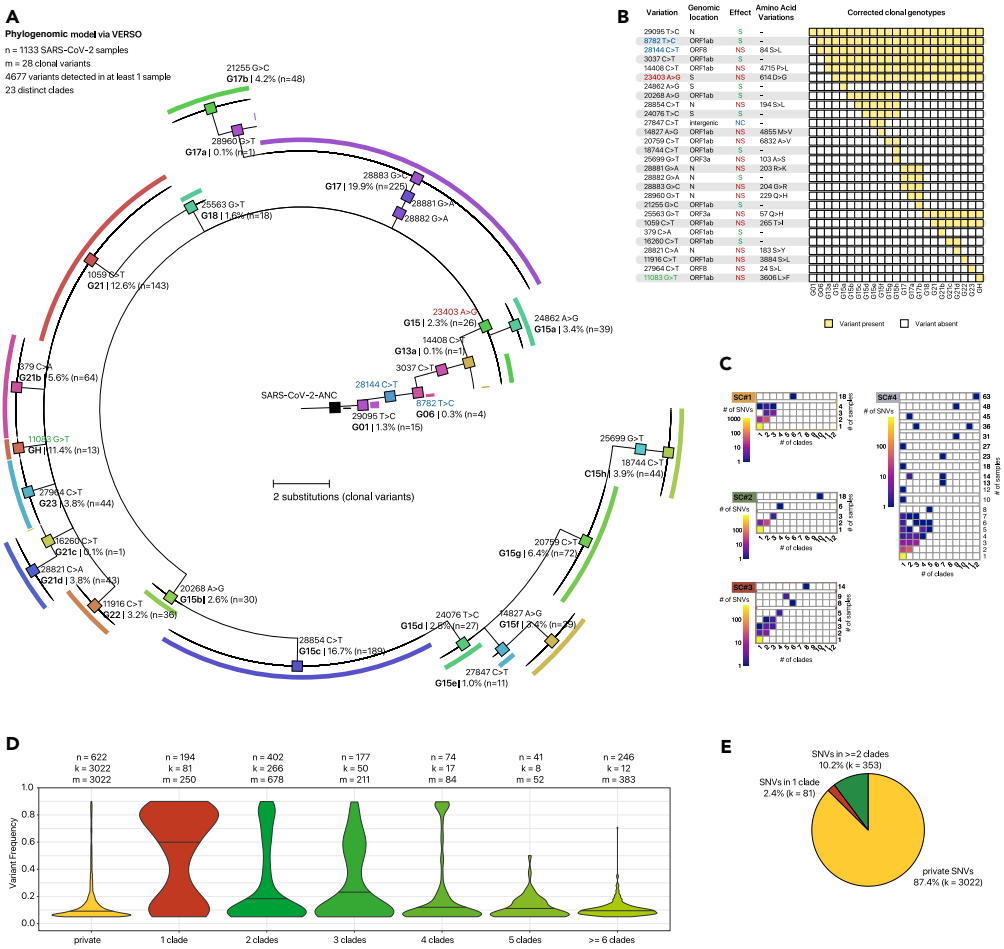


Figure 5. Phylogenomic model of 1133SARS-CoV-2 samples of via VERSO – data set #1 (NCBI BioProject: PRJNA645906)

(A)The phylogenetic tree returned by VERSO (Ramazzotti et al., 2021) considering 28 clonal variants (VF > 0.90) detected in at least 3% of the 1133 samples of the data set is displayed. Colors mark the 23 distinct clades identified by VERSO, which are associated to corrected clonal genotypes. Genotype labels are consistent with (Ramazzotti et al., 2021), whereas in Supplementary File S3, one can find the mapping with the lineage nomenclature proposed in Rambaut et al. (2020). Samples with identical corrected clonal genotypes are grouped in polytomies (visualization via FigTree (Rambaut, 2009)). The black colored sample represents the SARS-CoV-2-ANC reference genome.

(B) Heatmap returning the composition of the 23 corrected clonal genotypes returned by VERSO. Clonal SNVs are annotated with mapping on ORFs, synonymous (S), nonsynonymous (NS) and non-coding (NC) states, and related amino acid substitutions. Variants g.8782T>C (*ORF1ab*, synonymous) and g.28144C>T (*ORF8*, p.84S>L) are colored in blue, variant g.23403 A>G (*S*, p.614 D>G) in red, homoplastic variant g.11083G>T (*ORF1ab*, p.3606L>F) in green.

(C) Heatmaps displaying the count of minor variants with respect to the number of clades and samples in which they are found, grouped by signature-based cluster (e.g., at row 3 and column 5, the color represents the number of SNVs found in 3 clades and 5 samples).

(D) Violin plots returning the VF distribution of all minor variants, with respect to the number of clades in which they are found (the first violin plot is associated to variants privately detected in single samples). *n* returns the number of samples, *k* the number of distinct SNVs, *m* the number of non-zero entries of the VF matrix.

(E) Pie chart returning the proportion of minor variants privately detected in single samples, detected in multiple samples of the same clade, and in multiple samples of independent clades.

As extensively discussed in Ramazzotti et al., 2021, while all the clonal variants of a host are most likely transmitted during an infection, the extent of transmission of minor variants is still baffling and is highly influenced by bottlenecks, founder effects, and stochasticity (Gutierrez et al., 2012; Domingo et al., 2012). Simultaneous infections of the same host from multiple individuals harboring distinct viral lineages (also

named superinfections) might in principle affect variant clonality, yet their occurrence is extremely rare (Lythgoe et al., 2020).

For this reason, we quantified the number of minor variants (VF $\leq$90%)

1. privately detected in single samples and which are most likely spontaneously emerged via host-related mutational processes;
2. found in multiple samples of the same clade, which might be either (*a*) spontaneously emerged or (*b*) transferred from other hosts via infection chains; and
3. observed in multiple samples of independent clades (i.e., homoplasies) and which might be due to (*a*) positive selection of the variants due to some functional advantage, in a scenario of parallel/convergent evolution, (*b*) mutational hotspots, i.e., SVNs falling in mutation-prone sites or regions of the viral genome, (*c*) phantom mutations due to sequencing artifacts (Bandelt et al., 2002), and (*d*) complex transmission dynamics involving founder effects and stochasticity, which may allow certain minor variants to transit to clonality, eventually leading to a clonal genotype transmutation (see above).

In our case, we observe that 87.4% of minor variants are observed as private of single samples, 2.4% in multiple samples of the same clade, and 10.2% are detected in samples belonging to distinct clades (Figure 5E). Importantly, significantly different VF distributions are observed, and, especially, an approximately monotonic decrease of the median VF is detected with respect to the number of clades in which minor variants are found (Figure 5D). Important conclusions can be drawn from these results.

Apparently, the large majority of minor SVNs spontaneously emerges in single samples, likely due to signature-based mutational processes. Yet, the VF distribution of private minor SNVs suggests that, as expected, most of such variants are indeed purified in the population.

Accordingly, the hypothesis of transmission of minor variants during infections is supported by the significantly larger VF of (the fewer) minor variants found in multiple samples of the same clade, as this effect is most likely due to transmission bottleneck effects (Gutierrez et al., 2012; Domingo et al., 2012).

In addition, the progressively smaller VF of minor variants observed in samples of independent clades and which are likely more distant in the infection chain hints at the noteworthy presence of mutational hotspots and of phantom mutations related to sequencing artifacts (Bandelt et al., 2002). In all scenarios, the presence of positively selected variant cannot be excluded but requires ad hoc investigations.

Interestingly, one can refine the analysis by focusing on distinct signature-based clusters, for instance, by pinpointing variants likely related to mutational hotspots or phantom mutations: see, e.g., variant g.8651A>C (*ORF1ab*, p.2796M>L) which is observed in 63 samples and 12 clades (Figure 5C).

Validation – data sets #2 – 5

We employed 4 independent data sets (NCBI BioProjects: `PRJNA625551`, `PRJNA633948`, `PRJNA636748`, and `PRJNA647529`; see Methods for details) to validate the presence of the discovered mutational signatures. Specifically, we performed signature assignment with respect to the discovered signatures on 141, 23, 17, and 14 high-quality samples showing $\geq$6 always minor variants in each data set, respectively.

Three signature-based clusters are found for all data sets and explain more that 97% of the variance in all cases, with highly significant p values (see Figure 6). Such clusters are related to combinations of signatures consistently to the analysis presented in the text and display alike distributions of minor SVNs (see Figure S6).

These important results prove the generality of our findings and strongly support the hypothesis of distinct mutational processes active in distinct groups of samples.

DISCUSSION

Standard (phylo)genomic analyses of viral consensus sequences might miss useful information to investigate the elusive mechanisms of viral evolution within hosts and of transmission among hosts. In this respect,

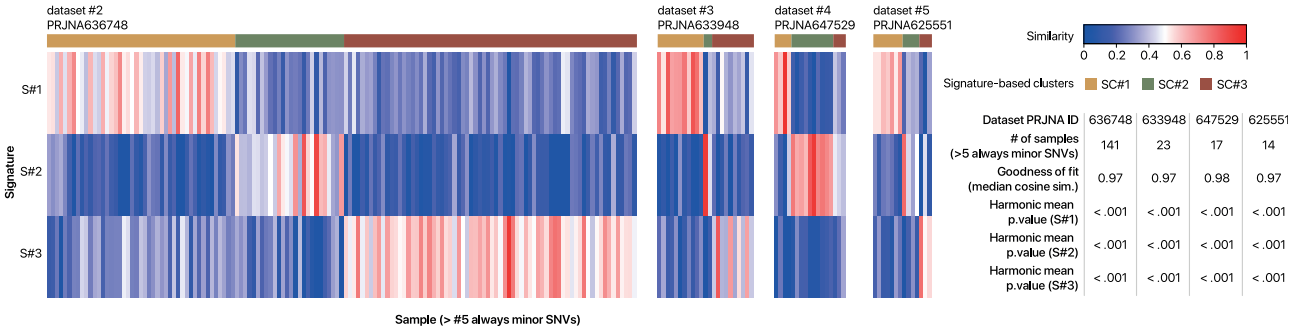


Figure 6. Validation on data sets #2 – 5 (NCBI BioProject: PRJNA636748, PRJNA633948, PRJNA647529, and PRJNA625551)

Heatmap returning the clustering of 141, 23, 17, and 14 samples of data sets #2 – 5 with ≥6 always minor variants (of the data set), computed via k-means on the low-rank latent NMF matrix on the three signatures discovered on data set ≈13% (see Methods). The goodness of fit in terms of median cosine similarity between observations and predictions and the harmonic mean p value of the one-sided Mann-Whitney U test on bootstrap re-sampling are shown for all signatures (see Methods).

raw sequencing data of viral samples can be effectively employed to deliver a high-resolution picture of intra-host heterogeneity, which might underlie different clinical outcomes and affect the efficacy of antiviral therapies. This aspect is vital especially during the critical phases of an outbreak, as experimental hypotheses are urgently needed to deliver effective prognostic, diagnostic, and therapeutic strategies for infected patients.

We here presented one of the largest up-to-date quantitative analyses of intra-host genomic diversity of SARS-CoV-2, which revealed that the large majority of samples present a complex genomic composition, likely due to the interplay between host-related mutational processes and transmission dynamics.

In particular, we here proved the existence of mutually exclusive viral mutational signatures, i.e., nucleotide substitution patterns, which show that different hosts respond to SARS-CoV-2 infections in different ways, likely ruled by APOBEC, ROS, or ADAR-related processes.

The corrected-for-signatures *dN*/*dS* analysis shows that such numerous low-frequency variants tend to be purified in the population whereas, conversely, a certain number of variants appear to consolidate. In particular, due to the still obscure combination of bottleneck effects and selection phenomena, certain variants appear to transit to clonality in the population, eventually leading to the definition of new clonal genotypes. Once become clonal, mutations tend to accumulate in the population, as proven by a statistically significant increase of genomic diversity, and might be used to reconstruct robust models of viral evolution via standard phylogenetic approaches.

The analysis of homoplasies, i.e., minor variants shared across distinct clades and unlikely due to infection events, demonstrates that a high number of mutations can independently emerge in multiple samples, due to mutational hotspots often related to signatures or, possibly, to positive (functional) selection.

In addition, the relatively higher VF of minor variants shared by multiple samples of the same clades supports the hypothesis of transmission during infections.

To conclude, we advocate the release of a larger number of raw sequencing data sets, especially in combination with clinical data, in order to investigate the relation among the discovered host-specific processes and clinical outcomes.

LIMITATIONS OF THE STUDY

Reference genome

Different reference genomes have been employed for variant calling in the investigation of the origin and evolution of SARS-CoV-2. For instance, sequence EPI_ISL_405839 was used, e.g., in Bastola et al. (2020) and sequence EPI_ISL_402125, e.g., in Andersen et al. (2020). As detailed in the Methods section, here we employed as reference the sequence SARS-CoV-2-ANC, which was identified in Ramazzotti et al., 2021

as a likely ancestral SARS-CoV-2 genome. Clearly, the use of different, albeit mostly overlapping, reference genomes can influence downstream analyses and, especially, the inference of the first evolutionary steps of the phylogenomic model, which should be therefore considered with caution. However, in our specific case, the employment of any of such reference genomes does not impact the identification and characterization of mutational signatures since the SNVs that distinguish such sequences are found as clonal in at least one sample of the data set and, accordingly, are excluded from the analysis.

Quasispecies composition

As discussed in the Introduction section, the analysis of raw sequencing data might be used to characterize the quasispecies architecture of single samples. To this end, a plethora of sophisticated computational methods for the characterization of the quasispecies composition of single samples is available, e.g., (Prosperi and Salemi, 2012; Giallonardo et al., 2014; Töpfer et al., 2014; Barik et al., 2018), and was recently reviewed in Knyazev et al. (2020). In the phylogenomic analysis included in this work, we decided to restrict the analysis on clonal variants that, by definition, are present in most of (or all) the quasispecies of a given sample. This allows us to provide a coarse-grained picture of the main steps of SARS-CoV-2 evolution and, at the same time, to investigate the possible transmission of minor variants, which are related to scarcely prevalent and rare quasispecies. It would be worth investigating how the combination of more sophisticated methods for quasispecies deconvolution and of our approach for mutational signatures analysis may improve the overall comprehension of SARS-CoV-2 diversity, adaptability, and evolution.

Data set quality

It was recently noted that some currently available SARS-CoV-2 data sets might present quality issues, especially with respect to low-frequency variants (De Maio et al., 2020). For this reason, the results of any computational pipeline should be, in principle, validated on data sets for which the ground truth is known. In our case and given the current shortage of SARS-CoV-2 benchmark data sets, we decided to validate the discovery and characterization of the mutational signatures on 4 different data sets, generated from independent laboratories worldwide, so to ensure the generality of the results obtained via our framework (see the Validation section).

Indels

The evolution of SARS-CoV-2 is characterized by the presence of a significant number of insertions and deletions (Koyama et al., 2020), which are being cataloged via COV-Glue (Singer et al., 2020). In this work, we focused on the analysis of SNVs, as this allows us to discover and characterize statistically significant host-related mutational signatures. Despite being beyond of the scope of the current work, it might be worth investigating the origination, evolution, and transmission of indels as well.

RESOURCE AVAILABILITY

Lead contact

Alex Graudenzi, Institute of Molecular Bioimaging and Physiology, Consiglio Nazionale delle Ricerche (IBFM-CNR), via F.lli Cervi, 93, 20,090 Segrate, Milan, Italy. alex.graudenzi@ibfm.cnr.it.

Material availability

This study did not generate new unique reagents.

Data and code availability

The source code used to replicate all the analyses is available at this link: https://github.com/BIMIB-DISCo/SARS-CoV-2-IHMV. VERSO can be downloaded at this link: https://github.com/BIMIB-DISCo/VERSO. Additional supplemental items are available from Mendeley Data at: https://doi.org/10.17632/vwc9jx5jfm.2.

METHODS

All methods can be found in the accompanying Transparent methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102116.

ACKNOWLEDGMENTS

This work was partially supported by the Elixir Italian Chapter and the SysBioNet project, a Ministero dell'Istruzione, dell'Università e della Ricerca initiative for the Italian Roadmap of European Strategy Forum on Research Infrastructures and by the Associazione Italiana per la Ricerca sul Cancro (AIRC)-IG grant 22082. We thank Marco Antoniotti, Giulio Caravagna, Chiara Damiani, Lucrezia Patruno, and Francesco Craighero for helpful discussions.

AUTHOR CONTRIBUTIONS

A.G., D.M., F.A., R.P., and D.R. designed and developed the study. A.G., D.M., F.A., and D.R. defined, implemented, and executed the computational analyses. A.G., D.M., F.A., R.P., and D.R. analyzed the data and interpreted the results. A.G., D.R., and R.P. supervised the study. All authors wrote the manuscript, discussed the results, and commented on the manuscript.

DECLARATION OF INTERESTS

The authors declare that they have no competing interests.

Received: September 8, 2020
Revised: November 9, 2020
Accepted: January 22, 2021
Published: February 19, 2021

REFERENCES

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L., et al. (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Ng, A.W.T., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., et al. (2020). The repertoire of mutational signatures in human cancer. Nature *578*, 94–101.

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., and Garry, R.F. (2020). The proximal origin of SARS-CoV-2. Nat. Med. *26*, 450–452.

Bandelt, H.J., Quintana-Murci, L., Salas, A., and Macaulay, V. (2002). The fingerprint of phantom mutations in mitochondrial DNA data. Am. J. Hum. Genet. *71*, 1150–1160.

Barik, S., Das, S., and Vikalo, H. (2018). QSdpR: viral quasispecies reconstruction via correlation clustering. Genomics *110*, 375–381.

Bastola, A., Sah, R., Rodriguez-Morales, A.J., Lal, B.K., Jha, R., Ojha, H.C., Shrestha, B., Chu, D.K., Poon, L.L., Costello, A., et al. (2020). The first 2019 novel coronavirus case in Nepal. Lancet Infect. Dis. *20*, 279–280.

Brunet, J.P., Tamayo, P., Golub, T.R., and Mesirov, J.P. (2004). Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. U S A *101*, 4164–4169.

Capobianchi, M.R., Rueca, M., Messina, F., Giombini, E., Carletti, F., Colavita, F., Castilletti, C., Lalle, E., Bordi, L., Vairo, F., et al. (2020). Molecular characterization of SARS-CoV-2 from the first case of COVID-19 in Italy. Clin. Microbiol. Infect. *26*, 954–956.

Chan, J.M., Carlsson, G., and Rabadan, R. (2013). Topology of viral evolution. Proc. Natl. Acad. Sci. U S A *110*, 18566–18571.

Daniloski, Z., Guo, X., and Sanjana, N.E. (2020). The D614G mutation in SARS-CoV-2 spike increases transduction of multiple human cell types. bioRxiv. https://doi.org/10.1101/2020.06.14.151357.

David, S.S., O'Shea, V.L., and Kundu, S. (2007). Base-excision repair of oxidative DNA damage. Nature *447*, 941–950.

Deng, X., Gu, W., Federman, S., du Plessis, L., Pybus, O.G., Faria, N., Wang, C., Yu, G., Bushnell, B., Pan, C.Y., et al. (2020). Genomic surveillance reveals multiple introductions of SARS-CoV-2 into northern California. Science *369*, 582–587.

Domingo, E., Martínez-Salas, E., Sobrino, F., de la Torre, J.C., Portela, A., Ortín, J., López-Galindez, C., Pérez-Breña, P., Villanueva, N., Nájera, R., et al. (1985). The quasispecies (extremely heterogeneous) nature of viral RNA genome populations: biological relevance-a review. Gene *40*, 1–8.

Domingo, E., Sheldon, J., and Perales, C. (2012). Viral quasispecies evolution. Microbiol. Mol. Biol. Rev. *76*, 159–216.

van Dorp, L., Acman, M., Richard, D., Shaw, L.P., Ford, C.E., Ormond, L., Owen, C.J., Pang, J., Tan, C.C., Boshier, F.A., et al. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect. Genet. Evol. *83*, 104351.

Van den Eynden, J., and Larsson, E. (2017). Mutational signatures are critical for proper estimation of purifying selection pressures in cancer somatic mutation data when using the dn/ds metric. Front. Genet. *8*, 74.

Forster, P., Forster, L., Renfrew, C., and Forster, M. (2020). Phylogenetic network analysis of SARS-CoV-2 genomes. Proc. Natl. Acad. Sci. U S A *117*, 9241–9243.

Giallonardo, F.D., Töpfer, A., Rey, M., Prabhakaran, S., Duport, Y., Leemann, C., Schmutz, S., Campbell, N.K., Joos, B., Lecca, M.R., et al. (2014). Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. Nucleic Acids Res. *42*, e115.

Di Giorgio, S., Martignano, F., Torcia, M.G., Mattiuz, G., and Conticello, S.G. (2020). Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Sci. Adv. *6*, eabb5813.

Grubaugh, N.D., Hanage, W.P., and Rasmussen, A.L. (2020). Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear. Cell *182*, 794–795.

Gutierrez, S., Yvon, M., Pirolles, E., Garzo, E., Fereres, A., Michalakis, Y., and Blanc, S. (2012). Circulating virus load determines the size of bottlenecks in viral populations progressing within a host. PLoS Pathog. *8*, e1003009.

Knyazev, S., Hughes, L., Skums, P., and Zelikovsky, A. (2020). Epidemiological data analysis of viral quasispecies in the next-generation sequencing era. Brief. Bioinform. bbaa101.

Korber, B., Fischer, W., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E., Bhattacharya, T., Foley, B., et al. (2020). Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell *184*, 812–827.e19.

Koyama, T., Platt, D., and Parida, L. (2020). Variant analysis of SARS-CoV-2 genomes. Bull. World Health Organ. *98*, 495.

Li, X., Wang, W., Zhao, X., Zai, J., Zhao, Q., Li, Y., and Chaillon, A. (2020). Transmission dynamics and evolutionary history of 2019-nCoV. J. Med. Virol. *92*, 501–511.

Lokman, S.M., Rasheduzzaman, M., Salauddin, A., Barua, R., Tanzina, A.Y., Rumi, M.H., Hossain, M.I., Siddiki, A.Z., Mannan, A., and Hasan, M.M. (2020). Exploring the genomic and proteomic variations of SARS-CoV-2 spike glycoprotein: a computational biology approach. Infect. Genet. Evol. *84*, 104389.

Lu, J., du Plessis, L., Liu, Z., Hill, V., Kang, M., Lin, H., Sun, J., François, S., Kraemer, M.U., Faria, N.R., et al. (2020). Genomic epidemiology of SARS-CoV-2 in guangdong province, China. Cell *181*, 997–1003.e9.

Lucas, M., Karrer, U., Lucas, A., and Klenerman, P. (2001). Viral escape mechanisms–escapology taught by viruses. Int. J. Exp. Pathol. *82*, 269–286.

Lythgoe, K.A., Hall, M.D., Ferretti, L., de Cesare, M., MacIntyre-Cockett, G., Trebes, A., Andersson, M., Otecko, N., Wise, E.L., Moore, N., et al. (2020). Shared SARS-CoV-2 diversity suggests localised transmission of minority variants. bioRxiv. https://doi.org/10.1101/2020.05.28.118992.

De Maio, N., Walker, C., Borge, R., Weilguny, L., Slodkowick, G., and Goldmand, N. (2020). Issues with SARS-CoV-2 Sequencing Data. https://virological.org/.

Molteni, C., Principi, N., and Esposito, S. (2014). Reactive oxygen and nitrogen species during viral infections. Free Radic. Res. *48*, 1163–1169.

Ni, M., Chen, C., Qian, J., Xiao, H.X., Shi, W.F., Luo, Y., Wang, H.Y., Li, Z., Wu, J., Xu, P.S., et al. (2016). Intra-host dynamics of ebola virus during 2014. Nat. Microbiol. *1*, 16151.

Nishikura, K. (2010). Functions and regulation of RNA editing by ADAR deaminases. Annu. Rev. Biochem. *79*, 321–349.

Novella, I.S., Domingo, E., and Holland, J.J. (1995). Rapid viral quasispecies evolution: implications for vaccine and drug strategies. Mol. Med. Today *1*, 248–253.

O'Toole, A., McCrone, J., and Scher, E. (2020). Pangolin 2.0. https://github.com/cov-lineages/pangolin.

Park, D., Huh, H.J., Kim, Y.J., Son, D.S., Jeon, H.J., Im, E.H., Kim, J.W., Lee, N.Y., Kang, E.S., Kang, C.I., et al. (2016). Analysis of intrapatient heterogeneity uncovers the microevolution of middle east respiratory syndrome coronavirus. Mol. Case Stud. *2*, a001214.

Plante, J.A., Liu, Y., Liu, J., Xia, H., Johnson, B.A., Lokugamage, K.G., Zhang, X., Muruato, A.E., Zou, J., Fontes-Garfias, C.R., et al. (2020). Spike mutation D614G alters SARS-CoV-2 fitness. Nature. https://doi.org/10.1038/s41586-020-2895-3.

Poon, L.L., Song, T., Rosenfeld, R., Lin, X., Rogers, M.B., Zhou, B., Sebra, R., Halpin, R.A., Guan, Y., Twaddle, A., et al. (2016). Quantifying influenza virus diversity and transmission in humans. Nat. Genet. *48*, 195.

Popa, A., Genger, J.W., Nicholson, M.D., Penz, T., Schmid, D., Aberle, S.W., Agerer, B., Lercher, A., Endler, L., Colaço, H., et al. (2020). Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. Sci. Transl. Med. *12*, eabe2555.

Prosperi, M.C., and Salemi, M. (2012). QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. Bioinformatics *28*, 132–133.

Ramazzotti, D., Angaroni, F., Maspero, D., Gambacorti-Passerini, C., Antoniotti, M., Graudenzi, A., and Piazza, R. (2021). VERSO: a comprehensive framework for the inference of robust phylogenies and the quantification of intra-host genomic diversity of viral samples. Patterns, 100212, https://doi.org/10.1016/j.patter.2021.100212.

Rambaut, A. (2009). Figtree v1. 3.1. http://tree.bio.ed.ac.uk/software/figtree/.

Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., and Pybus, O.G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. *5*, 1403–1407.

Reshi, M.L., Su, Y.C., and Hong, J.R. (2014). RNA viruses: ROS-mediated cell death. Int. J. Cell Biol. *467452*.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. *61*, 539–542.

Rose, R., Nolan, D.J., Moot, S., Feehan, A., Cross, S., Garcia-Diaz, J., and Lamers, S.L. (2020). Intra-host site-specific polymorphisms of SARS-CoV-2 is consistent across multiple samples and methodologies. medRxiv. https://doi.org/10.1101/2020.04.24.20078691.

Seemann, T., Lane, C.R., Sherry, N.L., Duchene, S., Gonçalves da Silva, A., Caly, L., Sait, M., Ballard, S.A., Horan, K., Schultz, M.B., et al. (2020). Tracking the covid-19 pandemic in Australia using genomics. Nat. Commun. *11*, 4376.

Sharma, S., Patnaik, S.K., Taggart, R.T., Kannisto, E.D., Enriquez, S.M., Gollnick, P., and Baysal, B.E. (2015). Apobec3a cytidine deaminase induces rna editing in monocytes and macrophages. Nat. Commun. *6*, 1–15.

Shen, Z., Xiao, Y., Kang, L., Ma, W., Shi, L., Zhang, L., Zhou, Z., Yang, J., Zhong, J., Yang, D., et al. (2020). Genomic diversity of SARS-CoV-2 in coronavirus disease 2019 patients. Clin. Infect. Dis. *71*, 713–720.

Shu, Y., and McCauley, J. (2017). GISAID: global initiative on sharing all influenza data–from vision to reality. Eurosurveillance *22*, 30494.

Simmonds, P. (2020). Rampant C -> U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses: causes and consequences for their short-and long-term evolutionary trajectories. MSphere *5*.

Singer, J., Gifford, R., Cotten, M., and Robertson, D. (2020). CoV-GLUE: a web application for tracking SARS-CoV-2 genomic variation. Preprints. https://doi.org/10.20944/preprints202006.0225.v1.

Tang, X., Wu, C., Li, X., Song, Y., Yao, X., Wu, X., Duan, Y., Zhang, H., Wang, Y., Qian, Z., et al. (2020). On the origin and continuing evolution of SARS-CoV-2. Natl. Sci. Rev. *7*, 1012–1023.

Töpfer, A., Marschall, T., Bull, R.A., Luciani, F., Schönhuth, A., and Beerenwinkel, N. (2014). Viral quasispecies assembly via maximal clique enumeration. PLoS Comput. Biol. *10*, e1003515.

Wölfel, R., Corman, V.M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M.A., Niemeyer, D., Jones, T.C., Vollmar, P., Rothe, C., et al. (2020). Virological assessment of hospitalized patients with COVID-2019. Nature *581*, 465–469.

Woo, P.C., Wong, B.H., Huang, Y., Lau, S.K., and Yuen, K.Y. (2007). Cytosine deamination and selection of cpg suppressed clones are the two major independent biological forces that shape codon usage bias in coronaviruses. Virology *369*, 431–442.

World Health Organization (WHO) (2020). Coronavirus Disease 2019 (COVID-19): Situation Report. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.

Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., et al. (2020). A new coronavirus associated with human respiratory disease in China. Nature *579*, 265–269.

Xiao, K., Zhai, J., Feng, Y., Zhou, N., Zhang, X., Zou, J.J., Li, N., Guo, Y., Li, X., Shen, X., et al. (2020). Isolation of SARS-CoV-2-related coronavirus from malayan pangolins. Nature *583*, 286–289.

Xu, D., Zhang, Z., and Wang, F.S. (2004). Sars-associated coronavirus quasispecies in individual patients. N. Engl. J. Med. *350*, 1366–1367.

Zhou, B., Thao, T.T.N., Hoffmann, D., Taddeo, A., Ebert, N., Labroussaa, F., Pohlmann, A., King, J., Portmann, J., Halwe, N.J., et al. (2020a). SARS-CoV-2 spike D614G variant confers enhanced replication and transmissibility. bioRxiv. https://doi.org/10.1101/2020.10.27.357558.

Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., et al. (2020b). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature *579*, 270–273.

4

Multiscale modelling and simulation of multicellular systems

In this chapter, I discuss the attempts of characterizing the heterogeneity of multicellular systems via multiscale modelling and simulation.

After a brief introduction on Cellular Potts Model and Flux Balance Analysis (section 4.1), the Flux Balance Cellular Automata (FBCA) approach is introduced, alongside with its extension to model nutrients diffusion is also presented (section 4.2).

4.1 Background: Cellular Potts Model and Flux Balance Analysis

The general goal of this research was to conceive an expressive multiscale model for the simulation of the interactions of cell in multicellular systems and, particularly, of the cancer-microenvironment interplay. Such framework needs to be flexible, also allowing for the possible integration with omics data. As a consequence, we designed a new hybrid model where the metabolism of individual cells drives the cellular population dynamics, by coupling the Cellular Potts Models with the simulation of metabolic models via Flux Balance Analysis. More details are provided in the following sections.

Cellular Potts Model – CPM. The Cellular Potts Model (CPM, also known as the Glazier-Graner-Hogeweg model) is a type of Cellular Automaton, which is widely used to investigate tissue morphogenesis [124], tissue engeneering [203], and cancer disease

[102, 183]. It is used to simulate growth, proliferation, movement and apoptosis of single cell in complex environments [106].

Briefly, CPM represents biological cells as a set of contiguous lattice sites identified with the same numerical identifier. The representation allows cells to have arbitrary dimensions and shapes. The evolution of the system is driven by a Hamiltonian functional (H) which specifies the cells' properties relevant to their evolution as the simulation progresses. H comprises many additive terms representing different biological properties or phenomena based on the problem under investigation. Often, Hamiltonian functional includes cell-cell adhesion properties or target size. The former represent the tendency of cells to be near other cells of the same types, whereas the latter defines the area/volume (in terms of the number of lattice sites) that each cell pursue to reach, modelling the growth tendency.

Once the functional is defined, the system evolution proceeds stochastically based on energy minimization using a dynamic Monte Carlo simulation algorithm. In particular, a random pair (source and target) of adjacent lattice sites are randomly chosen in each simulation step. In the move the state of the source is assigned to the target. For example, if the pair of lattice sites are part of two different cells, then one increases its size while the other is shrinked. The move is finally accepted or rejected based on the change of the system's energy.

Notice that lattice sites can also represent other biological components, e.g. extracellular matrix. For instance, some works interfaced CPM with simulations of a range of biological processes [49, 134].

Flux Balance Analysis – FBA. Flux Balance Analysis (FBA) is a widely used technique for studying networks of biochemical reactions. FBA compute the flux distribution of metabolites through such networks, allowing, for example, the growth rate prediction of an organism or the secretion rate of metabolites [44].

In FBA, biochemical reactions are mathematically represented via a numerical matrix, including each reaction's stoichiometric coefficients. The stoichiometric matrix imposes constraints on the direction of the flux through the network. The flux of each reaction is constrained in two ways: (i) an equation balances the reaction inputs and outputs, and (ii) an inequality imposes its upper and lower bounds (i.e., maximum and minimum flux rates). All the constraints define the space of feasible flux distributions of a given metabolic network.

Importantly, FBA assumes that the system is at the *steady-state*: fluxes and concentrations are supposed to be constant in time, so the sum of all fluxes is always equal to 0. Moreover, one also needs to define a set of boundary conditions, i.e., fluxes of metabolites as input and output from the system, which correspond to the uptaken and secreted metabolites from cells or tissue. A significant advantage of this assumption is that it

is possible neglect the knowledge of many difficult-to-measure parameters required in kinetics models.

The last requirement needed to perform FBA is the definition of a proper biological *objective function* for the problem under investigation. Since proliferation is a hallmark of cancer disease, the optimization of biomass production is often chosen as the system objective function. The mathematical representations of the metabolic reactions, the constraints and the objective function define a system of linear equations, which can be efficiently solved by exploiting linear programming.

Note that FBA is extensively applied to model the metabolism of complex tissues or diseases, e.g., [99]. Interestingly, the lower and upper bound of reaction fluxes can be used as scaffold for the analysis of high throughput data to allow integration of expression data as, e.g., in [132], as well as in some of our previous works [175] and in a new work introduced and discussed in section 5.2.

4.2 Flux Balance Cellular Automata – FBCA

The method developed during the PhD project is named FBCA (Flux Balance Cellular Automata), and is a multiscale modelling framework that combines a biophysically plausible representation of cell morphology and interactions via Cellular Potts Model, and a model of cellular metabolic activity, via Flux Balance Analysis. In this framework, the population dynamics is driven by a cell growth function determined by the optimization of an independent metabolic model assigned to each cell. This algorithmic choice allowed us to achieve an optimal trade-off between expressivity and computational complexity.

The workflow can be briefly summarized as follows: (*i*) we first assign a metabolic model to each synthetic cell; (*ii*) we apply FBA by considering (*A*) the sum of nutrient concentrations in cell's surrounding lattice sites as model boundary conditions, and (*B*) biomass optimization as objective function; (*iii*) the biomass produced by each metabolic model is linked to a Hamiltonian functional term, representing the tendency to grow. This term considers the biomass accumulated and the actual size of a given cell to determine its tendency to increase; (*iv*) when a cell reaches a target area, it divides into two daughter cells simulating a mitosis process. Thus, cells that can produce a higher biomass level have an increased growth tendency than the others.

Despite the abstractions underlying CPM and FBA, we proved that FBCA could reproduce complex phenomena observed in real-world biological systems, such as cell migration, tissue colonisation and homeostasis, allowing us to perform in-depth quantitative analyses of crucial properties in a variety of in-silico experimental settings. In fact, we could evaluate different properties emerging from the simulation. For instance, we could define metabolic heterogeneous cell populations and observe how the space got colonized. We could also evaluate how the microenvironment determines the heterogeneous

distribution of metabolic behaviours, e.g., in cells far or near nutrient sources.

Here, we report two articles which include the definition and application of FBCA for the investigation of real-world phenomena, and related extensions.

In particular, in article paper P#8 we introduce the algorithmic framework and investigate two in-silico experimental scenarios.

- *Fermentative vs. oxidative behaviour.*
  We first populated the initial lattice with two distinct metabolic sub-populations: one in which cells are allowed to intake lactate, but not to secrete it (oxidative cells), and vice versa (fermentative cells). We did this by setting the constraints of their metabolic models accordingly. We performed a large number of simulations to observe how the competition for space drives the emergent behaviour, in environments with either uniform and heterogeneous nutrient distributions.

- *Subpopulations with distinct metabolic fitness.*
  In the second scenario, we define three metabolic subpopulations, whose biomass production rate is a specific function of the oxygen level present in the microenvironment. For example, one population grows faster in low-oxygen regions, while it is disadvantaged in the high-oxygen area. The idea was to simulate a complex scenario with non-uniform nutrient distribution and biomass production. In addition, we simulated a perturbation by removing the most proliferative population and observing how the surviving cells recolonize the space, mimicking the case in which a therapy is effective with one population only, leaving the resistant ones unharmed.

In paper P#9, the FBCA framework was extended to account for nutrients diffusion. To this end, following the positive results obtained in [27], we proposed to compute the diffusion of arbitrary nutrients by averaging their concentration in the neighborhood of each lattice site. Importantly, we modelled metabolic interactions by allowing the metabolite secreted by cells to diffuse over the lattice and by positioning discrete nutrient sources along the lattice. We evaluated the differences of two nutrients diffusion models, namely: (*i*) *impermeable cells* or (*ii*) *permeable cells*. In the former, lattice sites occupied by cells cannot contain nutrients. In the latter, lattice sites can contain both cells and nutrients. Notice that we considered a specific diffusion coefficient for each metabolite (in the range $(0, 1]$), which reduces the concentration changes to mimic the different size of the molecules (e.g., Oxygen with respect to Glucose).

In the article, we investigated various experimental settings to evaluate the impact of the nutrients diffusion and of the shape of the microenvironment, especially with respect to the emergence of metabolic behaviour. In detail, we analyzed the difference between a closed environment, e.g., representing a simplified cell culture flask, and a tissue-like environment, including a schematic representation of blood vessels. In both

cases, our model was able to reproduce real-world phenomena, and we could evaluate the competition between metabolic subpopulations.

Overall, our analyses demonstrated that the choice of the most appropriate experimental setting depends on the aim of the research. For instance, the impermeable cells scenarios allowed us to improve the characterization of nutrient consumption and secretion dynamics, despite a substantial increase in computational time.

Let us finally point out that, in its original formulation, FBCA does not allow to consider any omics measurement, e.g., from single-cell RNA-seq data. The preliminary attempts of defining and applying a data-driver multiscale modelling framework are discussed in the next chapter.

Journal of Cellular Automata, Vol. 15, pp. 75–95
Reprints available directly from the publisher
Photocopying permitted by license only

©2019 Old City Publishing, Inc.
Published by license under the OCP Science imprint,
a member of the Old City Publishing Group.

FBCA, A Multiscale Modeling Framework Combining Cellular Automata and Flux Balance Analysis

ALEX GRAUDENZI[1,2,5], DAVIDE MASPERO[2,3,5] AND CHIARA DAMIANI[2,4,*]

[1]*Institute of Molecular Bioimaging and Physiology of the Italian National Research Council (IBFM-CNR), Segrate, Milan, Italy*

[2]*Dept. of Informatics, Systems and Communication, University of Milan-Bicocca, Milan, Italy*

[3]*Dept. of Research, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy*

[4]*SYSBIO Centre of Systems Biology, Milan, Italy*

[5]*Equal contributors*

Received: February 25, 2019. Accepted: March 23, 2019.

Multiscale computational models are powerful instruments to investigate the properties of biological systems, provide explanations about their complex behaviours and predict their likely future evolution. We here investigate the relation among the metabolic properties of cells and the emerging population dynamics of multi-cellular systems, such as tissues and organs, especially in abnormal circumstances, such as cancer origination and development.

To this end, we introduce FBCA (Flux Balance Cellular Automata) a multiscale modeling framework that combines a cellular automaton representation of the spatial/morphological dynamics of multi-cellular systems, i.e., the Cellular Potts Model, with a model of the metabolic activity of individual cells, as modeled via Flux Balance Analysis. With this framework, it is possible to investigate, both qualitatively and quantitatively, the dynamics and the spatial behavior of cell sub-populations in a variety of experimental scenarios.

In particular, we here show the results of a simplified model of intestinal crypt, which is the locus in which colorectal cancer is supposed to originate. We show that competition and selection phenomena are indeed largely driven by the metabolic properties of the cell sub-populations populating the crypt, leading to often non predictable dynamics. Finally, we present a scenario in which cancer recurrence is explained by the presence of non-dominant drug-resistant subclones.

Keywords: Cellular potts model, metabolic networks, flux balance analysis, population dynamics, multi-cellular systems

* Contact author: E-mail: chiara.damiani@unimib.it

1 INTRODUCTION

The properties of multi-cellular systems, such as tissues and organs, can be effectively investigated via computational models, which allow to test a huge number of experimental scenarios and parameter settings *in silico*, and produce statistically robust quantitative results and predictions on complex biological phenomena [11, 30].

Many modeling approaches have been developed in the last years at this aim, including compartment models, in which population dynamics is analyzed via mean-field approximations [2, 3], and *off-* and *in-lattice* models, which account for the spatial and mechanical properties of cells [4, 14, 18, 21, 24, 27, 33].

Cellular automata (CA), in particular, are widely used to represent in a very efficient way cell displacement, movement and interactions, and are often employed in multiscale models, which describe processes and phenomena occurring at different space/time scales [19, 31]. For instance, CA-based multiscale models including gene regulatory networks dynamics were effectively used to investigate the necessary conditions for homeostasis [14], or the clonal expansion of cancer sub-populations [24].

However, there is an apparent shortage of effective modeling approaches to investigate the influence of metabolism on cell population dynamics in multi-cellular systems. Even though increasing experimental evidences suggest that metabolic regulation and reprogramming play a central – and still undeciphered – role in a broad number biological complex phenomena, such as cancer development [15, 32], most models of metabolism are currently limited to the simulation of steady state behaviours of individual cells (or of an *average* cell), and of basic interactions via exchange of nutrients [20]. Cell population dynamics is usually not modeled, nor are considered the biophysical properties of cells and their interactions in tissues or organs. For instance, in [16] a steady state condition is assumed for the population composition, while in [9] a single snapshot of the composition of a population in time is depicted.

To fill this gap, in this work we propose a modeling framework specifically aimed at investigating the relation among high-level spatial properties of a generic multi-cellular systems and the metabolism of its constituting cells.

FBCA (Flux Balance Cellular Automata) is a multiscale modeling framework originally introduced in [1], which includes two distinct and interacting levels:

1. A spatial/morphological level, modeled via the Cellular Potts Model (CPM) [13], in which biological cells are represented by sets of

contiguous cells over a lattice, and the overall dynamics is probabilistically driven according to an energy minimization criterion, as provided by an Hamiltonian function. In CPM framework cells can expand, move, undergo mitosis, die and interact with each other. CPM has been used in several works and proved to reproduce complex emergent properties of real multi-cellular systems (see, e.g., [26]).

2. A metabolic network level, in which the metabolic activity of each individual cell is represented via Flux Balance Analysis, which is by far the most used approach to simulate the dynamics of individual metabolic networks [6]. FBA relies on Linear Programming to determine a metabolic flux distribution (i.e. the rate of turnover of molecules through each reaction) that maximizes/minimizes a predefined objective function, given constraints on: i) the stoichiometry of reactions; ii) the steady state assumption for internal metabolites; iii) constraints on the domain of the metabolic fluxes, as derived from experimental measurements or from reaction thermodynamics.

The growth rate at the spatial level (and consequently the cell replication pace) is determined as a function of the biomass increase, computed for each cell via FBA computation. Conversely, the emergent spatial dynamics at the spatial level influences the distribution of nutrients among cells, which is essential for cell survival and growth.

Our modeling approach is rooted in statistical physics and complex systems, as we aim at designing the simplest possible model (with fewer parameters) able to reproduce complex phenomena of multicellular systems, which might be experimentally validated. Therefore, we keep the *a priori* assumptions at a minimum and we investigate the *emergent* dynamical properties of the system, taking advantage of extensive simulations in different experimental settings, as proposed for instance in [10, 23], where the dynamical interaction of simplified models of gene regulatory networks was investigated by employing a cellular automata-based representation of space.

To our knowledge, this is the first attempt to connect the dynamical behavior of metabolic networks to biophysically realistic spatial and morphological properties of real multi-cellular systems. Notice that in [29] the authors introduced a multiscale model of colonic carbohydrate metabolism and bacterial population dynamics, with a simplified geometrical representation of the gut, which however does not take into account any biophysical property of the cells and of the surrounding space and, thus, cannot account for morphological and space competition dynamics.

In particular, we here focused on the representation of a generic intestinal crypt, as this particular biological structure has been largely characterized

and is supposed to be the locus in which colorectal cancers originate. Furthermore, the geometrical structure of crypts, which are single-layer one-side open cylinders, allows to employ simplified representation of space.

We investigated various distinct experimental scenarios, related to competition phenomena in environments with limited space and resources, and to the possible emergence of positively selected (cancer) sub-populations, characterized by specific and advantageous metabolic properties. We also present a preliminary study in which cancer recurrence is determined by the expansion of non-dominant drug-resistant subclones, following the wipe-out of the dominant clone due to therapy.

We remark that, in [1] a preliminary version of FBCA was introduced, in which a proof-of-principle of its usage and applicability was presented. In this work, in addition to testing distinct biologically realistic experimental scenarios, we included a more plausible representation of spatial interaction among cells, based on the Differential Adhesion Hypothesis, a neighbourhood-based limitation of biomass production, a more complex metabolic nutrient diffusion scheme, based on the intake/secretion activity of single cells, as well as a basic model of cancer therapy.

In Section 2 the FBCA computational framework is presented; in Section 3 the simulation settings are described; in Section 4 the results of the simulations in the distinct experimental scenarios are presented; finally Section 5 contains the discussion about the usefullness of the approach and possible future developments.

2 METHODS

FBCA is a multiscale modeling framework that includes a spatial representation of a generic multi-cellular system, as modeled via CPM, and a model of metabolism of its constituting cells, simulated via FBA.

Simulation of the metabolic activity of individual cells

The metabolic network of a generic cell σ is defined as a set $\mathcal{M} = \{m_1, \ldots, m_N\}$ of metabolites in the system and the set $\mathcal{R} = \{r_1, \ldots, r_M\}$ of chemical reactions taking place among them. Reactions are be defined as:

$$R_j : \sum_{i=1}^{N} \alpha_{ji} M_i \longleftrightarrow \sum_{i=1}^{N} \beta_{ji} M_i, \tag{1}$$

where $\alpha_{ji}, \beta_{ji} \in \mathbb{N}$ are stoichiometric coefficients associated, respectively, with the i-th reactant and the i-th product of the j-th reaction, with $i =$

$1, \ldots, N$, $j = 1, \ldots, M$. Let $[M_i]$ be the abundance of reactant M_i and v_j the flux of reaction R_j, i.e., the net value between forward and backward reaction rate.

Because a steady state is assumed for the abundance of each metabolite, i.e., $d[M_i]/dt = 0 \ \forall i$, Linear Programming is applied to identify the flux distribution $\vec{v} = (v_1, \ldots, v_M)$ that maximizes (or minimizes) the objective $Z = \sum_{j=1}^{M} w_j v_j$, where w_j is a coefficient that represents the contribution of flux j in vector $\vec{v}$ to the objective function Z.

In our simulations, we typically set the maximization of the rate of biomass production as objective function. It is standard practice in FBA computations [20] to approximate this rate with the flux of a pseudo-reaction, representing the conversion of biomass precursors into biomass.

Given a $N \times M$ stoichiometric matrix S, whose element s_{ji} takes value: i) $-\alpha_{ji}$ if metabolite M_i is a reactant of reaction R_j, ii) $+\beta_{ji}$ if metabolite M_i is a product of reaction R_j, and iii) 0 otherwise.

In order to determine the biomass $\mathcal{B}_\sigma$ produced (and then accumulated) by cell σ in the unit of time (MCS), we solve the following Linear Programming Problem.

$$\begin{aligned} &\text{maximize } \ \mathcal{B}_\sigma \\ &\text{subject to } S\vec{v} = \vec{0}, \ \vec{v}_L \leq \vec{v} \leq \vec{v}_U \end{aligned} \tag{2}$$

where $\vec{v}_L$ and $\vec{v}_U$ are two vectors specifying, respectively, the lower and upper bounds of the admitted interval of each flux v_j. A negative lower bound indicates that flux is allowed in the backward reaction. The exchange of matter with the environment is represented as a set of exchange reactions, in the form $M_i \longleftrightarrow \emptyset$, enabling a predefined set of species to be inserted in or removed from the network.

Cellular automata representation of tissue morphology

FBCA employs a simplified geometrical representation of a general tissue, based on the CPM [13], a cellular-automaton modeling framework often used to model energy-driven spatial pattern formation [26] (Figure 1).

In this specific case we use a 2D representation of space, which is suitable to model single-layer tissues, yet the model could be easily extended to the 3D scenario. More in detail, the space is a rigid 2D grid with square lattice sites, opened and rolled out onto a rectangular $h \times w$ lattice L through periodic boundary condition, to mimic the morphology of intestinal crypts (i.e., approximately a lower-side opened cylinder constituted of single-layer epithelial cells).

A *biological* cell, identified with σ_i, is delimited by connected domains: the space occupied by cell σ_i is denoted as $\mathcal{C}(\sigma_i)$ and consists of all lattice

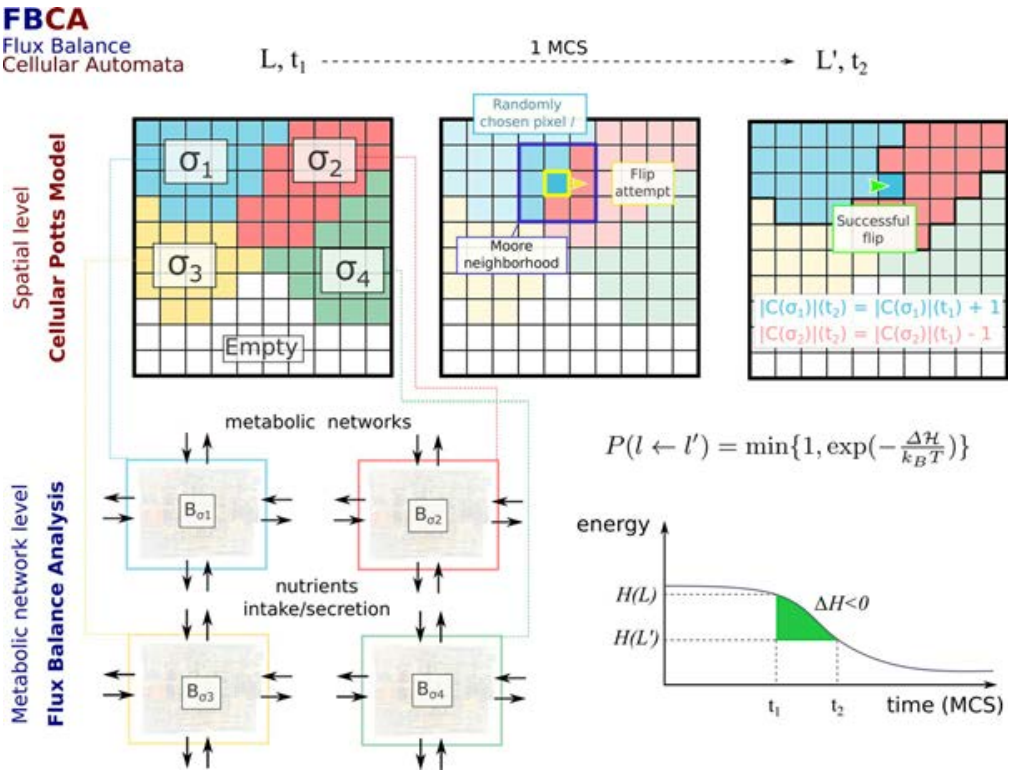


FIGURE 1
FBCA scheme. In FBCA, biological cells (σ_1, σ_2, σ_3 and σ_4) are represented as sets of contiguous lattice sites over a lattice L with periodic boundary condition and opened at the lower side. Each cell includes an individual metabolic network, which is used in FBA computation to determine the biomass gain at each time step, according to the nutrients distributed over the lattice, which will be used to compute the Hamiltonian function in equation 5. An example MCS step is shown: at time t_1 a random lattice site in L is chosen, belonging to cell σ_1; a flip attempt is attempted with a randomly chosen lattice site in its Moore neighborhood and evaluated via the Hamiltonian function. In this case the flip is accepted and the lattice site is transferred from cell σ_2 to cell σ_1 at time t_2.

sites $l \in L$ with value σ_i:

$$\mathcal{C}(\sigma_i) = \{l = \sigma_i | l \in L\} \tag{3}$$

For every disposition of cells, an Hamiltonian energy function $\mathcal{H}$ is evaluated, to account for the energy required for each mutual interaction, as well as other physical quantities. The dynamics of the system is driven by a discrete-time stochastic process (time unit: *Monte-Carlo Step* - MCS), in which cells are rearranged in order to minimize the Hamiltonian energy of the whole lattice. To this end, lattice sites of a given cell are probabilistically chosen to be *flipped* in favor of another cell in its neighborhood, and this allows cells to move over the lattice.

The update procedure can be summarized as follows: a lattice site l is selected with uniform probability in lattice L; another random lattice site l'

Symbol	[SC 1]	[SC 2]	Description
–	1 lattice site = $= 1\mu m$	1 lattice site = $= 1\mu m$	Conversion of space unit
–	1 MCS = 1/10 hour	1 MCS = 1/10 hour	Conversion of time unit
h	155 lattice sites	150 lattice sites	Height of the lattice
w	100 lattice sites	105 lattice sites	Width of the lattice
k	4	4	Number of lattice spin attempts per lattice site per MCS
$\mathcal{N}$	1	1	Moore neighborhood size
λ	1	4	Area rigidity constraint
$k_B T$	3	3	Temperature and Boltzmann constant
$A_{mitosis}$	50 lattice sites	50 lattice sites	Mitosis area for all cells
$J_{Normal_A - Normal_A}$	4	4	Hamiltonian adhesion factor among Cells of the same type
$J_{Normal_A - Normal_B}$	4	8	Hamiltonian adhesion factor among Cells of different type
$J_{Normal - Empty}$	0.5	2	Hamiltonian adhesion factor among Cells and empty space
$\mathcal{F}$	0.02	0.02	Area/biomass conversion factor
$[O_2]$	6 fmol / lattice site l	[0.1, 2.25] fmol / lattice site l	Oxygen abundance
$[Glc]$	0.5 fmol / lattice site l	0.4 fmol / lattice site l	Glucose abundance
$[Lact]$	0.5 fmol / lattice site l	relies on overall uptake and secretion	Lactate abundance
$[Gln]$	20 fmol / cell	0.4 fmol / lattice site l	Glutamine abundance
$[Arg]$	20 fmol / cell	20 fmol / cell	Arginine abundance

TABLE 1
Parameter settings. Most parameters for the simulations presented in this work have been chosen in accordance with existing literature and allow for a biophysically plausible representation of intestinal crypts [14, 24, 33].

is chosen in its Moore neighborhood $\mathcal{N}(l)$; the lattice site l' is then assigned to the cell including l with probability:

$$P(l \leftarrow l') = \min\{1, \exp(\frac{-\Delta\mathcal{H}}{k_b T})\} \tag{4}$$

where $\Delta\mathcal{H}$ is the Hamiltonian difference if the flip is accepted. A $h \times w \times k$ number of flips is attempted at each MCS (the parameters of the simulations are provided in Table 1). In equation 4 the Boltzmann distribution is used to drive cells to the configuration with minimum energy, whereas the factor $k_b T$ accounts for the amplitude of the cell membrane fluctuations.

In our framework, the Hamiltonian function has two main components, which subsume: i) the growth tendency of each cell, and ii) the Differential Adhesion Hypothesis (DAH) [28].

In order to account for the DAH, according to which cells of different types tend to segregate and form distinct compartments, we include in our model the possibility of having different cell types, plus a further abstract type, i.e., the empty space, along the lines of [14, 33]. The intuition is that cells will first tend to fill the empty space, if available in their surroundings, and then stay close to cells of the same type.

Cell growth is defined as a function of the biomass $\mathcal{B}_{\sigma_i}$ produced and accumulated by each cell, and computed via Flux Balance Analysis at each MCS (see below). In particular, at each MCS, cell σ_i will tend to grow toward an objective area $A_{target}(\mathcal{B}_{\sigma_i})$. To link the biomass growth, which is measured in pico grams (pg), to the target area, which is measured in lattice sites (1 lattice site is equal to $1\mu m$), a conversion factor is needed and defined (see Table 1). This defines the multiscale link between the spatial and the metabolic levels.

The Hamiltonian function is then defined as:

$$\mathcal{H}(L) = \frac{1}{2} \sum_{\sigma_i, \sigma_j \in \mathcal{N}} J(\tau(\sigma_i), \tau(\sigma_j))(1 - \delta(\sigma_i, \sigma_j)) + \lambda \sum_i [|\mathcal{C}(\sigma_i)| - A_{target}(\mathcal{B}_{\sigma_i})]^2 \tag{5}$$

where i and j are lattice sites $\in L$, σ_i is the cell at site i, δ is the Kronecker delta, $\tau(\sigma_i)$ is the cell type of cell σ_i, $J(\tau(\sigma_i), \tau(\sigma_j))$ is the amount of energy required to stick tied cells σ_i and σ_j according to the DAH (which in our case will first favor the migration toward empty space and then toward cells of the same type – see Table 1), $|\mathcal{C}(\sigma_i)|$ is the current area of cell σ_i in lattice sites, and $\lambda > 0$ is a Lagrange multiplier that accounts for the capacity to deform a cell membrane.

Cells grow via the accumulation of biomass, as for equation 5, up to an objective area $A_{mitosis}$, which is initially set as double than the area of cells in the initial lattice configuration; when $|\mathcal{C}(\sigma_i)| = A_{mitosis}$, cell σ_i is divided in two daughter cells, by splitting its space along a randomly chosen direction (either horizontal or vertical), thus modeling *symmetric* cell division. Daughter cells will initially have area (approximately) equal to $\frac{A_{mitosis}}{2}$ and will inherit the metabolic network of the parent cell.

As we are modeling intestinal crypts, we recall that the lower boundary of the lattice is open: the expulsion of cells in the intestinal lumen is modeled by deleting the cells that reach the lower boundary from the lattice. Therefore, cell migration toward the open boundary is expected, due to cell growth and duplication dynamics.

Implementation. This version of FBCA has been implemented in MATLAB, so to exploit both the COBRA Toolbox [25] for FBA computation and matrix calculus for CPM computations. The computation time for a single MSC in a standard simulation setting is ~ 3 secs (ASUS NOTEBOOK;

CPU:Intel Core i7-4710HQ CPU @ 2.50GHz, 4 core; RAM: 16.0 GB; WINDOWS 10 pro 64-bit).

3 SIMULATION SETTINGS

The parameter settings used in the simulations are reported in Table 1.

In all simulations, we employed as metabolic network model associated to each cell the model of central carbon metabolism HMRCORE introduced in [12] and used in [9], composed of 240 metabolites and 272 reactions. We also assume a constant supply of nutrients across the lattice: at each MCS, each cell is supplied with an amount of nutrients (i.e., upper bound of intake flux) that is proportional to the area of the cell. $|C(\sigma_i)|$ is the number of lattice sites occupied by cell σ_i; let $[M_j^l]$ be the abundance assumed for metabolite j in site $l \in \mathcal{C}(\sigma_i)$; the upper bound $U_j^{\sigma_i}$ of the exchange reaction of metabolite j for cell σ_i is set as: $U_j^{\sigma_i} = \sum_{l=1}^{|C(\sigma_i)|}[M_j^l]$.

In an earlier work [1], we tested the FBCA framework in a control scenario, in which we populated the crypt with two sub-populations, standing for simplified cancer cells (maximizing biomass) and normal cells (maximizing ATP). We here present the results of extensive simulations in two distinct and biologically plausible experimental scenarios.

Scenario 1 [SC 1] - Fermentative vs. oxydative behaviour
Lactate is supposed to play a pivotal role in cancer metabolism. In fact, it is secreted by cells under the Warburg effect, it is often found in tumor microenvirorment, and it is usually exchanged in the interplay involving stroma and cancer cells [8]. For this reason, in this analysis we populated the crypt with two different metabolic sub-populations, one in which cells are allowed to intake lactate, but not to secrete it (*oxydative* cells) and the other viceversa (*fermentative* cells). We refer to these two populations respectively as "*type 1*" and "*type 2*". We simulated the population dynamics in two distinct environmental settings:

- [**A**] uniform distribution of nutrients across the lattice (see Figure 2);
- [**B**] non-uniform distribution of nutrients, as obtained by including two areas of 3875 lattice sites, positioned at the left/right sides of the lattice, are characterized by limited oxygen abundance ($[O_2] = 0.5$ fmol / lattice site), thus mimicking an hypoxic area (see Figure 3).

In all initial configurations (i.e., MCS = 0) cells are drawn as squares of 5×5 lattice sites, and fill the whole lattice. Example of initial configurations

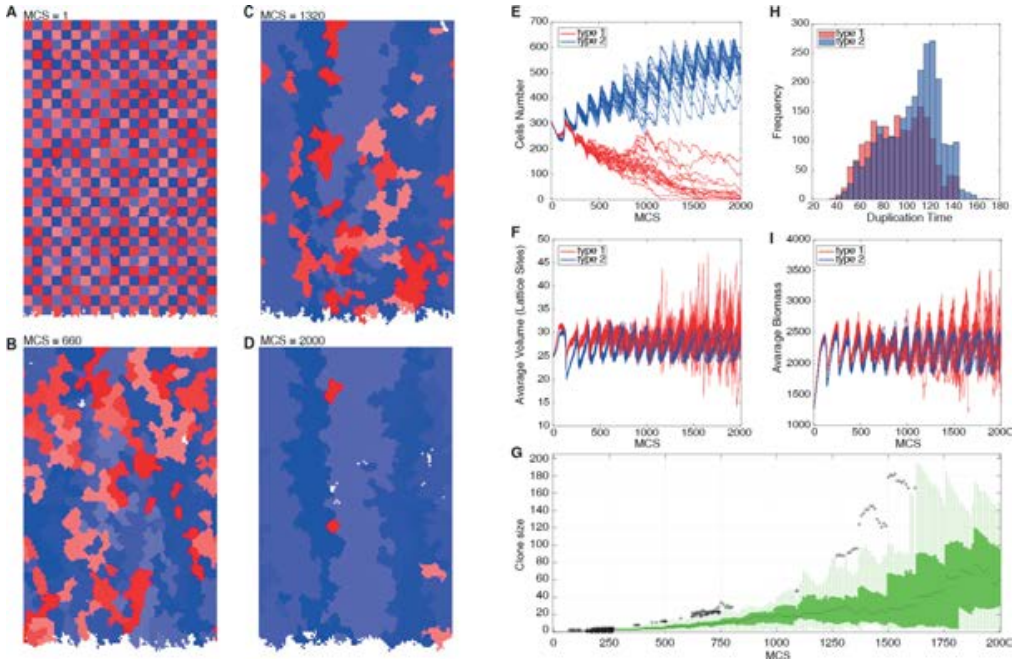


FIGURE 2
[SC 1 - A]. (A-D) Snapshots of FBCA dynamics respectively at 1, 660, 1320 and 2000 MCSs of an example simulation; shapes with red tonalities indicate *type 1* cells, shapes with blue tonalities indicate *type 2* cells; identical color refers to the same clonal population. (E) Total number of cells of each type as a function of time; one curve for each of the 20 simulations. (F) Average cell volume for each cell type as a function of time; one curve for each of the 20 simulations. (G) Box-plot of the distribution of clonal populations size for $MCS \in [0, 10, 20, ..., 2000]$ with respect to the example simulation in panels A-D. (H) Distribution of cell duplication time in the example simulation displayed in panels A-D; red histograms correspond to *type 1*, blue histograms correspond to *type 2*. Transparency is used to make both series visible: when bars overlap a darker color is displayed. (I) Average cell biomass for each cell type as a function of time; one curve for each of the 20 simulations.

of the lattice are shown in Figure 2[A]. It is clear that the square shape is a strong simplification, yet the energy minimization criterion that underlies the CPM simulation ensures that cells reach a rounded and more physically sound shape in a few MCSs (despite some possible and expected defects in cell boundaries).

For each scenario, we executed 20 distinct simulations with random initial configurations. At the initial condition ($MCS = 0$), 620 cells with an individual area of 25 lattice sites (5×5) and a biomass of 1250 pg are disposed on the lattice. Half of the cells are assigned to *type 1* and the other half to *type 2*, and are alternatively disposed on the lattice (i.e., both the left and the right neighbours of a cell are of a distinct type).

Notice that in this scenario we do not consider the cell sub-populations as different with respect to DAH, meaning that cells will tend to fill the empty surrounding space, if available, but will show no preference in staying close to cell of the same type or different (see Table 1 – [SC 1]).

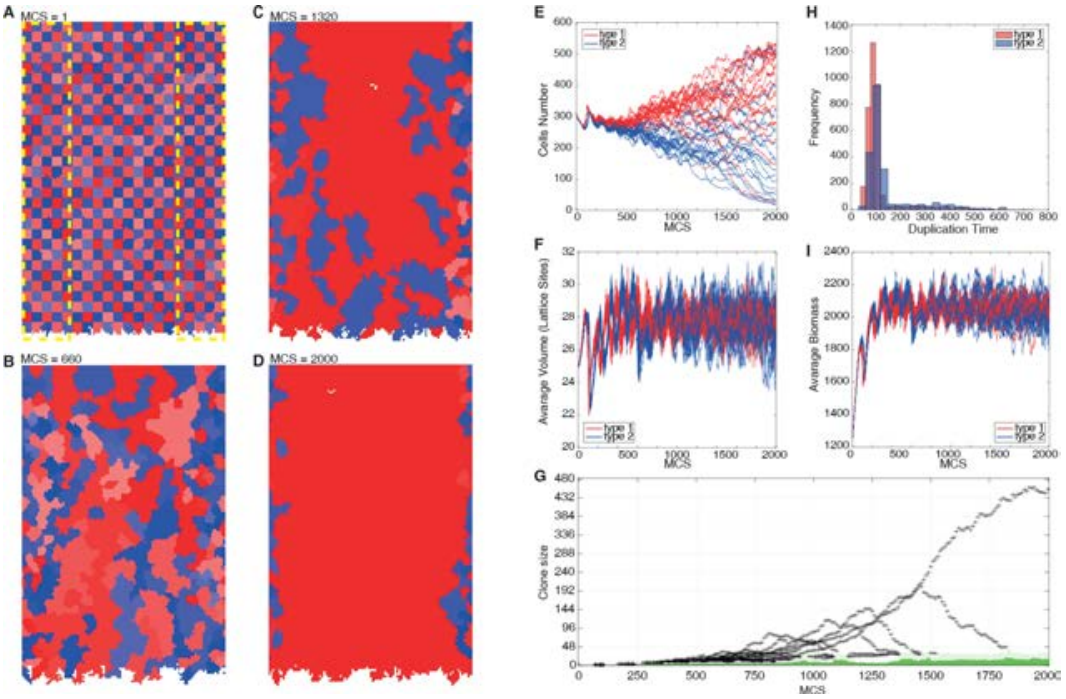


FIGURE 3
[SC 1 - B]. (A-D) Snapshots of FBCA dynamics respectively at 1, 660, 1320 and 2000 MCSs of an example simulation; shapes with red tonalities indicate *type 1* cells, shapes with blue tonalities indicate *type 2* cells; identical color refers to the same clonal population. The yellow dashed boxes at the left/right sides of the lattice indicates the hypoxic areas. (E) Total number of cells of each type as a function of time; one curve for each of the 20 simulations. (F) Average cell volume for each cell type as a function of time; one curve for each of the 20 simulations. (G) Box-plot of the distribution of clonal populations size for $MCS \in [0, 10, 20, ..., 2000]$ with respect to the example simulation in panels A-D. (H) Distribution of cell duplication time in the example simulation displayed in panels A-D; red histograms correspond to *type 1*, blue histograms correspond to *type 2*. Transparency is used to make both series visible: when bars overlap a darker color is displayed. (I) Average cell biomass for each cell type as a function of time; one curve for each of the 20 simulations.

Scenario 2 [SC 2] - Supbpopulations with distinct metabolic fitness

During cancer development, different cell sub-populations compete for limited resources, such as space and nutrients, and are positively selected in relation to their overall fitness, measurable in terms of survival and reproduction rate. In this scenario, we aim at investigating the competition dynamics that emerges among cell sub-populations with markedly different metabolic properties, in an environment in which space is limited and nutrients are heterogeneously distributed.

To this end, we first selected three reactions involved in distinct key metabolic pathways, i.e., (1) pyruvate fermentation to lactate, (2) proline biosynthesis and (3) cholesterol biosynthesis.

We then defined three distinct metabolic models: in each model we set a limiting flux constraint for one of such reactions, leaving the other unchanged.

In particular, in the metabolic model of cell *type 1* – respectively *2* and *3* – we reduced the maximum flux of reactions (1) by 70% – respectively of reaction (2) by 100%, and of reaction (3) by 70%.

In Figure 4[A] one can see the ratio between biomass produced in each model and that generated without flux limitations, with respect to different

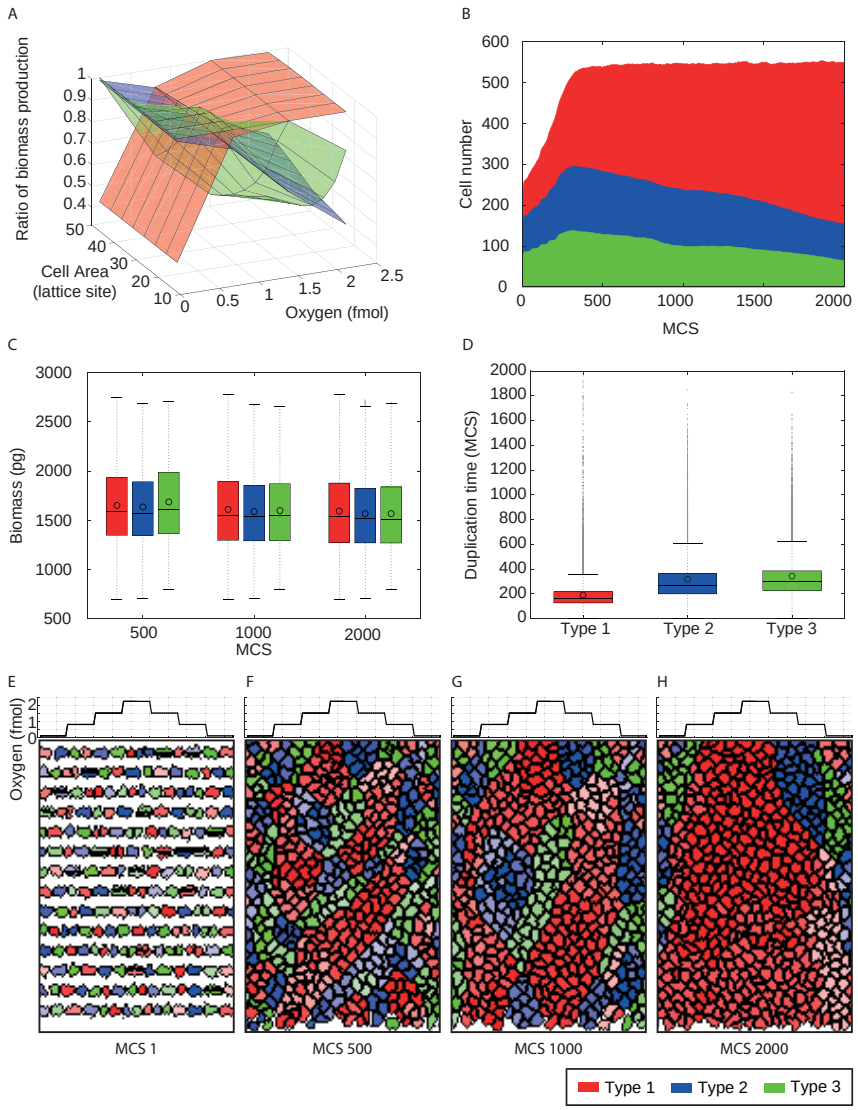


FIGURE 4
[SC 2 - A]. (A) Ratio among the biomass produced by the distinct metabolic models of scenario [SC 2] and that generated with metabolic models without limitations, with respect to different combination of oxygen and cell size. (B) Variation of the average population size of the three metabolic sub-populations described in Section 3, as computed over 20 distinct simulations. (C) Box-plot of the distribution of the biomass with respect to the three metabolic sub-populations, computed at different time points: $MCS = 500, 1000, 2000$. (D) Distribution of the duplication times recorded for all cells across the whole simulation, in all simulation runs. (E-H) Snapshots of FBCA dynamics respectively at 1, 500, 1000 and 2000 MCSs of an example simulation. Shapes with red, blue and green tonalities indicate *type 1*, *type 2* and *type 3* cell types, respectively; identical color refers to the same clonal population.

values of oxygen and cell size. In detail, the biomass production of *type 1* cell will be limited in hypoxic zones, where the production of lactate is advantageous. Instead, the activity of *type 2* cells will be reduced in zones with high oxygen level, as the limited pathway is involved in the biosynthesis of metabolites that are used in biomass production (i.e., proline). Finally, *type 3* cells have similar properties to *type 2* cells, as the ratio of biomass production decreases when oxygen concentration increases. However, unlike *type 2* cells, the biomass production of *type 3* cells also decreases in accordance to the cell size, i.e., the metabolic activity of small cells is enhanced with respect to larger ones.

In this scenario, cells are assigned with a random initial area in the interval [25, 50] (lattice sites), and a corresponding value of biomass such that $\rho = \frac{1}{\mathcal{F}}$. Moreover, we fill only half of the lattice, leaving empty space among the cells in horizontal stripes, as one can see in Figure 4[E], in order to reduce the competition pressure for space. Thus, at the initial condition ($MCS = 0$), the lattice will include around 250 cells of different dimensions.

Moreover, we considered the distinct cell sub-populations as different with respect to the DAH, i.e., cells will first tend to fill the surrounding empty space, if available, and, otherwise, will tend to stay closer to cells of the same type (see Table 1 – [SC 2])

Notice also that in scenario [SC 1], biomass could accumulate in each cell independently from the actual cell's density ($\gamma = 1$). This assumption may lead to an unrestrained increase in cell density $\rho(\sigma_i)$. Therefore, following [17] we introduced a rudimentary sensing mechanism for cell population density, as follows. The biomass accumulated at each time step $\mathcal{B}(\sigma_i)^{MCS}$ is reduced by the following factor:

$$\gamma = \begin{cases} 1, & \text{if } \rho(\sigma_i) \leq \dfrac{1}{\mathcal{F}} \\ 1 - \min(1, 2(\mathcal{F} \cdot \rho(\sigma_i) - 1))^2, & \text{otherwise} \end{cases}$$

In this way, biomass cannot accumulate ($\mathcal{B}(\sigma_i)^{MCS} = \mathcal{B}(\sigma_i)^{MCS-1}$) if the cell density $\rho(\sigma_i)$ exceeds its initial value by 1.5 times.

We simulated the dynamics of the system in two distinct conditions.

[**A**] In this first scenario, glucose and glutamine are uniformly distributed across the lattice (see Table 1 – [SC 2]); also lactate is uniformly distributed across the lattice, but its total amount is influenced by the consumption/secretion of each cell, and the concentration at the initial condition is set equal to 0 fmol/lattice site. Instead, oxygen is distributed with four different concentration levels that are progressively

increased from the borders toward the center of the lattice, as shown in Figure 4 [E-H]

[**B**] The distribution of nutrients is identical to case [SC 2 - A]. However, in this setting we simulated a basic therapy that is able to quickly wipe-out all the cells of the dominant clone, in order to analyze the possible recurrence of cancer. More in detail, at time point $MCS = 2000$ we set the death rate for cells of the most abundant cell type equal to 0.5 (i.e., at each MCS, any given cell of that type has a probability of 0.5 of dying). In a few MCSs the corresponding cell sub-populations is expectedly wiped-out from the lattice, leaving space and nutrients for the other ones.

4 RESULTS

[SC 1 - A] Proliferative cells with fermentative metabolism colonize the space.

With respect to the scenario [SC 1 - A], we simulated the dynamics of the system for 2000 MCS (= 200 hours). In Figure 2[A-D], screenshots of the lattice of an example simulation are displayed in four distinct moments, i.e., MCS = 1, 660, 1320 and at the end of the simulation (MCS = 2000). Both *type 1* (red tonalities) and *type 2* (blue tonalities) cells are highly proliferative, but at slightly different rates: *type 1* produces 11% more biomass than *type 2*. Notice that all daughter cells maintain the same color of the parent, so each clone is characterized by a unique and identifiable color.

A visible result is that CPM simulation ensures that cells reach a cell-like shape in a few MCSs. It is also apparent that, after a certain transient, a few clones tend to colonize space in vertical stripes, and such pattern remarkably reproduces the complex phenomenon of vertical cell migration in real-world intestinal crypts [14].

Surprisingly, although in this scenario nutrients are homogeneously distributed, and in spite of the minor advantage of *type 1* in terms of biomass growth rate, *type 2* cells tend to colonize all space after a period of time, in all simulation runs. It can indeed observed in Figure 2[E] that the number of cells of *type 1* constantly decrease, in all 20 runs, approaching values close to zero after around 1500 MCSs. This phenomenon, which may depend on the properties of the tissue and on the competition for limited space, deserves further investigations.

As opposed to standard FBA analysis, our model allows to compute the duplication time, i.e., the number of MCSs passed before mitosis for any given cell. The histogram in Figure 2[H] shows the distribution of the duplication times recorded for any cells in the lattice over 2000 MCSs, for the example simulation run displayed in panels A-D. One can see that the

distribution of *type 2* cells is slightly shifted to the right, meaning that, on average, such cells tend to duplicate at a slower pace.

With FBCA it is also possible to analyze the variation of the clonal population size distribution (i.e., the number of cells generated from a unique ancestor cell) in time. For instance, in Figure 2[G] one can see that larger clones are expectedly emerging during the example simulation displayed in panel A-D, yet reaching a median value around 60 at the end of the simulation: starting from the initial condition, in which 620 clones of size 1 are present on the lattice, at the end of the simulation (MCS = 2000) around 10 distinct clones (on average) are left.

[SC 1 - B] Competition for space in heterogeneous nutrients environments.

In this scenario, the population dynamics becomes more complex, as an heterogeneous distribution of nutrients is mimicked by introducing the hypoxic areas depicted in Figure 3[A].

In Figure 3[A-D], one can see that *type 1* cells (red) tend to migrate towards and occupy the highly-oxygenated area, whereas *type 2* cells (blue), which are allowed to have a fermentative metabolism succeed to proliferate in hypoxic areas.

Remarkably, two distinct dynamical behaviours emerge in different simulation runs: i) colonization by one cell type, ii) coexistence of both cell types (i.e., *homeostasis*). In the former case, one cell type ends up in colonizing the lattice; as one can see from Figure 3[E], in most cases *type 1* cells dominate, but in a relevant number of cases *type 2* tend to colonize the space. Complete extinction of a cell type was never observed, but this is most likely due to the limited simulation time. More interestingly, in a certain number of cases, the two cell populations reach a dynamical equilibrium (i.e., homeostasis), in which they both coexist in a stable proportion during the simulation time, despite distinct metabolic properties, different growth rates and the heterogeneous distribution of nutrients.

Notice also that the distribution of the observed duplication time displays a long right tail, likely due to the non-homogeneity of nutrients, i.e., cells in hypoxic areas tend to grow at a slower pace.

Finally, it is interesting to notice from Figure 3[G] that in the simulation displayed in panels A-D certain clones tend to cyclically dominate the lattice (outliers in the box-plots), until a huge clone consisting of around 430 cells emerges at the end of the simulation.

[SC 2 - A] Metabolic fitness determines crypt colonization.

The goal of this analysis is to determine whether limitations in the activity of selected metabolic reactions can influence the emerging population

dynamics, especially in presence of limited space and heterogeneous distribution of nutrients.

In Figure 4[B] one can see the average variation of the cell population size in time. *Type 1* sub-population dominates the overall dynamics in all simulations, proving a remarkable selective advantage based on its metabolic properties.

The biomass production of *type 1* cells is indeed limited when there is shortage of oxygen. As in this case the hypoxic zones are positioned at the lateral sides of the lattice, *type 1* cells tend to colonize the central part of the lattice, which is larger and richer in oxygen. Conversely, as the growth of *type 2* and *3* cells is limited in presence of high oxygen, such cells tend to displace along the hypoxic regions of the lattice, where they compete for the same limited space and resources.

It is interesting to notice that the growth limitation constraint related to cell size that characterizes *type 3* cells does not influence the overall population dynamics, as *type 2* and *3* sub-populations display very similar variations in time.

Notice also that the selective advantage of type 1 cell is not reflected in the variation of the average biomass distribution, which is quite similar for all the sub-populations across the simulations (see Figure 4[C]); it is, instead, due to a faster pace in duplication events 4[D].

This interesting results prove that metabolic fitness can be evaluated only with respect to environmental conditions, and especially to the spatial constraints and the availability of nutrients of the considered tissue.

[SC 2 - B] Resistant subclones determine cancer recurrence.
Intra-tumor heterogeneity is one of the major causes of therapy failure and tumour relapse, as small resistant cancer subclones may emerge and expand once a certain therapy has successfully killed the dominant – and usually detectable – subclone. In this analysis we aim at investigating the emergent behaviour of a crypt populated by metabolically different sub-populations (taken from [SC 2 - A] simulations), in which an effective therapy gets rid of the dominant subclone only.

In all simulations from scenario [SC 2 - A], *type 1* sub-population emerged as dominant, so once targeted by the therapy at $MCS = 2000$, all *type 1* cells are quickly wiped-out from the lattice, leaving space and nutrients for *type 2* and *3* sub-populations (see Figure 5[B,F]).

It is interesting to notice that, in distict simulations, either *type 2* or *type 3* sub-populations end up in colonizing the crypt (see Figure 5[A-D, I] and [E-H, L] for two example simulations, respectively, and Figure 6 [A], for a summary of all simulations). In this respect, one may suppose that, despite the difference in metabolic properties of the two cell types, the underlying

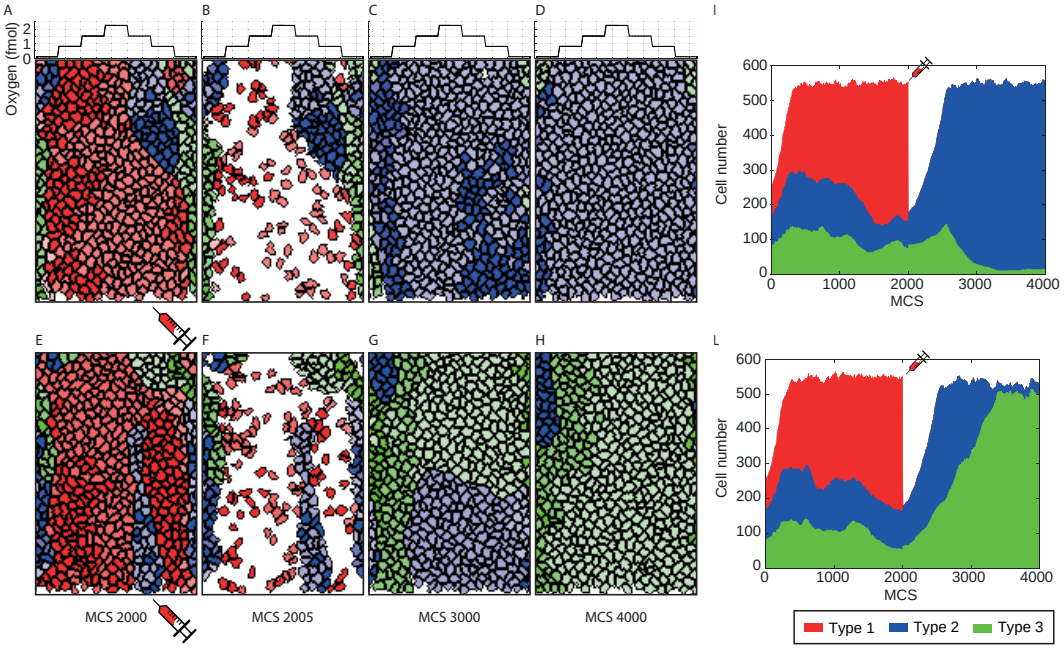


FIGURE 5
[SC 2 - B](A-D) and (E-H) Snapshots of FBCA dynamics at 2000, 2005, 3000 and 4000 MCSs of two example simulations (first and second row, respectively). The simulations start from the last MCS of randomly selected simulations presented in Figure 5, and model a therapy occurring at the first time step ($MCS = 2000$). Shapes with red, blue and green tonalities indicate *type 1*, *type 2* and *type 3* cell types, respectively; identical color refers to the same clonal population. (I-L) Variation of the cell number of the distinct sub-populations in time, with respect to the two example simulations. In the first row, as a consequence of the therapy, which quickly eliminates all *type 1* cells from the lattice, *type 2* sub-population colonizes the crypt, after a certain transient. In the second row, *type 3* sub-population dominates.

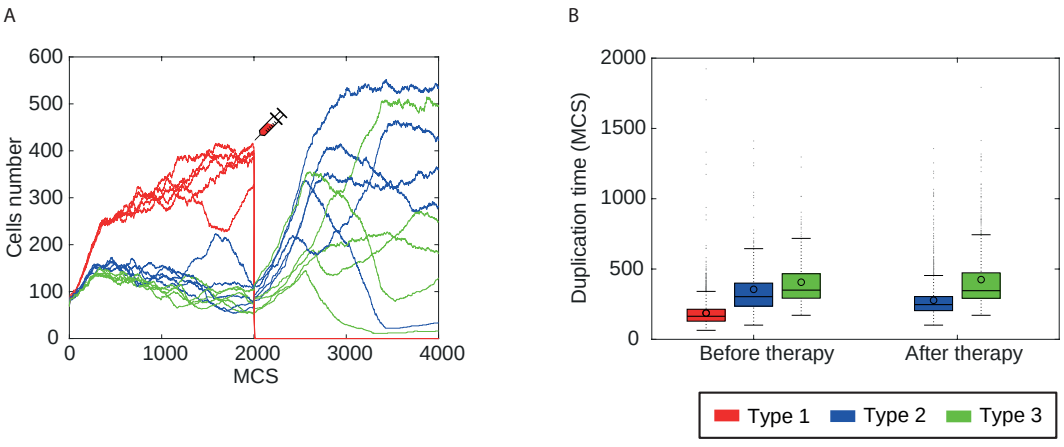


FIGURE 6
[SC 2 - B] (A) Variation of the cell number of the distinct cell sub-populations, with respect to 5 distinct simulation runs (including the two simulations presented in Figure 5). (B) Box-plots depicting the distribution of the duplication times recorded before and after the therapy, with respect to all the simulations.

stochasticity of the system dynamics is indeed largely responsible for the final fate of the system. In both scenario, the crypt is colonized in around $500 MCSs$ (around 50 hours), quickly filling the space left empty by *type 1* cells.

Also, one can notice that the distributions of the duplication times recorded before and after the therapy are quite similar for *type 3* cells, whereas *type 2* cells display a significant faster duplication pace. This phenomenon is likely due to the biomass growth limitation related to cell size, which characterizes *type 3* cells, preventing a more noteworthy increment in the duplication rate.

To conclude, we here proved that our modeling approach can link the metabolic properties of cancer cells with the complex emerging behaviour of a biologically realistic tissue, hence allowing to investigate the behaviour of possibly undetected drug-resistant subclones, which may be responsible for cancer recurrence. Such aspects could not have been tackled by focusing on the individual dynamics of metabolic networks only, and require effective frameworks to integrate distinct biological levels and phenomena, including the effects of stochasticity.

5 DISCUSSION

FBCA is a multiscale modeling framework that combines a biophysically plausible representation of cell morphology and interactions, via Cellular Potts Model, and a model of cellular metabolic activity, via Flux Balance Analysis.

Despite the abstractions underlying both modeling approaches, we proved that FBCA can reproduce complex phenomena observed in real world biological systems, such as cell migration, tissue colonization and homeostasis, allowing to perform in-depth quantitative analyses of key properties of cell populations in a variety of simulated experimental settings. In particular, the encouraging results on cancer recurrence modeling could be further investigated.

One the one hand, it could be possible to integrate our approach with cancer evolution models derived from genomic data as proposed, e.g., in [5, 22]. For instance, one might characterize the cancer sub-populations resulting from the distinct evolutionary trajectories with specific metabolic models. By integrating such models with a detailed morphological characterization of tissues and organs, it will be possible to deliver predictive models of cancer spatial evolution, in which sub-populations are driven by specific metabolic properties and interactions.

On the other hand, our framework allows to effectively test realistic metabolic therapies in silico, by targeting specific reactions within the

metabolic networks of the single cells, and by simulating the emerging population dynamics in realistic experimental scenarios. This will provide a powerful automated instrument for metabolic therapy design and testing, and represents one of the major advantages of employing our modeling approach.

Many extensions of FBCA are underway in order to model more biologically realistic processes and phenomena. For instance, metabolic communication among cells can be easily modeled with FBCA, by allowing the metabolites secreted by cells to diffuse over the tissue. In addition, the diffusion of nutrients via spacial gradients can be introduced in FBCA, and this will allow to explore scenarios that might be experimentally validated, e.g., in cell culture. Besides, it will be possible to characterize the features and properties of each cell by employing the increasingly available *single-cell* -omics data, as proposed for instance in [11].

The overall approach has a remarkable potential in several distinct application domains, ranging from cancer research to metabolic engineering. For instance, FBCA might be used to simulate the impact on tissue morphology of mutations in metabolic genes accumulating through successive clonal expansions.

Efforts to speed up the execution time are ongoing, focused on the parallelization of the CPM computation and distribution of FBA computation.

ACKNOWLEDGMENTS

The institutional financial support to SYSBIO - within the Italian Roadmap for ESFRI Research Infrastructures - is gratefully acknowledged. CD received funding from FLAG-ERA grant ITFoC.

REFERENCES

[1] Alex Graudenzi, Davide Maspero, and Chiara Damiani. (2018). Modeling spatio-temporal dynamics of metabolic networks with cellular automata and constraint-based methods. In Giancarlo Mauri, Samira El Yacoubi, Alberto Dennunzio, Katsuhiro Nishinari, and Luca Manzoni, editors, *Cellular Automata*, 16–29, Cham. Springer International Publishing.

[2] Matthew Bjerknes. (1996). Expansion of mutant stem cell populations in the human colon. *Journal of theoretical biology*, 178(4):381–385.

[3] Bruce M Boman, Jeremy Z Fields, Oliver Bonham-Carter, and Olaf A Runquist. (2001). Computer modeling implicates stem cell overproduction in colon cancer initiation. *Cancer research*, 61(23):8408–8411.

[4] Peter Buske, Jörg Galle, Nick Barker, Gabriela Aust, Hans Clevers, and Markus Loeffler. (2011). A comprehensive model of the spatio-temporal stem cell and tissue organisation in the intestinal crypt. *PLoS computational biology*, 7(1):e1001045.

[5] Giulio Caravagna, Alex Graudenzi, Daniele Ramazzotti, Rebeca Sanz-Pamplona, Luca De Sano, Giancarlo Mauri, Victor Moreno, Marco Antoniotti, and Bud Mishra. (2016). Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proceedings of the National Academy of Sciences*, 113(28):E4025–E4034.

[6] Paolo Cazzaniga, Chiara Damiani, Daniela Besozzi, Riccardo Colombo, Marco S Nobile, Daniela Gaglio, Dario Pescini, Sara Molinari, Giancarlo Mauri, Lilia Alberghina, *et al.* (2014). Computational strategies for a system-level understanding of metabolism. *Metabolites*, 4(4):1034–1087.

[7] Chiara Damiani, Riccardo Colombo, Daniela Gaglio, Fabrizia Mastroianni, Dario Pescini, Hans Victor Westerhoff, Giancarlo Mauri, Marco Vanoni, and Lilia Alberghina. (2017). A metabolic core model elucidates how enhanced utilization of glucose and glutamine, with enhanced glutamine-dependent lactate production, promotes cancer cell growth: The warburq effect. *PLOS Computational Biology*, 13(9):e1005758.

[8] Chiara Damiani, Davide Maspero, Marzia Di Filippo, Riccardo Colombo, Dario Pescini, Alex Graudenzi, Hans Victor Westerhoff, Lilia Alberghina, Marco Vanoni, and Giancarlo Mauri. (2019). Integration of single-cell rna-seq data into population models to characterize cancer metabolism. *PLoS computational biology*, 15(2):e1006733.

[9] Chiara Damiani, Marzia Di Filippo, Dario Pescini, Davide Maspero, Riccardo Colombo, and Giancarlo Mauri. (2017). popfba: tackling intratumour heterogeneity with flux balance analysis. *Bioinformatics*, 33(14):i311–i318.

[10] Chiara Damiani, Stuart A Kauffman, Roberto Serra, Marco Villani, and Annamaria Colacci. (2010). Information transfer among coupled random boolean networks. In *Lecture Notes in Computer Science. International Conference on Cellular Automata*, pages 1–11. Springer.

[11] Giovanni De Matteis, Alex Graudenzi, and Marco Antoniotti. (2013). A review of spatial computational models for multi-cellular systems, with regard to intestinal crypts and colorectal cancer development. *Journal of mathematical biology*, 66(7):1409–1462.

[12] Marzia Di Filippo, Riccardo Colombo, Chiara Damiani, Dario Pescini, Daniela Gaglio, Marco Vanoni, Lilia Alberghina, and Giancarlo Mauri. (2016). Zooming-in on cancer metabolic rewiring with tissue specific constraint-based models. *Computational biology and chemistry*, 62:60–69.

[13] François Graner and James A Glazier. (1992). Simulation of biological cell sorting using a two-dimensional extended potts model. *Physical review letters*, 69(13):2013.

[14] Alex Graudenzi, Giulio Caravagna, Giovanni De Matteis, and Marco Antoniotti. (2014). Investigating the relation between stochastic differentiation, homeostasis and clonal expansion in intestinal crypts via multiscale modeling. *PLoS One*, 9(5):e97272.

[15] Douglas Hanahan and Robert A Weinberg. (2011). Hallmarks of cancer: the next generation. *cell*, 144(5):646–674.

[16] Ruchir A Khandelwal, Brett G Olivier, Wilfred FM Röling, Bas Teusink, and Frank J Bruggeman. (2013). Community flux balance analysis for microbial consortia at balanced growth. *PloS one*, 8(5):e64567.

[17] Davide Maspero, Alex Graudenzi, Satwinder Singh, Dario Pescini, Giancarlo Mauri, Marco Antoniotti, and Chiara Damiani. (2019). Synchronization effects in a metabolism-driven model of multi-cellular system. In Stefano Cagnoni, Monica Mordonini, Riccardo Pecori, Andrea Roli, and Marco Villani, editors, *Artificial Life and Evolutionary Computation*, pages 115–126, Cham. Springer International Publishing.

[18] Philip J Murray, Alex Walter, Alexander G Fletcher, Carina M Edwards, Marcus J Tindall, and Philip K Maini. (2011). Comparing a discrete and continuum model of the intestinal crypt. *Physical biology*, 8(2):026011.

[19] Roberto Serra, Marco Villani, Chiara Damiani, Alex Graudenzi, and Annamaria Colacci. (2008). The diffusion of perturbations in a model of coupled random boolean networks. In Hiroshi Umeo, Shin Morishita, Katsuhiro Nishinari, Toshihiko Komatsuzaki, and Stefania Bandini, editors, *Cellular Automata*, pages 315–322, Berlin, Heidelberg. Springer Berlin Heidelberg.

[20] Jeffrey D Orth, Ines Thiele, and Bernhard Ø Palsson. (2010). What is flux balance analysis? *Nature biotechnology*, 28(3):245.

[21] Joe Pitt-Francis, Pras Pathmanathan, Miguel O Bernabeu, Rafel Bordas, Jonathan Cooper, Alexander G Fletcher, Gary R Mirams, Philip Murray, James M Osborne, Alex Walter, *et al.* (2009). Chaste: a test-driven approach to software development for biological modelling. *Computer Physics Communications*, 180(12):2452–2471.

[22] Daniele Ramazzotti, Giulio Caravagna, Loes Olde Loohuis, Alex Graudenzi, Ilya Korsunsky, Giancarlo Mauri, Marco Antoniotti, and Bud Mishra. (2015). Capri: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics*, 31(18):3016–3026.

[23] Marco Villani RobertoSerra and Chiara Damiani. (2008). The diffusion of perturbations in a model of coupled random boolean networks. In *Lecture Notes in Computer Science. Cellular Automata: 8th International Conference on Cellular Automata for Research and Industry, ACRI 2008, Yokohama, Japan, September 23-26, 2008, Proceedings*, volume 5191, page 315. Springer Science & Business Media.

[24] Simone Rubinacci, Alex Graudenzi, Giulio Caravagna, Giancarlo Mauri, James Osborne, Joe Pitt-Francis, and Marco Antoniotti. (2015). Cognac: a chaste plugin for the multiscale simulation of gene regulatory networks driving the spatial dynamics of tissues and cancer. *Cancer informatics*, 14:CIN–S19965.

[25] Jan Schellenberger, Richard Que, Ronan MT Fleming, Ines Thiele, Jeffrey D Orth, Adam M Feist, Daniel C Zielinski, Aarash Bordbar, Nathan E Lewis, Sorena Rahmanian, *et al.* (2011). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2. 0. *Nature protocols*, 6(9):1290–1307.

[26] Marco Scianna and Luigi Preziosi. (2013). *Cellular Potts Models: Multiscale Extensions and Biological Applications*. CRC Press.

[27] Abbas Shirinifard, J Scott Gens, Benjamin L Zaitlen, Nikodem J Popławski, Maciej Swat, and James A Glazier. (2009). 3d multi-cell simulation of tumor growth and angiogenesis. *PloS one*, 4(10):e7190.

[28] Malcolm S Steinberg. (1962). On the mechanism of tissue reconstruction by dissociated cells, i. population kinetics, differential adhesiveness, and the absence of directed migration. *Proceedings of the National Academy of Sciences*, 48(9):1577–1582.

[29] Milan JA Van Hoek and Roeland MH Merks. (2017). Emergence of microbial diversity due to cross-feeding interactions in a spatial model of gut microbial metabolism. *BMC systems biology*, 11(1):56.

[30] IMM Van Leeuwen, HM Byrne, OE Jensen, and JR King. (2006). Crypt dynamics and colorectal cancer: advances in mathematical modelling. *Cell proliferation*, 39(3):157–181.

[31] Joseph Walpole, Jason A Papin, and Shayn M Peirce. (2013). Multiscale computational models of complex biological systems. *Annual review of biomedical engineering*, 15:137–154.

[32] Patrick S Ward and Craig B Thompson. (2012). Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. *Cancer cell*, 21(3):297–308.

[33] Shek Yoon Wong, K-H Chiam, Chwee Teck Lim, and Paul Matsudaira. (2010). Computational model of cell positioning: directed and collective migration in the intestinal crypt epithelium. *Journal of The Royal Society Interface*, 7(Suppl 3):S351–S363.

Fundamenta Informaticae 171 (2020) 279–295
DOI 10.3233/FI-2020-1883
IOS Press

The Influence of Nutrients Diffusion on a Metabolism-driven Model of a Multi-cellular System

Davide Maspero[*], **Chiara Damiani**[†], **Marco Antoniotti,**[‡] **Alex Graudenzi**[§]
Department of Informatics Systems and Communication
Univ. of Milano-Bicocca, Milan, Italy

Giulio Caravagna
Data Scientist
Institute of Cancer Research
ICR, London, UK

Daniele Ramazzotti
Department of Pathology
Stanford University
Stanford, CA 94305, USA

Marzia Di Filippo, Marco Vanoni[£]
Department of Biotechnology and Biosciences
Univ. Milano-Bicocca, Milan, Italy

Riccardo Colombo
Department of Biomedical and Clinical Sciences "L. Sacco"
Univ. of Milan, Milan, Italy

Dario Pescini[C]
Department of Statistics and Quantitative Methods
Univ. of Milano-Bicocca, Milan, Italy
dario.pescini@unimib.it

Abstract. The metabolic processes related to the synthesis of the molecules needed for a new round of cell division underlie the complex behaviour of cell populations in multi-cellular systems, such as tissues and organs, whereas their deregulation can lead to pathological states, such as cancer. Even within genetically homogeneous populations, complex dynamics, such as population oscillations or the emergence of specific metabolic and/or proliferative patterns, may arise, and this aspect is highly amplified in systems characterized by extreme heterogeneity.

[*]Also affiliated at: Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[†]Also affiliated at: SYSBIO Centre of Systems Biology, Univ. of Milano-Bicocca, Milan, Italy
[‡]Also affiliated at: NeuroMI Milan Center for Neuroscience, Univ. of Milano-Bicocca, Milan, Italy
[§]Also affiliated at: Institute of Molecular Bioimaging and Physiology, Italian National Research Council, Milan, Italy.
[£]Also affiliated at: SYSBIO Centre of Systems Biology, Univ. of Milano-Bicocca, Milan, Italy.
[C]Address for correspondence: Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi, 8 - 20126 Milan.

To investigate the conditions and mechanisms that link metabolic processes to cell population dynamics, we here employ a previously introduced multi-scale model of multi-cellular system, named FBCA (Flux Balance Analysis with Cellular Automata), which couples biomass accumulation, simulated via Flux Balance Analysis of a metabolic network, with the simulation of population and spatial dynamics via Cellular Potts Models.
In this work, we investigate the influence that different modes of nutrients diffusion within the system may have on the emerging behaviour of cell populations. In our model, metabolic communication among cells is allowed by letting secreted metabolites to diffuse over the lattice, in addition to diffusion of nutrients from given sources. The inclusion of the diffusion processes in the model proved its effectiveness in characterizing plausible biological scenarios.

Keywords: Multi-scale modeling, Cellular Potts Model, Flux Balance Analysis, Diffusion, Cancer development

1. Introduction

Increasing experimental evidences are suggesting that the deregulation of metabolism is one of the key actors in tumor origination and development [1, 2]. In this respect, most current computational strategies to investigate metabolic deregulation are based on the simulation of the steady state behavior of an average cell belonging to a (heterogeneous) population [3, 4].

Yet, cancer (sub)population evolve and compete in a (micro)environment with usually limited resource (e.g., oxygen and nutrients) and with specific spatial properties, which significantly differs in distinct tissues and organs [5]. For this reason, properties based on average measurements may be scarcely significant, as they are not representative of the specific features of single cells and subpopulations, as well as of their interactions with the surrounding environment [6].

This aspect has important translational repercussions, as the intra-tumor heterogeneity resulting from such complex interplay is one of the major causes of drug resistance, therapy failure and relapse [7, 8, 9]. Unfortunately, so far the emerging single-cell omics technologies are still unable to finely characterize the interactions among cell (sub)populations, especially because of many technical issues that compromise the overall resolution [10, 11].

For this reason, effective computational multi-scale models are increasingly needed to account for processes and phenomena occurring at different time/ space scales and involving populations of interacting cells, with the final goal of identifying the conditions that may lead to complex behaviours, such as tissue patterning, cellular migration, homeostasis, and the emergence of pathological states [12, 13]. In this respect, many attempts have been proposed to simulate the spatial/morphological dynamics of multi-cellular systems, by employing biophysically plausible models of cells and their interactions, e.g., within tissues and organs [14, 15, 16]. In general, investigating the emerging spatio-temporal behaviour of the heterogeneous populations may facilitate the development of more effective intervention strategies [17].

In [18], we developed FBCA (Flux Balance with Cellular Automata) as first attempt to connect the dynamical behavior of metabolic networks to biophysically realistic spatial and morphological properties of real multi-cellular systems. This model combines a spatial/morphological dynamical model of a general tissue, simulated via Cellular Potts Model (CPM) [19, 20, 21, 22], and a lower-level model of the metabolic activity of its constituting single cells, computed via Flux Balance Analysis (FBA) [3].

In [23], we extended the previous framework by adding a population density-sensing mechanism, that led to more realistic results.

In this work, we further extended the methodology to model metabolites diffusion through space. In particular, we modeled metabolic communication among cells by allowing metabolites (e.g., lactate secreted by cells) to diffuse over the tissue according to a local spatial gradient. The final goal of this study is to evaluate the impact of different models of nutrient diffusion on the overall cell population dynamics.

2. Methods

In order to model a generic multi-cellular system, such as a cell culture or a tissue, we adopted FBCA, which we previously introduced in [18, 23]. This computational methodology combines through a multi-scale model the spatial dynamics representation of the system via CPM [24], with a model of cell metabolism via FBA [3] (see Figure 1).

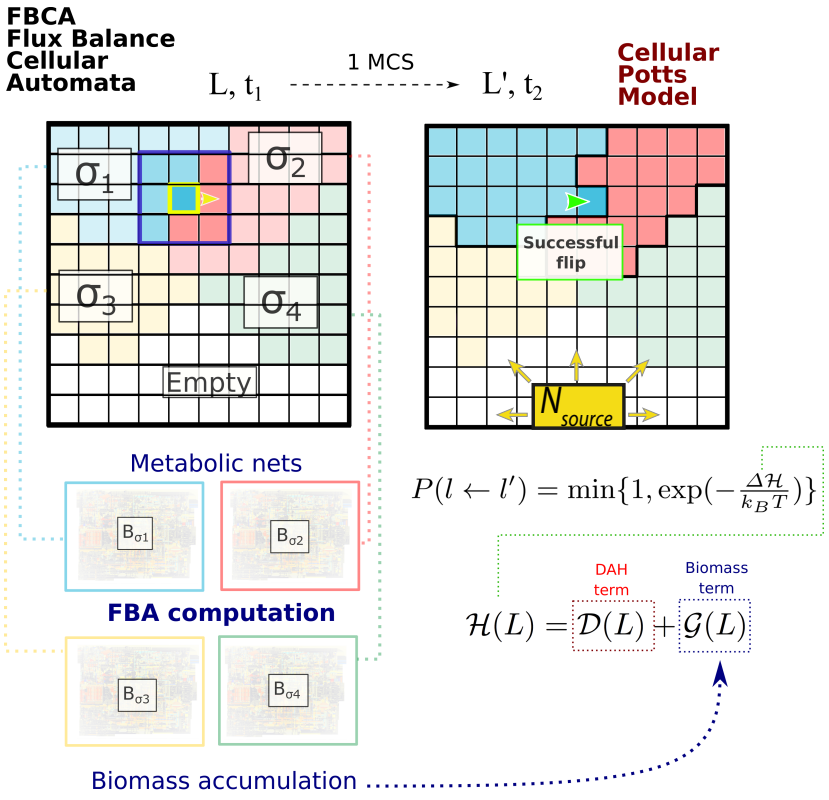


Figure 1. FBCAmodel is depicted. Image modified from [18]. The morphology of a generic tissue is modeled via Cellular Potts Model, in which biological cells are represented by sets of contiguous lattice sites, which evolve via flip attempts driven by a Hamiltonian function $\mathcal{H}(L)$. $\mathcal{H}(L)$ depends on two terms, the first accounting for the Differential Adhesion Hypothesis (DAH), the second for the growth tendency of each cell due to the accumulation of biomass (see Methods). Biomass is computed via Flux Balance Analysis on cell-specific metabolic networks. Nutrients N are diffused, from different sources, to sustain the cellular growth.

2.1. Simulation approach

According to CPM representation of multi-cellular system, biological cells are represented on a two dimensional lattice L, and their spatial dynamics is driven by an energy minimization criterion ruled by a Hamiltonian function $\mathcal{H}(L)$. A rectangular and rigid grid composed of $h \times w$ square sites $l^i \in L$ is used to represent the lattice. Given a set of cells tags $\mathcal{C} = \{c_c\}$ represented in the lattice L, where $c_c \in \mathbb{N}$, and a set of cellular types $\mathcal{T} = \{t_t\}$, where $t_t \in \mathbb{N}$, each biological cell is identified through the identifier c_c and associated to a specific cellular type t_t. The space occupied by cell c_c on the lattice is denoted as $C_c \subseteq L$ and corresponds to the set of contiguous lattice sites l^i associated to cell c_c. It is possible to represent on the lattice the empty space assigning to a lattice site the cellular type E (or the empty space).

To connect the information of lattice sites with the information of cells, it is possible to exploit $[N(l^i)]$, that is the concentration for a nutrient N at a specific lattice site l^i, together with the definition of two operators:

- $\sigma : L \to \mathcal{C}$ such that $\sigma(l^i) = c_c$. For short, we can say that $\sigma_i = c$, where c is the identifier of the cell that lies on lattice site l^i.
- $\tau : \mathcal{C} \to \mathcal{T}$ such that $\tau(\sigma_i) = \tau(\sigma(l^i)) = t_t$. For short, we can say that $\tau_i = t$, where t is the cellular type that is associated to the cell that lies on lattice site l^i.

The simulation
The simulation approach can be depicted as follows:

1. Initialize: L, $\mathcal{C}$, N
2. For each simulation step:
    - FBA computation
    - Biomass accumulation
    - Removal of cell death by lacking of nutrient
    - Minimization of the Hamiltonian function
    - Cell cycle phase evaluation
    - Nutrient diffusion

Each step of the second phase is detailed in the following.

FBA computation. FBA is exploited to compute the biomass production rate of any cell present in the lattice according to the corresponding nutrient availability. To this aim, a metabolic network model of human central carbon metabolism consisting of 272 reactions and 240 metabolites is associated to each cell c_c; this model was introduced in [25]. A metabolic network is defined as the set of metabolites and chemical reactions taking place in a cell, and it can be formalized as a $M \times R$ stoichiometric matrix S, where M is the number of metabolites and R is the number of reactions. Each element $s_{m,r}$ of the matrix S is the stoichiometric coefficient of metabolite m in reaction r, which expresses the number

of molecules of metabolite m that are transformed in reaction r. For any cell, uptake rate boundaries of extracellular nutrients are defined according to the simulated experimental setting. In general, the uptake rate of a given metabolite is constrained to be lower than the sum of the corresponding concentration values in all of the lattice sites in that cell closeness. For the sake of simulation, flux values of extracellular nutrients uptake reactions are assumed to be proportional to the concentration of the corresponding involved nutrient.

Given a metabolic network, Linear Programming is applied in FBA simulations to identify the flux distribution $\mathbf{v} = (v_1, \ldots, v_R)$ that maximizes or minimizes a specific objective function Z:

$$Z = \sum_{j=1}^{R} w_j v_j \tag{1}$$

where w_j is a coefficient that represents the contribution of flux v_j in vector $\mathbf{v}$ to Z, given a steady-state assumption for the abundance of each metabolite, i.e., $S\,\mathbf{v} = \mathbf{0}$. In this work, the Linear Programming Problem is solved at each *Monte-Carlo Step* (MCS) to maximize for each cell a pseudo-reaction that represents the conversion rate of biomass precursors into biomass, as follows:

$$\begin{aligned} &\text{maximize } \; v_b \\ &\text{subject to } S\,\mathbf{v} = \mathbf{0},\ \mathbf{v}_L \leq \mathbf{v} \leq \mathbf{v}_U \end{aligned} \tag{2}$$

where v_b is the biomass reaction flux of the cell c_c, whereas $\mathbf{v}_L$ and $\mathbf{v}_U$ represent the vectors containing, respectively, minimum and maximum allowed reaction fluxes.

Biomass accumulation. At each time step s, the current value of the biomass so far accumulated by a given cell c_c is represented by the map $\mathcal{B} : \mathbb{R} \times \text{time} \to \mathbb{R}$, that is $\mathcal{B}(\sigma_i, s) = \mathcal{B}(c, s)$. To update the current value for $\mathcal{B}(c, s)$, we add the biomass synthesis rate v_b obtained from the FBA simulation for the corresponding cell, to the biomass that is so far accumulated by the cell itself:

$$\mathcal{B}(c, s) = \mathcal{B}(c, s-1) + \gamma \cdot v_b \tag{3}$$

where γ is a dimensionless factor linking the contribution of cell density to biomass accumulation. $\mathcal{B}(c, s)$ is exploited to determine the corresponding *cell density* $\rho_c = \frac{\mathcal{B}(c,s)}{A_c}$.

In [18], we considered the scenario where, at each time step, a given cell produces biomass according only to the nutrient availability, disregarding other parameters and setting the factor γ equal to 1 to make the biomass accumulation independent of the current cellular density. However, in a more realistic scenario, proposed in [23], cells avoid an uncontrolled raising of their cell density ρ_c, we therefore modelled the limitation of biomass accumulation $\mathcal{B}(c, s)$ at each time step introducing a limiting factor $\varphi \in \mathbb{R}^+$ as follows:

$$\gamma = \begin{cases} 1, & \text{if } \rho_c \leq \varphi \\ 1 - \min(1, 2(\dfrac{\rho_c}{\varphi} - 1))^2, & \text{otherwise} \end{cases}$$

In this way, we prevent the increase of the biomass accumulation in case of ρ_c exceeding its initial value of 1.5 times.

Hamiltonian function. The Hamiltonian function $\mathcal{H}(L) : L \rightarrow \mathbb{R}$, is composed of two terms $\mathcal{H}(L) = \mathcal{D}(L) + \mathcal{G}(L)$ that, respectively, account for the Differential Adhesion Hypothesis (DAH) $\mathcal{D}(L)$ [26], and the growth tendency $\mathcal{G}(L)$ of each cell that is driven by biomass accumulation:

The first term $\mathcal{D}(L)$ is defined as follows:

$$\mathcal{D}(L) = \frac{1}{2} \sum_{i,j \in \mathcal{N}} \mathrm{J}(\tau_i, \tau_j)(1 - \delta_{\sigma_i, \sigma_j}) \tag{4}$$

where $\mathrm{J}(\tau_i, \tau_j)$ is the surface energy between cellular types τ_i and τ_j that is required to maintain adjacent two cell sites, $\delta_{\sigma_i, \sigma_j}$ is the Kronecker delta, and $\mathcal{N}$ is the Moore neighborhood.

The second term $\mathcal{G}(L)$ is defined as follows

$$\mathcal{G}(L) = \lambda \sum_{c \in \mathcal{C}} (A_c - \mathrm{A}_{\oplus}(\mathcal{B}(c, s)))^2 \tag{5}$$

where the plasticity coefficient $\lambda \in \mathbb{R}^+$ is a Lagrange multiplier that accounts for the capacity to deform a cell membrane, whereas A_c is defined as $A : L \rightarrow \mathbb{N}$ such that $A(\sigma(l^i)) = A(\sigma_i) = A(c_c) = A_c$ and corresponds to the current area of the cell c_c. The target area $\mathrm{A}_{\oplus}(\mathcal{B}(c, s))$ is function of the biomass $\mathcal{B}(c, s)$ that is accumulated by the cell c_c at each time step according to the current nutrient availability. The latter depends on the adopted diffusion process according to the considered environmental settings, as it will be discussed later in this Section in the paragraph Diffusion. The accumulated biomass $\mathcal{B}(c, s)$ is converted into the corresponding $\mathrm{A}_{\oplus}$ by using the conversion factor φ, $\mathrm{A}_{\oplus} = \frac{\mathcal{B}(c,s)}{\varphi}$, which is usually set to 50.

Cell cycle phase evaluation. We associate to each cell an initial equal area, which is called base area A_B, representing the corresponding initial area before a cell starts to grow. Each cell can grow until it reaches twice the initial A_B. After that, following the half splitting of the updated space along a randomly chosen horizontal or vertical direction, a daughter cell characterized by cell properties inherited from its parent is produced.

Diffusion. The nutrient diffusion process, driven by the resulting positive outcomes shown in [27], is here implemented averaging the nutrient concentrations in a neighborhood I of each lattice site l^i:

$$[N(l^i)] = \frac{D}{|I|} \sum_{j \in I} [N(l^j)] \tag{6}$$

where, the neighborhood I has different definition according to the biological setting to be mimicked, i.e. $I \coloneqq \mathcal{N} \cup l^i$ in case of permeable cells, or $I \coloneqq C_E \cap (\mathcal{N} \cup l^i)$ in case of impermeable cells. C_E refers to the lattice sites where nutrients cannot diffuse because they are occupied by impermeable cells. D is the diffusion coefficient chosen based on the nutrient species.

2.2. FBCA and diffusion

To investigate the ability of the proposed approach discussed in Section 2.1 to simulate the impact of diverse nutrient diffusion models on a cell population dynamics, we modelled two different biological scenarios corresponding, respectively, to a closed environment and to a tissue like environment.

Regarding the latter, we proposed two configurations relative to the geometry of the nutrient sources. These two configurations represent, respectively, a cross and a longitudinal section of the tissue like environment (see Figure 2 panels A and B). Finally, to evaluate different descriptions of nutrient diffusion across cellular membranes, namely cell membranes permeability, we modelled both cross and longitudinal section configurations by considering cells as permeable and impermeable.

Overall, all these investigated scenarios share the following properties: i) the lattice is a fully closed environment where the process of cellular death by starving is the only way to remove cells; ii) in [23], the limitation of biomass accumulation allowed to obtain outcomes that are close to the biological reality. For this reason, we maintain the assumption in this work; iii) metabolism of each cell is simulated by using a core model of human central carbon metabolism [25, 28, 29]; iv) in this study we consider one cellular type c and empty space E i.e. $\mathcal{T} = \{c, E\}$, with surface energies $\mathrm{J}(c, c) = 8$ and $\mathrm{J}(c, E) = 2$, mimicking the tendency to fill the empty space if available. Finally, simulation parameters are set as in [18], unless otherwise specified.

2.2.1. Nutrients distribution and geometry

The first simulated biological scenario corresponds to a closed environment. In this regard, we modelled a rectangular lattice space consisting of 150 $\times$ 100 sites. In this configuration, metabolites are uniformly distributed over the lattice and their concentration values are equally set in all the lattice sites. Furthermore, intercellular interactions only depend on nutrients exchange rather than on specific nutrient diffusion dynamics.

In the initialization phase, the lattice is populated with four cells having the same area. At each simulation time step: i) to reset a uniform nutrients distribution over the lattice, the amount of each nutrient at each lattice site, $[N(l^i)]$ is set to its mean value over the entire lattice; ii) nutrients uptake rate of each cell is set proportionally to its area A_c and constraints of each cell $\mathbf{v}_U, \mathbf{v}_L$ are set to $[N(l^i)] \cdot A_c$; iii) a FBA optimization is performed for each and every cell being in the lattice, and the secreted metabolites instantaneously diffuse; iv) finally, the nutrients amount in the lattice is updated. The simulation is halted before that cells saturate the closed environment (i.e. 1000 time steps) according to the biological inspiration.

In the second biological scenario, we conceived a rectangular lattice space of 175 $\times$ 115 sites to represent a tissue-like environment. Moreover, this lattice is completely closed to avoid the removal of cells from the system beyond the cell death. According to this space configuration, nutrient diffusion dynamics is considered and metabolites concentration in each lattice site $[N(l^i)]$ is set equal to the mean of the corresponding concentrations in its neighborhood according to Equation 6. In the initialization phase, the lattice is randomly populated with cells of different initial area.

Many projections are possible sections of a three dimensional space into a two dimensional one, but to pursue our intent to mimic realistic conditions of a tissue, we opted for the following two sections.

Cross section represents a transversal section of a generic biological tissue that consists of 5 square nutrient sources that are fixed placed within the lattice. These point sources represent blood vessels from which nutrients radially diffuse. As depicted in Figure 2A, vessels area is equal to 121 lattice sites per the top and bottom point sources, whereas it is equal to 81 for the three central ones.

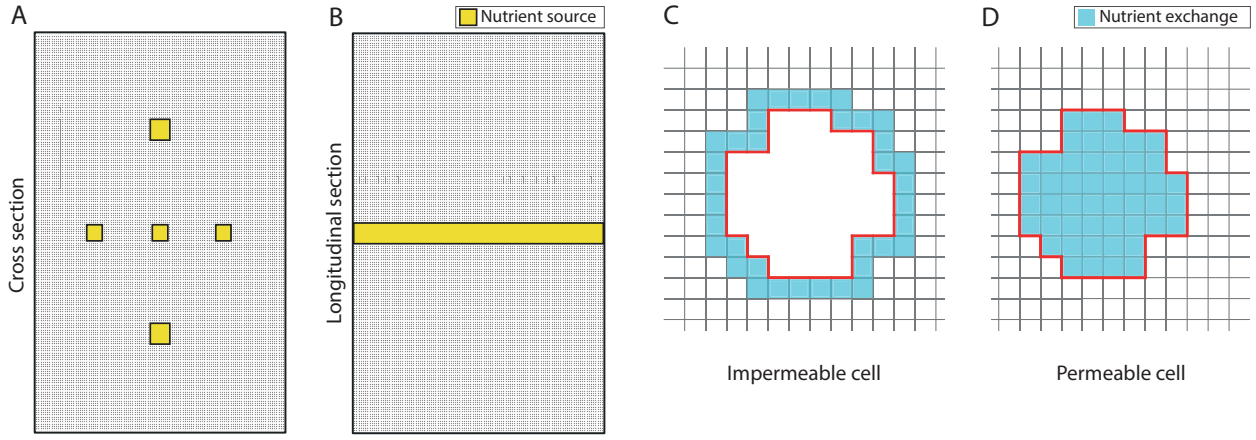


Figure 2. Schematic representation of the nutrient sources geometry and access. On the left: A) The tissue is represented from a cross section point of view. B) The tissue is represented from a longitudinal section point of view. Yellow rectangles and squares show the position and the area of the nutrient sources in the lattice. On the right: Cells have different nutrient availability (light blue surface) according to their permeability. C) Impermeable cells exchange nutrients with lattice sites of the empty space adjacent to cell membrane (red solid lines); D) permeable cells exchange nutrients with lattice sites within the cell membrane.

Longitudinal section represents a biological tissue traversed by a blood vessel. The blood vessel is here represented as a rectangle occupying all the central area of the lattice equal to 11×115 lattice sites. In this configuration, nutrients diffuse in the lattice with a linear gradient perpendicular to the vessel axis.

The geometry and the position of the nutrient source are the only elements discriminating the two tissue like configurations. At each diffusion step, each one of the considered nutrients concentration within lattice sites of the source area is set equal to a specific value, i.e., oxygen is equal to 100 fmol, glucose is equal to 50 fmol, and glutamine is equal to 50 fmol. The lactate is not supplied in the extracellular environment, but it may be just produced and then exchange from the cells. Once nutrients concentration values are updated in these lattice sites, the diffusion in the entire lattice occurs according to equation 6. If nutrients are not consumed by cells, they are removed from the simulated lattice through a constant flux value when the corresponding edges of the lattice itself are reached.

2.2.2. Cell membrane permeability

In this work, we devise a simulator suitable to investigate the interplay between cellular population dynamics and nutrient diffusion. In this framework two different length scales take place: the one of the nutrients that has dimensions of Å and the one of the cells that has length of μm. This, combined with the fact that we exploit lattices, bi-dimensional objects to represent cells, three-dimensional objects, suggested us to compare two different descriptions of nutrient diffusion across cellular membranes, permeable versus impermeable cells (see Figure 3).

Permeable cell. In this condition, cell and nutrients move on two distinct overlapping layers: the cell matrix, in which cells evolve, overlays the nutrient matrix, where metabolites freely diffuse disregarding the presence of the cells. In this configuration, cells have access to the nutrients and

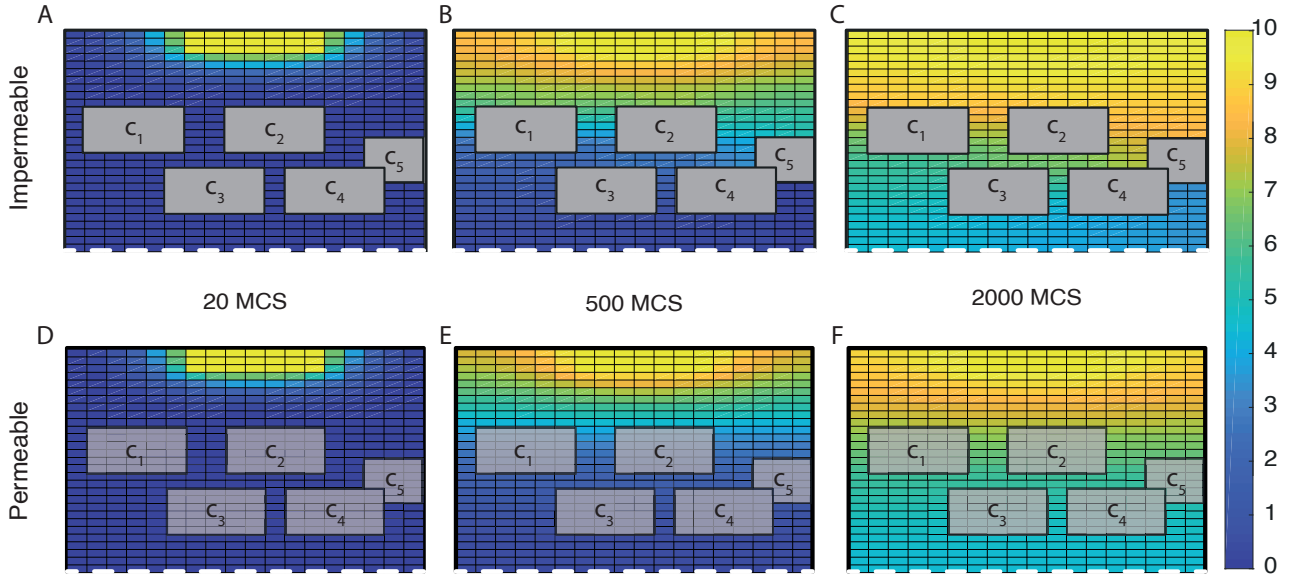


Figure 3. Nutrient diffusion simulation. In each frame the same environment is depicted with a lattice populated with five cells $(C_1, \ldots, C_5)$ where in each lattice site l^i the color correspond to the concentration of a generic nutrient $[N(l^i)]$ at that time step. In this dynamics there is no interaction between cells and nutrient but the space occupied by each cell. Comparing the upper part (impermeable setting) with the lower one (permeable setting), it is possible to recognize the shading effect, due to the cellular membrane, and the slower diffusion process due to the limited amount of space available to nutrients to the impermeable configuration.

metabolites of each lattice site over which they "float" (See Figure 2D). The set of lattice sites in which diffusion takes place $I = \mathcal{N} \cup l^i$ does not require that sites belong to the empty space. In the same way, the metabolites produced by each cell flow only into those lattice sites under the same cell surface C. Cells are completely permeable to nutrients and metabolites and the diffusion process takes place "independently" from the cells positions.

Impermeable cell. Cells and nutrients share a common layer, and nutrient diffusion is strongly affected by cells locations. In this scenario, cells are fully impermeable to nutrients. Consequently, nutrients cannot diffuse in the lattice sites that are occupied by cells (See Figure 2C). Set $I = C_E \cap (\mathcal{N} \cup l^i)$ adopted in Equation 6 requires to consider only lattice sites belonging to the empty space. In this configuration, cells only access to metabolites that are available in lattice sites adjacent to their external surface. In a similar manner, the metabolites produced by each cell flow only into lattice sites on the perimeter of the cell surface. However, in this way, two neighbouring cells that are separated by just a unique lattice site, share the set of nutrients in lattice sites that are adjacent to cellular external sides. To avoid the situation where a limiting nutrient concentration is consumed by both cells, we updated the corresponding nutrients concentration values after metabolic fluxes computation of each single cell. Moreover, to avoid bias, we randomized the order in which these optimizations were performed. Since nutrient diffusion and computation of the cellular dynamics are sequentially executed, if cell area increase includes a previously empty lattice site, nutrients concentration is shifted in that lattice site and then splitted among its neighborhood belonging to the set of empty sites.

2.3. Implementation

The scripts to perform all the analysis and the functions to simulate the spatial dynamics and the diffusion steps has been written from scratch in Matlab. We used the COBRA Toolbox[30] to optimize the metabolic networks and compute the biomass for each cell in the lattice. A complete MCS requires between ~ 3 s and ~ 6 s with a PC Intel Core i7-3770 CPU 3.40 GHz 64-bit capable, with 32 GB of RAM DDR3 1600 MT/s.

3. Results

3.1. Closed environment

As shown in Figure 4, when we modelled the closed environment setting, we observed that cells start to grow and to fill the empty available space. Nevertheless, cells dimension does not increases excessively because of the previously defined growth rules with a mitotic area set to 50.

Moreover, Figure 4 contains few snapshots of the metabolic history of the entire cells population suggesting a metabolic switch from glucose uptake to lactate consumption: cells turn from red to blue.

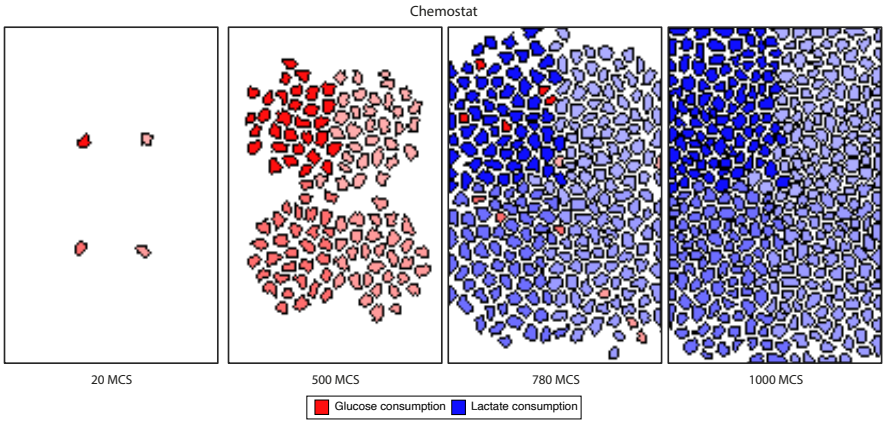


Figure 4. Cells population dynamics in the closed environmental setting. In this configuration, metabolites are uniformly distributed over the lattice and their concentration values are equally set in all the lattice sites. Intercellular interactions only depend on nutrients exchange rather than on a specific nutrient diffusion dynamics. The color assigned to each cell depends on its emerging metabolic trait. Red cells depend on glucose utilization, while the blue ones on lactate utilization. Color shades are intended to distinguish among cells and are not informative about fluxes values.

But, it is in Figure 5 that is possible to account for this phenomenon, at the beginning glucose and glutamine represent the main nutrients adopted from cells to grow, even if the initial consumed concentration of glucose is higher than that of glutamine. Moreover, from the red curve in Figure 5 A, that corresponds to the available amount of glucose over time and space, we observed that cell consumption of this nutrient does not follow a linear trend due to the initial low number of cells that are present in the lattice.

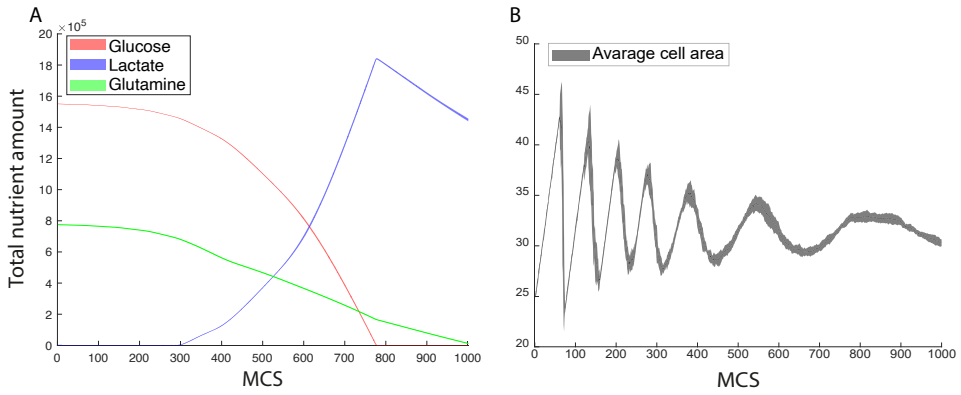


Figure 5. Nutrient dynamics and average area in the closed environment setting. A) The graph depicts the mean nutrients amount (solid line) and its standard deviation (shaded area) over 10 independent simulations. B) The graph depicts the mean cell area (solid line) and its standard deviation (shaded area) over 10 independent simulations.

These two nutrient sources are gradually consumed over MCS time step and lactate is produced. However, at the point where there is no more available glucose to grow, we observed that cells start to use lactate as alternative nutrient source. Accordingly, in the second and third panel of Figure 4, we observed a rapid switch of cellular metabolism from cells mainly growing on glucose to cells whose growth is dependent on the consumption of lactate. This rapid switch can occur because of the identical access of cells to nutrients. In line with the description of closed environment setting, nutrients access only changes among cells according to their dimension.

In Figure 5B, we plotted the average cell area over MCS. The starting setting of 4 cells having the same area showed very well the cellular growth process due to the initial sharp increase of cell average area. Moreover, we can appreciate how long it takes a cell to double, i.e., the mitosis process. As cells continue to grow, we observed that their number increases, and differences begins to emerge among them due to the no longer equal size. Accordingly, the average cell area results to be the mean of a more extended range of values.

3.2. Tissue like environment

In the tissue-like model, we consider that all cells can be either permeable or impermeable to nutrients. Our simulations consider either a longitudinal or a cross-section of a 3D lattice, therefore originating 4 independent simulation scenarios (Figure 6).

In all scenarios, nutrients are injected in the system only within a specialized lattice subset that we refer to as “vessel”, they mix only locally and cells can move and die (because of nutrient depletion). Since their survival chances collapses as they move away from nutrient availability, cells tend to stay close to the vessels in all four simulation scenarios over the course of the entire simulation (Figure 7).

Snapshots of the lattices at the end of simulations are shown in Figure 6. We assigned cells in the population to three groups according to their way of metabolizing lactate, taken as an indication of

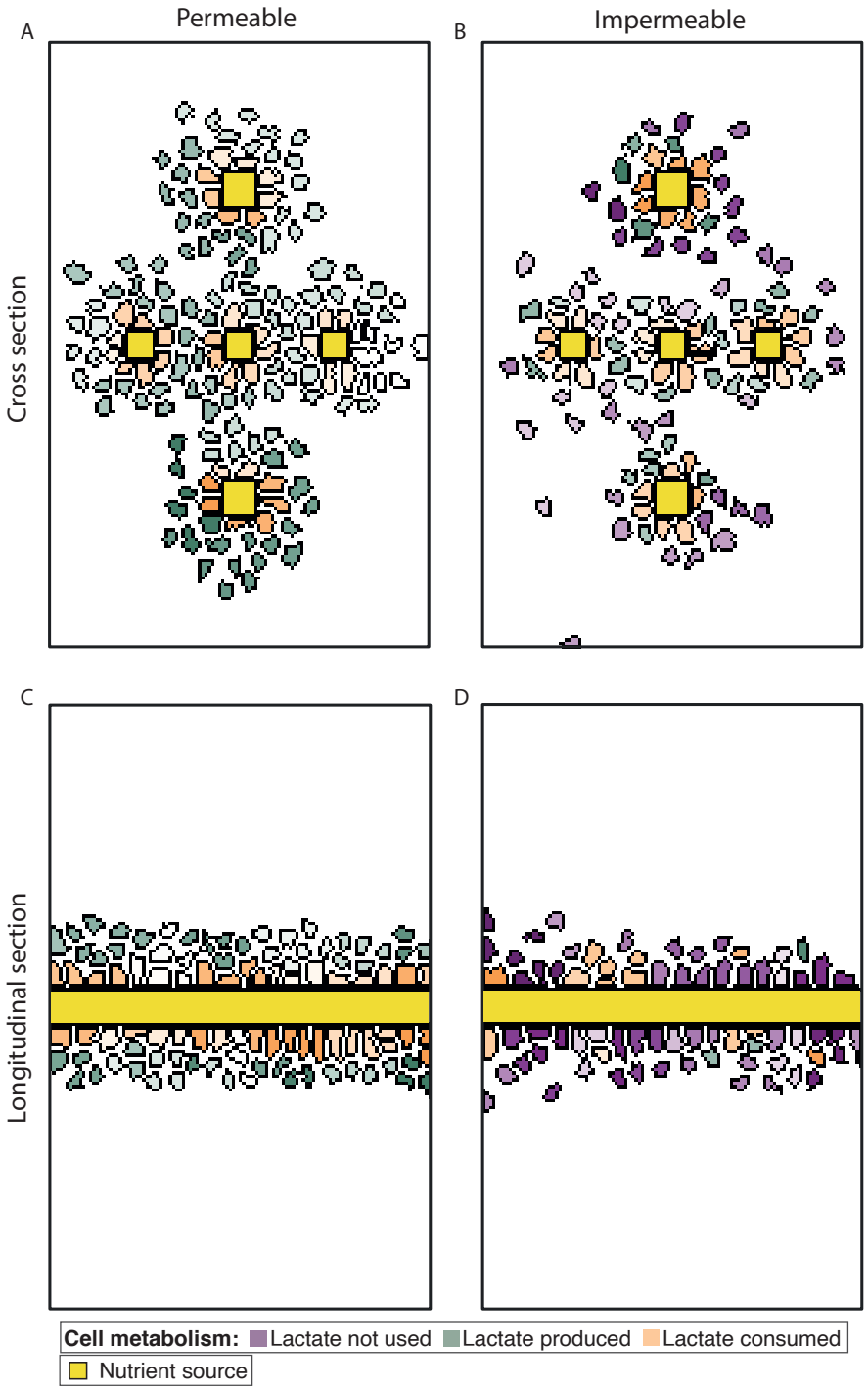


Figure 6. Snapshot of the final simulation step relative to cell population arrangement according to the four possible configurations of tissue environmental setting. The tissue is represented from a cross section (panels A,B) and from a longitudinal section point of view (panels C, D). The cells have permeable membrane (panels A, C) and impermeable membrane (panels B, D). The yellow lattice sites represent the nutrient sources in the lattice. Biological cells are differently colored according the corresponding lactate metabolism, i.e. green if lactate is produced, orange is lactate is consumed and purple if lactate is both produced or consumed with a flux less than 0.1 fmol.

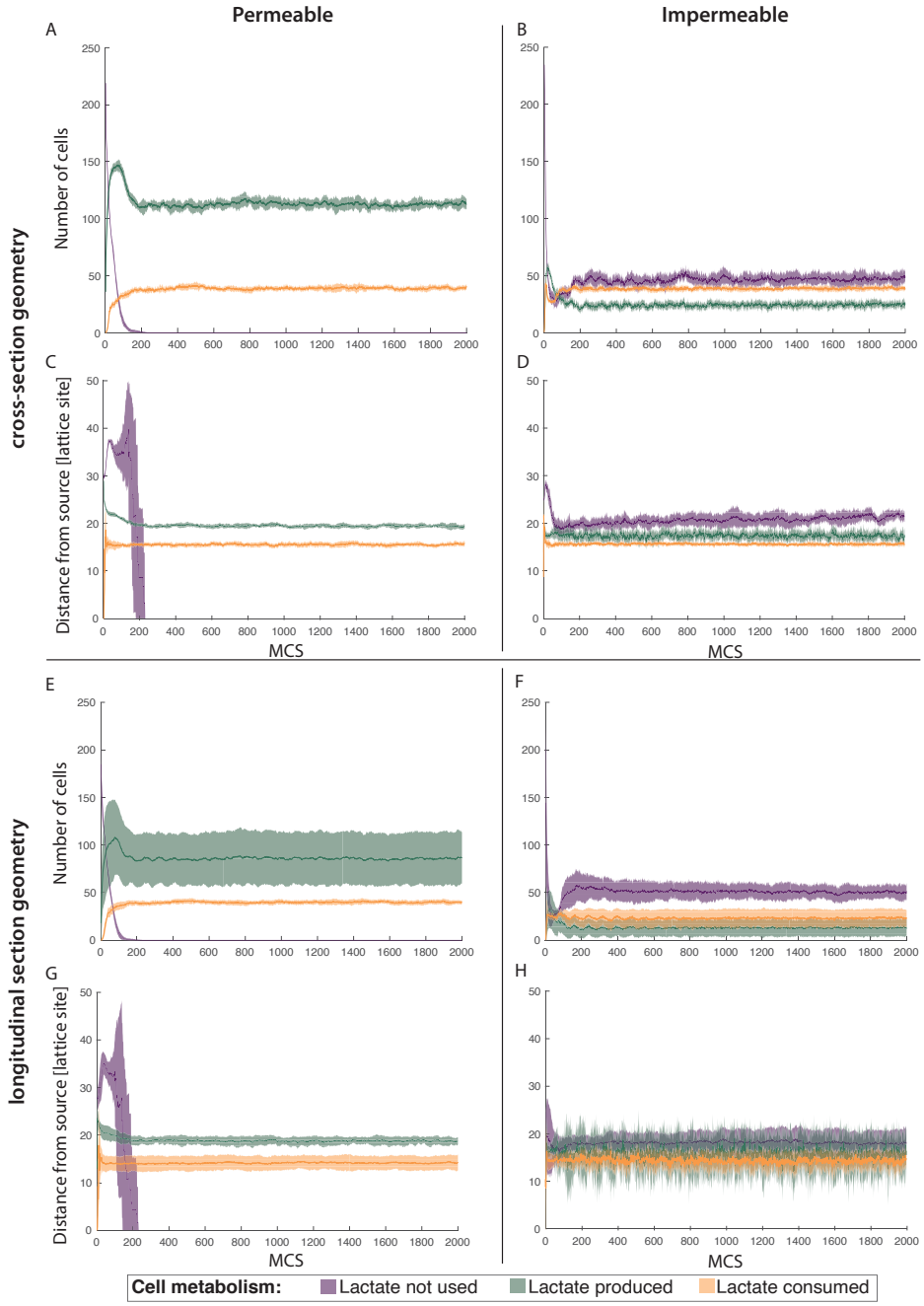


Figure 7. Comparison of the effect of cell permeability and simulation geometry on number and distance from source of cells of different metabolic phenotype. In panels A, B, C, D the geometry is shaped to represent a cross-section while in panels E, F, G, H it is shaped as a longitudinal section. Panels A, B, E and F depict the mean number of cells (solid line) and its standard deviation (shaded area) over 10 independent simulations. Panels C, D, G and H depict the mean distance of the cells from the closest nutrients source (solid line) and its standard deviation (shaded area) over 10 independent simulations, where the distance is calculated between the cell and the closest source barycentre. Biological cells are differently colored according the corresponding lactate metabolism, i.e. green if lactate is produced, orange is lactate is consumed and purple if lactate is both produced or consumed with a flux less than 0.1 fmol.

their metabolism: Outward lactate flux (green), Inward lactate flux (orange), No lactate flux (purple). Purple cells, corresponding to the ones that are not either consuming or producing lactate with a flux greater than 0.1 fmol, are missing among the permeable cells. These cells show a rapid drop in number in the early simulation cycles, completely disappearing at MCS 200, in both the cross-section (Figure 7 A to D) and longitudinal (Figure 7 E to H) simulation settings. The average number of lactate producing (outward flux, green) and their distance from the vessel was similar in both the cross-section and longitudinal simulations (panels A,B and E,F, respectively, of Figure 7), although the latter showed increased variability. In both cross- and longitudinal sections, lactate producing cells were present in higher number and farther to the vessels than lactate-consuming cells. Differently from permeable cells, the impermeable ones showed all the three types of lactate metabolism (Figure 6 and 7). Lactate producing (outward flux, green) cells represent the less abundant sub-population in both the longitudinal and cross-section simulation scenarios (panel B and F of Figure 7, respectively). Similar to the trend observed in permeable cells, longitudinal simulations of impermeable cells showed increased variability compared to cross-section variability (panels B, D and F, H of Figure 7), and a steady state was reached after about 200 MCS.

4. Discussion

In this work, we extended the previous FBCAmethodology presented in [18, 23] to evaluate the impact of different nutrient diffusion models on the population dynamics. In particular, we modeled metabolic interaction among cells by allowing secreted metabolite to diffuse over the lattice together with the diffusion of nutrients from different sources.

We proved that the inclusion of the diffusion process in the system dynamics can characterize realistic and complex biological processes and phenomena. The developed simulator allowed to perform in depth quantitative analyses of cell populations properties in two different biological scenarios, namely a cell culture in a closed environment and a tissue. In the first scenario, the simulator faithfully reproduced the diauxic growth of yeast in a well stirred flask, whereas the second scenario properly mimicked the behaviour of a tissue organization or tumor micro-environment niche. In both cases, already known phenomena have been reproduced, such as for example the switch of the carbon source exploited to grow when the mainly used one is totally consumed.

In addition, our simulator showed the ability to discriminate cells having different ways of nutrient access according to their permeability, i.e. permeable and impermeable cells. Simulations regarding tissue experimental setting revealed that the corresponding geometry considerably influence cell positions within the lattice. Nevertheless, our model faithfully represented the cellular structure that we modelled like an *in vivo* cross or longitudinal sectioning of a biological tissue. This ability could be, in the future, exploited to model more complex and close to reality blood vessels configurations.

Overall, the choice of the most suitable experimental setting is linked to the aim of the research. A higher interest on the population dynamics rather than on the investigation of how nutrients are consumed by cells leads to consider permeable cells a better approximation. In the opposite case, impermeable cells result to be more appropriate.

In the next future we plan to perform further analyses to investigate the influence of different initial set of nutrients on cellular growth dynamics, as well as to explore any difference with the current

outcomes. This strategy could assist the analysis of the differences emerged in this work between alternative experimental settings. In particular, the reason why a more heterogeneous situation occurred only when impermeable cells were considered, could be investigated.

In view of the outcomes we obtained in this work, we are confident that our simulator is enough plastic to be adapted to the simulation of various and more complex scenarios, geometry of the system and nutrient diffusion process.

References

[1] Hanahan D, Weinberg R. Hallmarks of Cancer: The Next Generation. *Cell*, 2011. **144**:646–674. doi:10.1016/j.cell.2011.02.013.

[2] Ward P, Thompson C. Metabolic Reprogramming: A Cancer Hallmark Even Warburg Did Not Anticipate. *Cancer Cell*, 2012. **21**:297–308. doi:10.1016/j.ccr.2012.02.014.

[3] Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? *Nature Biotechnology*, 2010. **28**:245. doi:10.1038/nbt.1614.

[4] Cazzaniga P, Damiani C, Besozzi D, Colombo R, Nobile M, Gaglio D, Pescini D, Molinari S, Mauri G, Alberghina L, Vanoni M. Computational Strategies for a System-Level Understanding of Metabolism. *Metabolites*, 2014. **4**:1034–1087. doi:10.3390/metabo4041034.

[5] Burrell RA, McGranahan N, Bartek J, Swanton C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 2013. **501**:nature12625. doi:10.1038/nature12625.

[6] Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, 2012. **12**:323. doi:10.1038/nrc3261.

[7] McGranahan N, Swanton C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer cell*, 2015. **27**(1):15–26. doi:10.1016/j.ccell.2014.12.001.

[8] Caravagna G, Graudenzi A, Ramazzotti D, Sanz-Pamplona R, Sano LD, Mauri G, Moreno V, Antoniotti M, Mishra B. Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proceedings of the National Academy of Sciences*, 2016. **113**:E4025–E4034. doi:10.1073/pnas.1520213113.

[9] Lipinski KA, Barber LJ, Davies MN, Ashenden M, Sottoriva A, Gerlinger M. Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. *Trends in Cancer*, 2016. **2**:49–63. doi:10.1016/j.trecan.2015.11.003.

[10] Navin NE. The first five years of single-cell cancer genomics and beyond. *Genome Research*, 2015. **25**:1499–1507. doi:10.1101/gr.191098.115.

[11] Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 2016. **17**:175–188. doi:10.1038/nrg.2015.16.

[12] Cristini V, Lowengrub J. Multiscale modeling of cancer: an integrated experimental and mathematical modeling approach. Cambridge University Press, 2010. doi:10.1017/cbo9780511781452.

[13] Deisboeck TS, Stamatakos GS. Multiscale cancer modeling. CRC Press, 2010. doi:10.1201/b10407.

[14] Matteis GD, Graudenzi A, Antoniotti M. A review of spatial computational models for multi-cellular systems, with regard to intestinal crypts and colorectal cancer development. *Journal of Mathematical Biology*, 2013. **66**:1409–1462. doi:10.1007/s00285-012-0539-4.

[15] Graudenzi A, Caravagna G, Matteis GD, Antoniotti M. Investigating the Relation between Stochastic Differentiation, Homeostasis and Clonal Expansion in Intestinal Crypts via Multiscale Modeling. *PLoS ONE*, 2014. **9**:e97272. doi:10.1371/journal.pone.0097272.

[16] Rubinacci S, Graudenzi A, Caravagna G, Mauri G, Osborne J, Pitt-Francis J, Antoniotti M. CoGNaC: A Chaste Plugin for the Multiscale Simulation of Gene Regulatory Networks Driving the Spatial Dynamics of Tissues and Cancer. *Cancer Informatics*, 2015. **14**:53–65. doi:10.4137/cin.s19965.

[17] Mina P, Bernardo Md, Savery NJ, Tsaneva-Atanasova K. Modelling emergence of oscillations in communicating bacteria: a structured approach from one to many cells. *Journal of the Royal Society, Interface*, 2013. **10**:20120612. doi:10.1098/rsif.2012.0612.

[18] Graudenzi A, Maspero D, Damiani C. Modeling spatio-temporal dynamics of metabolic networks with cellular automata and constraint-based methods. In: Giancarlo Mauri ADKNLM Samira El Yacoubi (ed.), Cellular Automata. ACRI 2018. Lecture Notes in Computer Science, volume 11115. Springer, Cham, 2018 pp. 16–29. doi:10.1007/978-3-319-99813-8_2.

[19] Graner F, Glazier JA. Simulation of biological cell sorting using a two-dimensional extended Potts model. *Physical Review Letters*, 1992. **69**(13):2013–2016. doi:10.1103/physrevlett.69.2013.

[20] Marée AF, Grieneisen VA, Hogeweg P. The Cellular Potts Model and biophysical properties of cells, tissues and morphogenesis. In: Single-cell-based models in biology and medicine, pp. 107–136. Springer, 2007. doi:10.1007/978-3-7643-8123-3_5.

[21] Scianna M, Preziosi L. Multiscale developments of the cellular Potts model. *Multiscale Modeling & Simulation*, 2012. **10**(2):342–382. doi:10.1137/100812951.

[22] Szabó A, Merks RM. Cellular Potts Modeling of Tumor Growth, Tumor Invasion, and Tumor Evolution. *Frontiers in oncology*, 2013. **3**:87. doi:10.3389/fonc.2013.00087.

[23] Maspero D, Graudenzi A, Singh S, Pescini D, Mauri G, Antoniotti M, Damiani C. Synchronization effects in a metabolism-driven model of multi-cellular system. volume 900. 2019 pp. 115–126. doi:10.1007/978-3-030-21733-4_9.

[24] Scianna M, Preziosi L. Cellular Potts Models: Multiscale Extensions and Biological Applications. CRC Press, 2013. doi:10.1201/b14075.

[25] Filippo MD, Colombo R, Damiani C, Pescini D, Gaglio D, Vanoni M, Alberghina L, Mauri G. Zooming-in on cancer metabolic rewiring with tissue specific constraint-based models. *Computational Biology and Chemistry*, 2016. **62**:60–69. doi:10.1016/j.compbiolchem.2016.03.002.

[26] Steinberg MS. On the mechanism of tissue reconstruction by dissociated cells, I. Population kinetics, differential adhesiveness, and the absence of directed migration. *Proceedings of the National Academy of Sciences*, 1962. **48**(9):1577–1582. doi:10.1073/pnas.48.9.1577.

[27] Dan D, Mueller C, Chen K, Glazier JA. Solving the advection-diffusion equations in biological contexts using the cellular Potts model. *Physical Review E*, 2005. **72**(4):041909. doi:10.1103/physreve.72.041909.

[28] Damiani C, Di Filippo M, Pescini D, Maspero D, Colombo R, Mauri G. popFBA: tackling intratumour heterogeneity with Flux Balance Analysis. *Bioinformatics*, 2017. **33**(14):i311–i318. doi:10.1093/bioinformatics/btx251.

[29] Graudenzi A, Maspero D, Di Filippo M, Gnugnoli M, Isella C, Mauri G, Medico E, Antoniotti M, Damiani C. Integration of transcriptomic data and metabolic networks in cancer samples reveals highly significant prognostic power. *Journal of Biomedical Informatics*, 2018. **87**:37–149. doi:10.1016/j.jbi.2018.09.010.

[30] Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S, Kang J, Hyduke DR, Palsson BØ. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nature protocols*, 2011. **6**(9):1290. doi: 10.1038/nprot.2011.308.

5 Data-driven multiscale modelling

In this chapter, the preliminary attempts of merging the two general methodological approaches discussed in the thesis (omics data analysis – Chapter 3 – and multiscale modelling – Chapter 4) are presented.

In particular, the inclusion of single-cell RNA-seq data into the FBCA modelling framework is shown, also discussing possible future developments.

5.1 Background: single-cell Flux Balance Analysis

As explained in section 3.1.1, one can project single-cell gene expression profiles onto metabolic networks to compute the RAS profiles. Instead of just comparing them, as done in [154], one can: (*i*) use RASs to constraint the upper and lower bounds of reactions in the metabolic model of a given cell; (*ii*) aggregate the single-cell metabolic models into a population-model, as done in [119]; (*iii*) compute the flux distribution using as objective function the optimization of the biomass at the population level.

All the steps resulted in the definition of the single-cell Flux Balance Analysis framework (scFBA) [175], which allows one to portray a snapshot of the single-cell metabolic phenotypes within a cell population at a given moment, by relying on unsupervised integration of scRNA-seq data. In the work, we proved that the integration of single-cell RNA-seq profiles allows one to point out the possible metabolic interactions via exchange of metabolites through the microenvironment, and to identify clusters of cells with different growth rates.

In general, this approach proved that it is possible to extract relevant metabolic

information even from single-cell expression profiles. However, scFBA allows only to obtain a snapshot of the cell metabolic states without considering cell spatial dynamics, nor nutrients diffusion. Since we are interested in studying the complex interactions between cells and the metabolic heterogeneity that emerges from them, we decided to extend the approach by using our FBCA framework as scaffold (see 4.2). Details are provided in the next section.

5.2 Integration of single-cell gene expression data into FBCA

In section 3.1.1, we described the methods developed to exploit gene expression profiles to characterize metabolic heterogeneity. In section 4.2, we defined the FBCA multiscale framework to simulate the spatial dynamics of a multicellular system, driven by the optimization of biomass production, as a function of the underlying metabolic model and nutrients availability. This chapter presents a preliminary attempt to merge the omics data analysis and multiscale modelling to characterize the complex emerging interactions among cancer cell subpopulations.

Notice that data-driven multi-scale modelling framework are becoming increasingly popular in the study of multicellular systems and, especially, cancer, as recently reviewed in [235]. We position our work in this sphere, which is expected to produce translational results in model-driven clinical oncology and precision/personalized medicine.

To this end, in paper P#10 we employed a scRNA-seq dataset to obtain single-cell RAS profiles, as proposed in [175]. Accordingly, we generated single-cell-specific metabolic models by constraining their upper and the lower bounds in accordance with the RAS profiles. Finally, we populated the initial configuration of the FBCA lattice with cells characterized by such metabolic models and a random spatial distribution. In particular, we employed the *Tissue-like* scenario presented in paper P#9. With this setting, space is limited and nutrients (e.g., Oxygen, Glucose, and Glutamine) diffuse from five nutrient sources with specific diffusion constant estimated from the literature. More details on how nutrients diffuse are provided in section 4.2 and in the paper.

After running a large number of simulations, we investigated two kinds of emerging properties. First, we marked cells based on the consumption or production of Lactate to define their metabolic behaviour (i.e., oxidative or fermentative). The idea is to observe if metabolic spatial patterns spontaneously emerge. The second analysis aims to detect metabolic subpopulations and how they are distributed across the space. To this aim, we first stratified cell metabolisms based on the optimal biomass that they can produce with identical nutrient availability. Then, we marked cells as *High*, *Low*, or *Average* proliferative and observed their spatial distribution at the end of the simulation. It might be sound to hypothesize that low-proliferative cells are depleted from the system, because of their lower proliferative behaviour. Instead, we noticed that this subpopulation can

indeed survive in regions distant from nutrient sources, where the scarcity of nutrients limits the overall proliferation and competition.

All in all, this preliminary result proves the effectiveness of the FBCA approach as a scaffold for the integration with single-cell omics measurements. However, some theoretical improvements may be useful to improve the realism of simulations.

First, one could increase the number of simulated cells, e.g., to integrate UMI-based datasets. To this end, despite the representation of biological cells as a set of lattice sites is helpful to consider their morphological properties, it is computationally expensive. A diverse spatial representation of the system might allow to simulate a significantly larger number of cells (from hundreds to thousands), e.g., via graph-based models [174, 227].

On the other hand, new data types might allow for a finer integration of omics data, e.g., via spatial transcriptomics (see Discussion).

Integration of Single-Cell RNA-Sequencing Data into Flux Balance Cellular Automata

Davide Maspero[1,2,6], Marzia Di Filippo[4], Fabrizio Angaroni[1], Dario Pescini[4,5], Giancarlo Mauri[1,5], Marco Vanoni[3,5], Alex Graudenzi[1,6( )], and Chiara Damiani[3,5( )]

[1] Department of Informatics, Systems and Communication, University of Milan-Bicocca, Milan, Italy
alex.graudenzi@unimib.it
[2] Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[3] Department of Biotechnology and Biosciences, University of Milan-Bicocca, Milan, Italy
chiara.damiani@unimib.it
[4] Department of Statistics and Quantitative Methods, University of Milan-Bicocca, Milan, Italy
[5] SYSBIO Centre of Systems Biology, University of Milan-Bicocca, Milan, Italy
[6] Institute of Molecular Bioimaging and Physiology, CNR, Segrate, Milan, Italy

Abstract. FBCA (Flux Balance Cellular Automata) has been recently proposed as a new multi-scale modeling framework to represent the spatial dynamics of multi-cellular systems, while simultaneously taking into account the metabolic activity of individual cells. Preliminary results have revealed the potentialities of the framework in enabling to identify and analyze complex emergent properties of cellular populations, such as spatial patterns phenomena and synchronization effects. Here we move a step forward, by exploring the possibility of integrating real-world data into the framework. To this end, we seek to customize the metabolism of individual cells according to single-cell gene expression profiles. We investigate the effect on cell metabolism of the interplay between: (a) the environmental conditions determined by nutrient diffusion dynamics; (b) the activation or deactivation of metabolic pathways determined by gene expression.

Keywords: Flux Balance Analysis · Cellular Potts Model · Single-cell RNA-seq

1 Scientific Background

The alteration of cellular metabolism plays a significant role in tumor origin and development. While the reprogramming of cancer metabolism potentially opens

Supported by SYSBIO and ITFOC.

© Springer Nature Switzerland AG 2020
P. Cazzaniga et al. (Eds.): CIBB 2019, LNBI 12313, pp. 207–215, 2020.
https://doi.org/10.1007/978-3-030-63061-4_19

new therapeutic opportunities, heterogeneity of cancer metabolism hinders the identification of effective treatments. Tumors with different tissue and cell type origin present extensive genetic and phenotypic variability, resulting in a differentiated aggressiveness and sensitivity to cytotoxic therapies. Such an inter-tumor heterogeneity is accompanied by intra-tumor heterogeneity, since multiple subclones having different genetic, epigenetic, and phenotypic features can characterize distinct regions of the same primary tumor. A complex metabolic interplay occurs among cancer cells, the host stroma, and cells of the immune system. Malignant cells may extract high-energy metabolites (e.g., lactate and fatty acids) from adjacent cells, contributing to treatment resistance. Therefore, effective therapeutic strategies should incorporate knowledge of cooperation and competition phenomena within cancer cell populations.

Many efforts have been devoted to investigate inter-tumor metabolic heterogeneity [1], which rely on tissue-specific steady-state modeling to unravel distinctive metabolic alterations of multiple cancer types that may represent potential targets for the development of personalized therapies. Fewer efforts have been instead put forward to investigate intra-tumor metabolic heterogeneity [2], which however require knowledge *a priori* of the composition of the cell population. This limit can be overcome by exploiting the information of gene expression at the single cell level today fully enabled by RNA-seq. In this regard, we have recently proposed the single-cell Flux Balance Analysis framework (scFBA) [3], which allows to portray a snapshot of the single-cell metabolic phenotypes within a cell population at a given moment, by relying on unsupervised integration of scRNA-seq data. scFBA does not explicitly model spatial organization and dynamics. Yet, cancer (sub)population evolve and compete in a (micro)environment with usually limited resource (e.g., oxygen and nutrients) and with specific spatial properties, which significantly differs in distinct tissues and organs. Therefore, modeling the spatio-temporal dynamics of heterogeneous cell populations may assist the development of strategies able to investigate processes and phenomena involving populations of interacting cells at different time/space scales. The simulation of the spatial/morphological dynamics of multicellular systems, such as tissues and organs, has recently progressed [4]. As a first attempt to combine the spatial/morphological dynamics of a multicellular model simulated via Cellular Potts Model (CPM) [5], and the metabolic activity of its constituting single cells computed via Flux Balance Analysis (FBA) [6], we developed FBCA (Flux Balance with Cellular Automata) [7,8,10]. In [9], we extended the previous framework by modeling the metabolic communication among cells, via diffusion of metabolites over the tissue according to a local spatial gradient.

In this work, we introduce the integration of single-cell RNA-seq (scRNA-seq) data into FBCA, aiming to evaluate the impact of different single-cell fluxomes on the overall cell population spatial dynamics. Metabolic behavior(s) characterizing every single cell in the population will emerge from the combination of the corresponding intracellular constraints dictated by single-cell transcriptomic data, together with its nutritional constraints. As a proof of principle, we

applied the methodology to the lung adenocarcinoma patient derive xenograft scRNA-seq data, already used in [3].

2 Materials and Methods

2.1 The FBCA framework

FBCA is a computational methodology which combines the spatial dynamics representation of the system through Cellular Potts Model (CPM) [5], with a model of cell metabolism by means of Flux Balance Analysis (FBA) [6]. For an extensive and formal description of the FBCA framework, the reader is referred to [7,8]. Here we just recall the general idea: the morphology of a generic tissue is modeled via CPM, in which biological cells are represented by sets of contiguous lattice sites, which evolve via flip attempts driven by a Hamiltonian function that accounts for the Differential Adhesion Hypothesis [5] and for the growth tendency of each cell due to the accumulation of biomass.

Accumulation of biomass is determined according to the biomass production rate of the metabolic network associated with each cell, according to the corresponding nutrient availability. A metabolic network is defined as the set of metabolites and chemical reactions taking place in a cell. The biomass production rate of a given metabolic network is computed at each Monte Carlo simulation Step (MCS) by means of FBA, according to the corresponding nutrient availability. The uptake rate of extracellular nutrients are constrained to be lower than the sum of the corresponding concentration values in the lattice sites belonging to that cell neighborhood. The concentration of nutrients in the lattice sites in turns depend on the diffusion of nutrients which is updated at each simulation step.

Given the constraints on uptake boundaries, FBA solves a Linear Programming problem to identify the flux distribution $\boldsymbol{v} = (v_1, \ldots, v_R)$ that maximizes or minimizes the biomass reaction flux v_b of the cell, given a steady-state assumption for the abundance of each metabolite, i.e., $S\ \boldsymbol{v} = \mathbf{0}$, and boundaries for allowed reaction fluxes.

At each time step s, the biomass synthesis rate is added to the current value of the biomass so far accumulated by a given cell. As described in [8], to avoid an uncontrolled raising of the cell density a limitation of biomass accumulation was introduced.

Each cell is assigned an initial equal area. When it reaches twice its initial area, it splits into two daughters, along a randomly chosen horizontal or vertical direction. Daughter cell inherit their parents' properties.

The nutrient diffusion process, as discussed in [9] is implemented by averaging the nutrient concentrations in a neighborhood of each lattice site. Different diffusion coefficients can be set for different nutrients.

At each simulation time step, the upper bounds of the nutrients uptake rates of each cell are set proportionally to its area.

2.2 Integration of ScRNA-Seq Data

In order to customize the cells according to scRNA-seq data, we first classify genes as "on" (1) if the corresponding Transcript Per Kilobase Milion (TPM) is greater than 0, as "off" (0) otherwise. We then solve the logic Gene-Protein-Reaction (GPR) rules included in the model, which determine the set of proteins that must be present for the reaction to carry flux, in a Boolean fashion, by considering the binary gene expression values. If the GPR is not satisfied (i.e., false), we limit the flux capacity of the reaction by setting the upper/lower bound as 0.01% of the original value.

We note that the Boolean simplification can be effective, to a first approximation, in investigating how the simple activation/inactivation of genes may influence the complex metabolic interplay involving cell subpopulations. Nevertheless, more refined data integration strategies, such as the employment of discretized or continues normalized expression values may be effective to this end, and will be included in further extensions of the method.

It is worth mentioning that, to reduce the computation time, we did not embrace the philosophy used in scFBA of setting the flux capacity as a continuous function of the expression of the associated genes, as this would require to perform a computationally demanding Flux Variability Analysis for each cell at each simulation step [3]. By doing so, we maintain the simulation time reported in [9].

It is also worth mentioning that we did not exploit the strategy used in scFBA of solving a unique mass balance problem for the entire population, given constraints on the uptake and consumption rate of the bulk, but each cell was modeled separately. Because space is explicitly modeled in FBCA, each cell must indeed have its own constraints, which depend in turn on the diffusion of nutrients in its neighborhood.

In our framework, nutrients diffusion and cells evolution take place with interleaved dynamics of different speeds (a MCS every ten diffusion steps). In order to avoid the case that cells starve at first steps of the simulation we let nutrients diffuse for 1500 steps before seeding the lattice with cells and letting them to enter in the dynamics process. Moreover, to guarantee the representativeness of each experimental cell phenotype, we seed the lattice with ten copies of each cell.

Datasets. In this work, we used the Lung Adenocarcinoma (LUAD) scRNA-seq dataset LCPT45Re obtained from the NCBI Gene Expression Omnibus (GEO) data repository under accession number GSE69405. The dataset is composed of 43 cells acquired from a xenograft, obtained by sub-renal implantation in mice of a surgical resection of a 37-mm irregular primary lung lesion in the right middle lobe of a 60-year-old untreated male patient.

2.3 Experimental Setting

We considered a rectangular lattice space of 150 × 100 sites to represent a tissue-like environment. The lattice is closed on all sides to avoid the washing

out of cells, which can disappear form the systems only as a result of cell death. The metabolism of each cell is simulated by using a core model of human central carbon metabolism [1]. As in [9], we considered a single cellular type and defined an Hamiltonian function to mimic the tendency of cells to fill the empty space if available. All simulation parameters are set as in [9], unless otherwise specified. In the initialization phase, the lattice is populated with ~ 500 cells with different and randomly assigned initial areas, in the range $[25, 50]$, as originally proposed in [8].

In order to introduce biophysically plausible nutrient sources in the modelling framework, in [9] we described two scenarios involving projections of three-dimensional blood vessels into the two-dimensional space of the lattice. Here we opted for the *Cross section* scenario, which represents a transversal section of a generic tissue that includes 5 squared nutrient sources positioned on the lattice (see Fig. 1).

The nutrient diffusion process is then simulated as follows. As in [9], at each diffusion step, the concentration of each nutrient in each lattice site of the nutrient source area is set equal to a specific value, i.e., oxygen is equal to $100[fmol]$, glucose is equal to $50[fmol]$, and glutamine is equal to $50[fmol]$. The lactate is not supplied in the extracellular environment, but it may be just produced and then exchanged with other cells. If nutrients are not consumed by cells, they are removed from the simulated lattice through a constant flux value when the corresponding edges of the lattice itself are reached.

In [9] we explored two different descriptions of nutrient diffusion across cellular membranes: permeable versus impermeable cells, and we concluded that, given that the population dynamics is only slightly affected by this choice, in order to speed up the computation time the second option is more convenient. Therefore, here we assume cell permeability, meaning that the diffusion process takes place "independently" from the cells positions. Cell and nutrients move on two distinct overlapping layers: the cell matrix, in which cells evolve, overlays the nutrient matrix, where metabolites freely diffuse disregarding the presence of the cells. In this configuration, cells have access to the nutrients and metabolites of each lattice site over which they "float" (See Fig. 1). More details can be found in [9].

In this work, we decided to perform 10 nutrient update steps at each step of the spatial dynamics (MCS), and to account for different diffusion rates of the distinct nutrients according to experimental estimations in literature. We also allow the nutrients to diffuse over the lattice for 1500 update steps before positioning the cells and start running the spatial simulation.

3 Results

As previously mentioned, we allow nutrients to diffuse for 1500 updates before positioning the cells. We also verified that after 1500 time steps all nutrients have diffused all over the lattice.

We consider this situation as our starting point (MCS=0). At MCS=0 we place the cells and we let them evolve. We ran 9 distinct simulations with the

same experimental setting. An example of simulation run is depicted in Fig. 1 for MCS 1, 500, 1000 and 2000. Cells are colored according to emerging metabolic properties: green cells consume lactate, where orange cells secrete it. It can be noticed that at time MCS 1, no cells consume lactate as it is not supplied with blood but it can just be secreted by cells in the lattice. It can be observed that at time 500 cells have lost their initial (unrealistic) squared shape and have filled up the lattice. Remarkably, at this point cells that consume lactate appear. As it is apparent also in later evolution steps, these cells tend to be confined at the extremities of the lattice, where nutrients are less abundant.

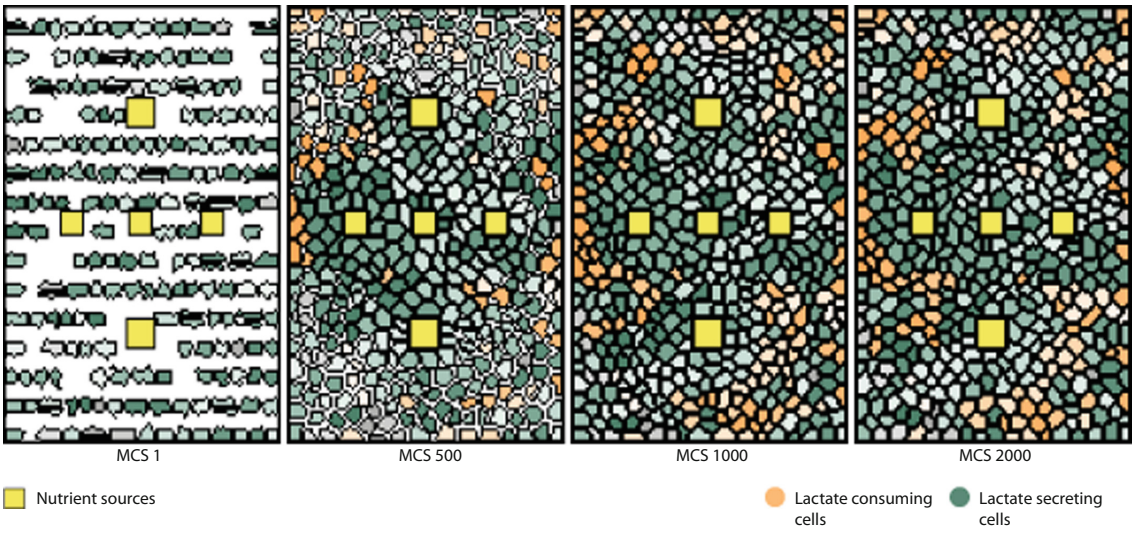


Fig. 1. Snapshots of four selected simulation steps (at time: $MCS = 1$, $MCS = 500$, $MCS = 1000$ and $MCS = 2000$) of a single simulation run. The tissue is represented from a cross section and the cells have permeable membrane. The yellow lattice sites represent the nutrient sources in the lattice. Biological cells are differently colored according the corresponding lactate metabolism, i.e., green if lactate is produced and orange is lactate is consumed. (Color figure online)

It is interesting to investigate whether this patterning is determined merely by nutrient gradients or also by the underling metabolic network, which should reflect the gene expression patterns. At this aim, we need to classify cells differently according to their metabolic network. As a first approximation, we considered as an indicator of differences in the metabolic network the differences in the theoretical capacity of the corresponding cell to make biomass given (the same) availability of all nutrients. We refer to this value as Optimal Biomass Production (OBP). Following the distribution of OPB values, we could easily divide cells into three non-overalapping groups: low OPB; medium OBP and high OBP (data not shown here).

In Fig. 2A cells are labeled with different colors according to the OBP group. The dots in Fig. 2 correspond to the barycenter of each cell in each of the 9 simulation runs at a given MCS. When observing MCS 1 (first panel from the left) it is apparent that most cells fall in the high OBP group. Cells belonging to

the tree groups are uniformly distributed across space. As the simulation proceed, it can be observed that cells tend to cluster according to their OBP, and hence to their gene expression. High OBP cells tend to colonize the surroundings of the blood vessels, whereas low and medium OBP cells are confined to the corners of the lattice. This interesting result proves that the metabolic behaviour of a cell is the result of an interplay involving gene expression patterns and the properties of the environment, in this case in terms of proximity to the nutrient source and availability of space. To further investigate such behaviour, we counted the number of cells included in each OBP group and placed at a given distance from the nutrient sources – computed in terms of lattice sites. In particular, we computed the proportion of cells of each group having distance d: $0 \leq d < 3$, $3 \leq d < 6$,$6 \leq d < 9$, ... As shown in Fig. 2B, at $MCS = 1$, the prevalence of the distinct groups is homogeneous with respect to the distance from the nutrient sources. However, as the simulation continues, the regions close to the nutrient sources get progressively colonized by the high OBP cells.

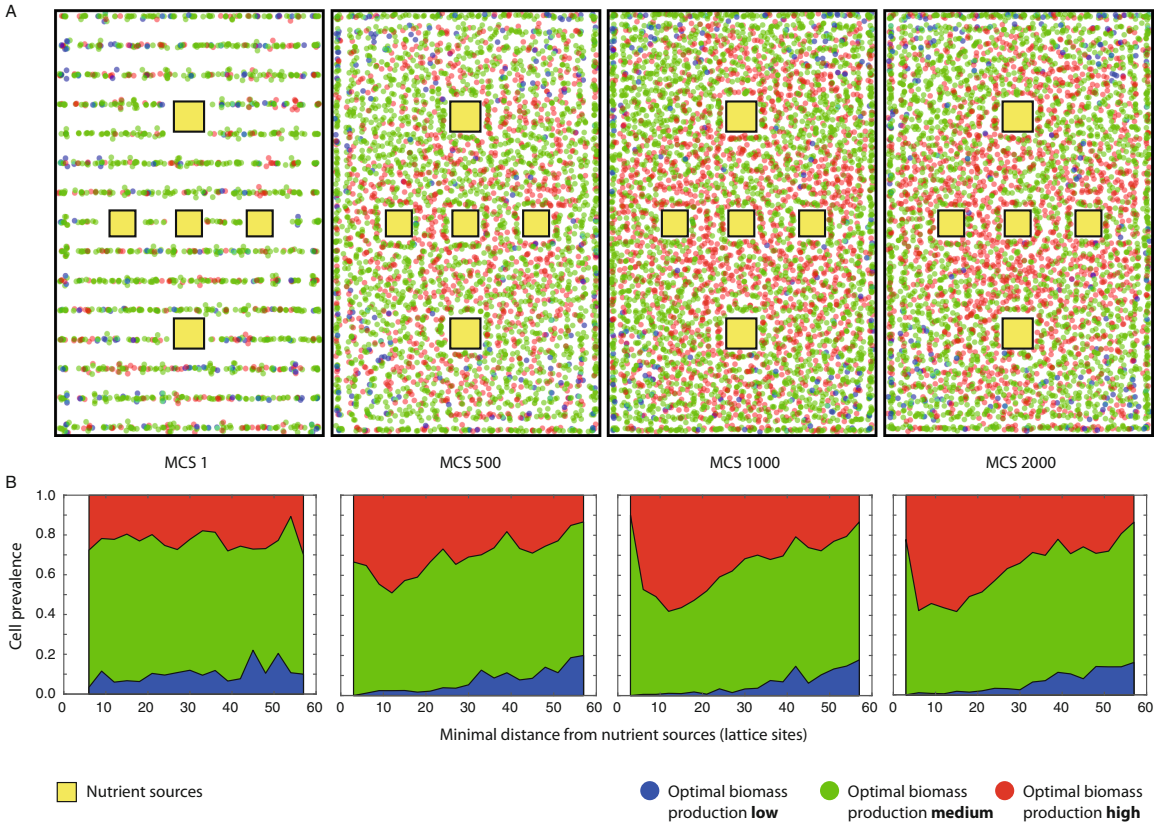


Fig. 2. A. The barycenters of the cells present in the lattice. B. Variation of the prevalence of the OBP cell populations at each time point. The distance is computed by considering for each cell only the nearest nutrient source. The plots correspond to time $MCS = 1$, $MCS = 500$, $MCS = 1000$ and $MCS = 2000$ in all 9 simulation runs are displayed. The colours are related the Optimal Biomass Production – OBP – groups, as defined in the main text: blue, green and red, corresponding to low, medium and high OBP, respectively. (Color figure online)

4 Conclusion

The results presented in Figs. 1 and 2 prove that FBCA has the potential to unravel the complex interplay between gene expression and nutrient gradients. Results are still preliminary, but suggest how a similar approach could be used to group cells in homogeneous clusters due to the projection of single-cell RNAseq data into cell-specific metabolic models. In [3] we showed how the scFBA methodology based exclusively on steady state modeling can be exploited to cluster cells with the same data, according to growth rate and metabolic phenotype. FBCA allows to refine such analysis by taking into account their spatial properties and the interaction with other cells and the environment. Of course, the former approach requires much less assumption and parameters, as scFBA needs only constraints on the rate of consumption/secretion of nutrients of the overall population, which can be promptly measured with current methodologies. On the contrary, FBCA requires information on the nutritional constraints of each single cell and is based on many assumption on the population dynamics. However, due to its high expressivity FBCA can describe multi-level complex phenomena that emerge specifically due to interaction and competition of cells in an environment with limited space and resources. This provides a powerful instrument to investigate intra-tumor metabolic heterogeneity in distinct in-silico scenarios, and surely deserves further investigations. We finally specify that a user-friendly tool for the simulation of the FBCA framework is currently under development and will be released in the near future.

Acknowledgements. The institutional financial support to SYSBIO.ISBE.IT within the Italian Roadmap for ESFRI Research Infrastructures and the FLAG-ERA grant ITFoC are gratefully acknowledged. Financial support from the Italian Ministry of University and Research (MIUR) through grant Dipartimenti di Eccellenza 2017 to University of Milano Bicocca is also greatly acknowledged.

References

1. Di Filippo, M., et al.: Zooming-in on cancer metabolic rewiring with tissue specific constraint-based models. Comput. Biol. Chem. **62**, 60–69 (2016)
2. Conde, M., do Rosario, P., Sauter, T., Pfau, T.: Constraint based modeling going multicellular. Front. Mol. Biosci. **3**:3 (2016)
3. Damiani, C., et al.: Integration of single-cell RNA-seq data into population models to characterize cancer metabolism. PLoS Comput. Biol. **15**(2), e1006733 (2019)
4. Graudenzi, A., Caravagna, G., De Matteis, G., Antoniotti, M.: Investigating the relation between stochastic differentiation, homeostasis and clonal expansion in intestinal crypts via multiscale modeling. PLoS ONE **9**(5), e97272 (2014)
5. Scianna, M., Preziosi, L.: Cellular Potts Models: Multiscale Extensions and Biological Applications. CRC Press, Boca Raton (2013)
6. Orth, J.D., Thiele, I., Palsson, B.: What is flux balance analysis? Nat. Biotechnol. **28**(3), 245 (2010)

7. Graudenzi, A., Maspero, D., Damiani, C.: Modeling Spatio-Temporal Dynamics of Metabolic Networks with Cellular Automata and Constraint-Based Methods. In: Mauri, G., El Yacoubi, S., Dennunzio, A., Nishinari, K., Manzoni, L. (eds.) ACRI 2018. LNCS, vol. 11115, pp. 16–29. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99813-8_2
8. Maspero, D., et al.: Synchronization Effects in a Metabolism-Driven Model of Multi-cellular System. In: Cagnoni, S., Mordonini, M., Pecori, R., Roli, A., Villani, M. (eds.) WIVACE 2018. CCIS, vol. 900, pp. 115–126. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-21733-4_9
9. Maspero, D., et al.: The influence of nutrients diffusion on a metabolism-driven model of a multi-cellular system. Fundam. Inform. **171**(1–4), 279–295 (2020)
10. Graudenzi, A., Maspero, D., Damiani, C.: FBCA, a multiscale modeling framework combining cellular automata and flux balance analysis. J. Cell. Automata **15** (1/2), 75–95 (2020)

6 Discussion

In this work, I presented my endeavours to exploit the potential of computer and data science in addressing a fundamental question in biomedical sciences, namely: *how can one successfully characterize the heterogeneity that emerges in a complex biological system?*

To pursue this objective, I have investigated in-depth the properties of many different biological systems and phenomena, so to design, develop and apply new effective computational methods that may generate experimental hypotheses with translational relevance, and which, therefore, might support experimentalists and clinicians. In the following, I briefly discuss some of the main achievements produced with respect to the four main topics of the work, i.e., (*i*) definition of omics data preprocessing pipelines, (*ii*) methods for omics data analysis and integration, (*iii*) multiscale modelling and simulation of multicellular systems, and (*iv*) data-driven multiscale modelling.

1. **Definition of omics data preprocessing pipelines**
   Next-Generation Sequencing technologies have generated an impressive amount of high-resolution data. Appropriate computational pipelines are crucial to extract and handle biological information contained in the data.

   In this regard, we considered two types of data, namely: (*i*) gene expression profiles and (*ii*) mutational profiles. The former are used to study phenotypic heterogeneity, while the latter to investigate genetic heterogeneity. In particular, we mostly focused on data produced by single-cell sequencing experiments that allow one to investigate heterogeneity in depth, despite a number of challenges related to

the high noise levels induced by technological and experimental limitations (see sections 2.2 and 2.3).

With regard to the preprocessing of gene expression data, we presented a thorough comparative assessment of 19 widely used denoising and imputation methods for scRNA-seq data (paper P#1). The goal was to deliver quantitative guidelines on which methods may be more effective and appropriate, for any given dataset type, sequencing protocol and downstream analysis.

The work related to pipelines for mutational profile generation led to the development of a variant calling pipeline from single-cell RNA-seq data (paper P#2). Specifically, we have shown that genotyping from scRNA-seq reads allows one to reconstruct the genomic identity of single cells with great accuracy, also leading to the detection of possible errors in experimental workflows. Therefore, similar pipelines might allow one to produce both somatic mutation and gene expression profiles to investigate intra-tumor heterogeneity from the very same data source.

Finally, with the VirMutSig pipeline (paper P#A1), we took advantage of the most recent advancements in workflow management to define a self-contained pipeline that allows the user to easily perform the discovery of viral mutational signatures. Importantly, the pipeline entirely runs in a Docker image, ensuring reproducibility of the results, along the line of Open Science practices.

2. **Methods for omics data analysis and integration**
   The second important goal of the thesis was to develop and improve methods to extract information from omics data, which, in our case, are both gene expression and mutational profiles. The methods must be robust and accurate to produce new translational knowledge on complex phenomena and diseases. In particular, we aimed at dissecting the heterogeneity that emerges from the interaction of the components of multicellular systems, as in cancer and viral evolution.

   - *Gene expression profiles*
     The methods developed for the analysis of gene expression profiles first proved that it is possible to extract knowledge about the metabolic states of a biological system, by relying on the related transcriptomic data. Leveraging on our previous works [154], we developed and published a tool for metabolic reaction enrichment analysis (MaREA) in the Galaxy Project repository, so to ensure maximum usability and reproducibility, two aspects that often are neglected.

     Furthermore, we exploited state-of-the-art machine learning algorithms to classify cancer and healthy samples, on the basis of the topological properties of the underlying metabolic networks (paper P#4). Interestingly, we

have shown that only 5 topological features, computed via standard network analyses, are sufficient to classify samples with great accuracy, pointing at the existence of key generic properties and regularities of real-world systems. Overall, inter-patient heterogeneity was exploited in a novel way, i.e., by considering the shape and connections of metabolic networks derived from gene expression profiles.

- *Mutational profiles*
  The main purpose of the computational strategies for mutational profiles was that of characterizing the evolution of a biological system. In fact, the evolutionary history of a system explains the progressive generation and alteration of the heterogeneity that is observed at any given time. In particular, here we considered two complex systems: cancer and viral evolution.

  In LACE (paper P#5), we developed the first method aimed at reconstructing cancer evolution from longitudinal single-cell sequencing data. The method improves over state-of-the-art approaches in different in-silico scenarios (e.g., with sampling limitations and high noise level), and is more expressive thanks to the introduction of the *longitudinal weighted clonal tree* formalism.

  Importantly, thanks to the robustness of its algorithmic framework, LACE allows one to use mutational profiles generated from single-cell RNA-seq data, allowing for a natural mapping of genotype and phenotype, which is one of the major results of this thesis in terms of data integration.

  From the translational perspective, LACE is the first method that explicitly allows to evaluate the impact of a therapy from single-cell sequencing longitudinal data, (e.g., collected before, during, and after therapy administration), and from different biological sources (e.g., patient-derived organoids, xenografts, cell cultures, etc.). This aspect is relevant, mostly considering that longitudinal datasets are expected to be increasingly common in the near future.

  In section 2.1, we explained that viral samples can be analyzed with NGS technologies to produce deep sequencing data and to retrieve the whole mutational spectrum, instead of being limited to the consensus sequence generated by classical methods such as Sanger sequencing, and used by the large majority of phylogenomic/phylodynamic studies.

  Accordingly, we designed a new framework, VERSO (paper P#6), which takes full advantage of such data with two specific purposes. First, VERSO improves the robustness of the inferred phylogeny in the case of sampling limitations and noisy observations, as witnessed during the current SARS-CoV-2 pandemic. Importantly, the underlying statistical frameworks borrows some key theoretical concepts employed for the inference of cancer evolution,

proving that a dialogue between usually disconnected fields is effective and needed.

Moreover, VERSO is the first framework that exploits low-frequency mutations to characterize intra-host heterogeneity and reconstruct likely chains of infections (validated by contact tracing data), which would not be possible with consensus sequences. Also, the identification of homoplastic events (i.e., emergence of the same mutation in disconnected clades) provides clues on positively selected variants, and may drive experimental research.

One of the limitation of both inference frameworks, which are based on Markov chain Monte Carlo sampling, is that the number of variables (i.e., mutations) cannot be too large with respect to observations (i.e., samples or single-cells), because this would preclude to reach the convergence in acceptable times and avoiding equivalent solutions. Therefore, we lately developed a new algorithm that exploits the solutions explored durig the MCMC sampling to compute a consensus tree via optimal branching (COB tree), instead of returning the maximum likelihood solution (section 3.2.1.1). Preliminary results display an improvement in the inference of the correct ordering among mutations, deserving further investigations.

In the last part of the work related to this topic, we focused on methods to characterise mutational processes regarding virus-host interactions (paper P#7). By applying a statistically robust approach based on Non-Negative Matrix Factorization, one can decompose the mutational spectra of viral samples to reconstruct mutational signatures. This strategy allowed us to identify 3 signatures of SARS-CoV-2, as well as to cluster samples into homogeneous groups, hence dissecting the host-related heterogeneity, which could be related to the different responses to the disease. Again, we showed how data science is essential to extract useful information from complex and heterogeneous data.

3. **Multiscale modelling and simulation of multicellular systems**
   During the PhD project, I have shown that it is possible to provide experimental and translational hypotheses via multiscale modelling and simulations.

   We first developed a new multiscale model that combines, for the first time, Flux Balance Analysis and Cellular Potts Model, to simulate cellular population dynamics, named Flux Balance Cellular Automata (FBCA). The spatial and population dynamics is driven by the rate of biomass production determined by a metabolic model and rely on the metabolites available in the surrounding neighborhood.

   In paper P#8, the model was introduced, and we demonstrated its utility in representing complex scenarios and quantifying emergent properties, like space competition among metabolically heterogeneous cell subpopulations, as well as the effect

of therapy administration.

In paper P#9, we increased the complexity by introducing the diffusion of nutrients in the environment. This new scenario allowed us to assess the emerging metabolic behaviour, instead of pre-determining it. For example, we could evaluate the emerging metabolic behaviour in terms of distance from nutrient sources. Importantly, thanks to optimized programming and parallelization, it was possible to simulate hundreds of cells simultaneously, which represent an important result in term of code scalability. Further improvements might be achieved by exploiting High Performance Computing and optimized cell population dynamics simulation frameworks [151]. The challenge, in this case, lies in the integration of the constraint-based metabolic model within such population model.

4. **Data-driven multiscale modelling**
   To fill the gap between data and theory, we finally focused on integrating single-cell RNA-seq data into the multiscale model FBCA. In paper P#10 we showed how FBCA is an excellent scaffold to integrate gene expression data obtained from real experiments. In this way, it is possible to characterize the heterogeneity among cells of a given tissue or of a tumour based on real measurements, and simulate the emerging behaviour. The final aim is to move from explanatory/descriptive models towards predictive ones, possibly anticipating the future evolution of the system/disease and, thus, providing general indications for effective therapies.

   This might, in turn, lead to the implementation of control-theoretic strategies for the definition of optimized therapies. In this respect, we have already investigated the effectiveness of control theory to determine the best therapeutic strategies for oncological patients. In appendix P#A2, we report our work describing the optimization of the therapy administration and schedule of Imatinib in patients with chronic myeloid leukemia (CML). Accordingly, the FBCA framework could be used to test different therapeutic strategies to be optimized via control theory, so to achieve different objectives such as maximization of drug efficacy and minimization of side effects.

To conclude, we proved that computational strategies are an essential instrument to make sense of the increasing amount of omics data, as well as to produce realistic in-silico models of real-world systems and phenomena.

A number of future developments are possible and are mostly related to the continuous technological advancements in omics data generation. Without the pretence of being exhaustive, I would like to point at two of the most interesting experimental breakthroughs: (*i*) technologies and protocols for the generation of multiple omics data from the same cell, and (*ii*) spatial transcriptomics.

In this work we have shown that it is possible and sound to project one omic layer

onto another, e.g., mapping gene expression onto metabolic profiles, or calling genomic variants from RNA-seq experiments. This somehow obviates the need for multiple omics measurements from the same cell, which are however starting to become available.

For istance, it is now possible to couple transcriptomic data with either genomic (e.g., G&T-seq [98]), proteomic (e.g., CITE-seq [139]), or epigenomic (e.g., scM&T-seq [111], scNMT-seq [146]) data from the same single-cell. Moreover, the coupled measurements of other omics layers was also recently developed, as with the scGET-seq protocol, which allows to obtain genetic and epigenetic information at the single cell level [255]. Please refer to [237] for up-to-date reviews on the topic. The investigation of the multicellular heterogeneity might clearly benefit from the datasets produced with such new protocols. For instance, inference of clonal structure and evolution could be enhanced considering also cell epigenetic states. Analogously, noise in the data might be reduced and corrected by considering the information extracted from different omics layers. To make another example, the RAS computation described in Section 3.1.1 assumes the gene expression level are a good proxy for the activity of a given metabolic reaction. This score could be more consistently estimated if one could also know the abundance of the corresponding enzyme, thus improving all downstream analyses (paper P#3, paper P#4, and paper P#10).

The omics technologies presented in this thesis require tissue dissociation to isolate cells, loosing the tissue morphology and any spatial information. Recent advancements in spatial molecular profiling technologies have allowed the complete molecular characterization of the cells, while maintaining intact their spatial and morphological context. Together with imaging data, spatial transcriptomic data provide exceptional possibilities to investigate tissue heterogeneity, spatial cell organization and characterize the microenvironment [128, 221, 256, 257]

In this regard, the diverse methods currently available for spatial transcriptomics are expected to generate highly detailed spatial maps of single-cell gene expression. As with other previous revolutions in systems biology, we need a parallel development of computational frameworks to integrate such measurements to characterize tissue heterogeneity. For examples, one might be interested in determining the tissue metabolic response to diverse inputs, e.g., drugs [131]. In this respect, our FBCA multiscale model, which already considers the metabolism as key aspect, could benefit from spatial transcriptomic data to obtain both the gene expression and the spatial distribution of single cells. Such information might be used to to define a realistic microenvironment, as well as the initial states of the simulation. All in all, this might allow one to fit the model parameters to better represent real-world scenarios, so to transform our descriptive modelling framework into a predictive one. Accordingly, this could be useful to provide prognostic, diagnostic or therapeutic insights about the complex disease under investigation.

6.1 Impact

Cancer is one of the primary causes of deaths worldwide. In 2020, 2.3 million women had been diagnosed with breast cancer and 685 000 deaths were recorded globally, making it the world's most prevalent cancer (source: World Health Organization [261]). As explained thoroughly in Section 1, the high levels of both inter- and intra-tumor heterogeneity that are observed in most cancer types frustrate the discovery and administration of effective therapies and, accordingly, impact the functioning of the whole healthcare system. In fact, the former induces the need for a patient-specific treatment that requires a deep characterization of the current and future state of the given tumor, whereas the latter underlies the emergence of drug-resistant mechanisms directly countering the efficacy of any therapy. To tackle this problem, it is essential to understand the dynamics of the interaction of cancer cell subpopulations among each other and with the surrounding microenvironment.

As shown in 3.1, a different biological phenomenon in which genetic heterogeneity plays a key role is the COVID-19 pandemic, which impacted our society on various levels, causing healthcare, economical and sociological emergencies. Both national and international institutions have not been fully able to contrast the spread of the virus, first due to the complex dynamics of geotemporal diffusion and, second, to the emergence of new viral variants, which affected the efficacy of vaccines and the severity of the disease [260].

The continuous advancements of biotechnology and of information and communication technology, together with the ever increasing amount of biological data generated and made available to the scientific community, has provided new tools to investigate the heterogeneity of complex biological systems.

This work, in particular, has showed that computer and data science are essential to make sense of the great wealth of omics data, so to deliver reliable explanations and predictions of any complex biological system and phenomenon, which will expectedly impact the current and future practices in many different fields. Such impact will be favoured also thanks to the high standards of reproducibility and diffusion, and to the adherence to the open science guidelines.

All the computational methods and pipeline described in Chapter 3 can be further improved to include more sophisticated assumptions, or extended to handle new data types, so to be used as a starting point for future research with diagnostic, prognostic and therapeutic aims. Furthermore, all proposed approaches may be translated to the study of distinct complex systems with similar properties in different research fields, in a sort of methodological exaptation.

Overall all, computer science is becoming more and more integrated with life sciences. New relevant biological discoveries probably will undoubtedly require effective

and scalable computational workflows to be accomplished. I believe that the outcome of the thesis provides an important part to this end.

A

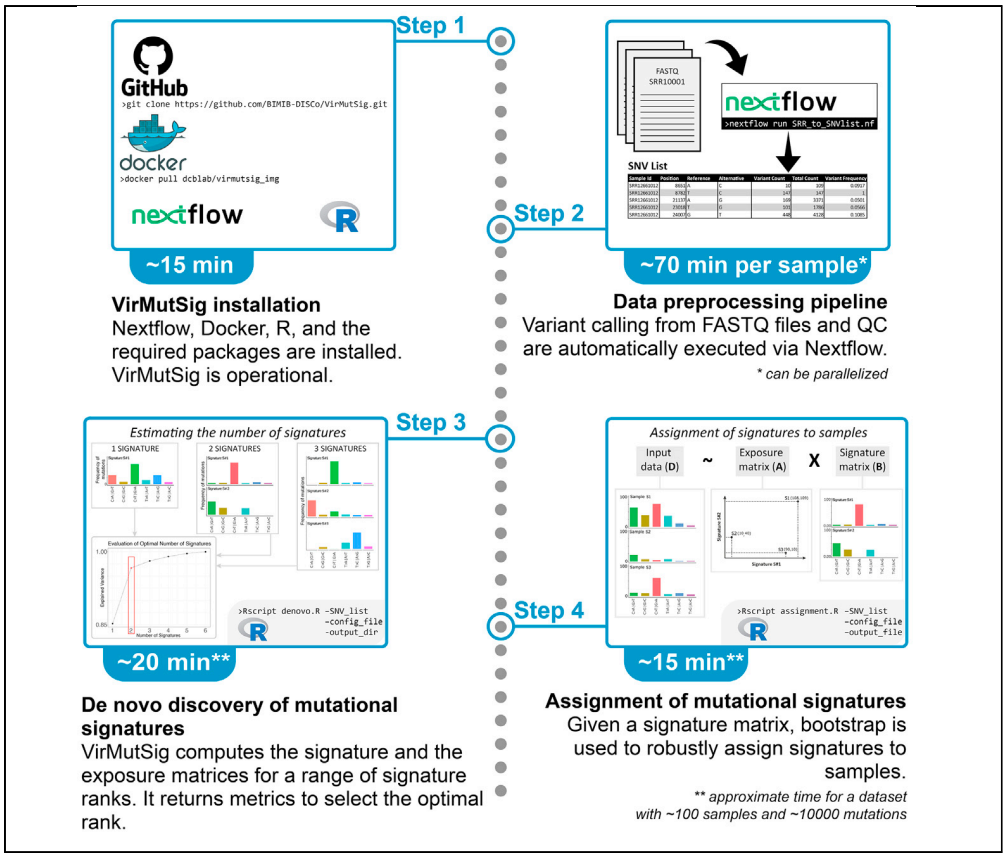
Appendix: Additional papers

STAR Protocols

CellPress
OPEN ACCESS

Protocol

VirMutSig: Discovery and assignment of viral mutational signatures from sequencing data



We describe the procedures to perform the following: (1) the *de novo* discovery of mutational signatures from raw sequencing data of viral samples and (2) the association of existing viral mutational signatures to the samples of a given data set. The goal is to identify and characterize the nucleotide substitution patterns related to the mutational processes that underlie the origination of variants in viral genomes. The VirMutSig protocol is available at this link: https://github.com/BIMIB-DISCo/VirMutSig.

Davide Maspero, Fabrizio Angaroni, Danilo Porro, Rocco Piazza, Alex Graudenzi, Daniele Ramazzotti

d.maspero@campus.unimib.it (D.M.)
alex.graudenzi@ibfm.cnr.it (A.G.)
daniele.ramazzotti@unimib.it (D.R.)

Highlights

Distinct mutational processes underlie the origination of viral variants

VirMutSig implements variant calling from raw sequencing data of viral genomes

The tool performs de novo discovery of mutational signatures (substitution patterns)

Robust assignment of signature activity to samples is performed via bootstrap

Maspero et al., STAR Protocols 2, 100911
December 17, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.xpro.2021.100911

STAR Protocols

CellPress
OPEN ACCESS

Protocol

VirMutSig: Discovery and assignment of viral mutational signatures from sequencing data

Davide Maspero,[1,2,6,*] Fabrizio Angaroni,[2] Danilo Porro,[1] Rocco Piazza,[4,5] Alex Graudenzi,[1,3,5,*] and Daniele Ramazzotti[4,5,7,*]

[1]Inst. of Molecular Bioimaging and Physiology, Consiglio Nazionale delle Ricerche (IBFM-CNR), Segrate, Milan, Italy

[2]Department of Informatics, Systems and Communication, Univ. of Milan-Bicocca, Milan, Italy

[3]Bicocca Bioinformatics, Biostatistics and Bioimaging Centre – B4, Milan, Italy

[4]Department of Medicine and Surgery, Univ. of Milan-Bicocca, Monza, Italy

[5]Senior author

[6]Technical contact

[7]Lead contact

*Correspondence: d.maspero@campus.unimib.it (D.M.), alex.graudenzi@ibfm.cnr.it (A.G.), daniele.ramazzotti@unimib.it (D. R.)

https://doi.org/10.1016/j.xpro.2021.100911

SUMMARY

We describe the procedures to perform the following: (1) the *de novo* discovery of mutational signatures from raw sequencing data of viral samples and (2) the association of existing viral mutational signatures to the samples of a given dataset. The goal is to identify and characterize the nucleotide substitution patterns related to the mutational processes that underlie the origination of variants in viral genomes. The VirMutSig protocol is available at this link: https://github.com/BIMIB-DISCo/VirMutSig.

For complete information on the theoretical aspects of this protocol, please refer to Graudenzi et al. (2021).

BEFORE YOU BEGIN

Problem description

The VirMutSig protocol aims at identifying mutational signatures from raw sequencing data of viral samples, as originally proposed in the context of cancer evolution in (Alexandrov et al., 2013) and in the analysis of SARS-CoV-2 in (Graudenzi et al., 2021). Mutational signatures represent the decomposition of the categorical distribution of nucleotide substitutions that are observed in the samples of a given dataset, and which might be due to distinct mutational processes. Such processes could be endogenous (e.g., APOBEC deaminase activity causes mainly C to T substitutions) or exogenous (e.g., tobacco smoke causes mainly C to A substitutions).

We formulated the problem of *de novo* signature discovery and assignment as a Non-negative Matrix Factorization (NMF) problem (Lal et al., 2021; Graudenzi et al., 2021).

In brief, the counts of nucleotide substitutions observed in all viral samples of a given dataset, after preprocessing and variant calling, result in an input data matrix $\mathbf{D}$, composed by n samples (rows) X m nucleotide substitution categories (e.g., the C to T category will include the count of all variants with such substitution in any given sample) (columns).

Formally, $\mathbf{D}$ is factorized in the following matrices:

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Signature matrix (B): this matrix specifies the composition of every detected signature in terms of the selected substitution categories.

Exposure matrix (A): it represents the coefficients of the linear combination of signatures present in a sample. This matrix evaluates the activity of the various signatures in each sample of the dataset.

As we are dealing with noisy data, we propose a stochastic optimization performed via a Monte Carlo Markov Chain (MCMC) to find the best A and B matrices such that AxB $\sim$ D (see the graphical abstract). The method computes A and B for a number of signatures in a user-defined range (e.g., from 2 to 10) and allows one to select the optimal number by assessing different metrics (see below).

Specifically, we release two R scripts named respectively: denovo.R and assignment.R.

1. denovo.R allows one to perform the inference of both A and B. In particular:
   a. B is randomly initialized.
   b. A is inferred given B via non-negative least squares (NNLS).
   c. B is then inferred given A via NNLS.
   Tasks "b" and "c" are repeated until convergence. The whole procedure is repeated from task "a" for a sufficient number of times (e.g., at least 100), by testing the rank of B (i.e., the number of signatures) within a user specified range.
   The output of this procedure is:

B, obtained as the consensus from all the proposed matrices; plots regarding different metrics, so to estimate the optimal number of signatures present in the dataset (see below).

2. The assignment.R script employs as input a B matrix that could be provided by the user or obtained with the script denovo.R. Then, it infers A similarly as done in the task "b" of the denovo.R script (see above).

Finally, a *bootstrap* procedure is performed to assess the statistical significance of the signature activities. To do this, we consider each sample as a categorical distribution over the given contexts, and we extract, with resampling, from such distributions the same number of mutations as observed in the sample. Then, given the bootstrapped dataset, we fit again A.

We repeat this process multiple times (e.g., 100) to obtain a distribution for each entry of A. So, a p-value of the significance of the activity of each signature in a given sample can be returned.

Overall structure of the VirMutSig protocol

The protocol is subdivided into 4 distinct parts:

Part 1) INSTALLATION

Part 2) DATA PREPROCESSING PIPELINE (via Nextflow)

Part 3) DE NOVO DISCOVERY OF MUTATIONAL SIGNATURES (denovo.R)

Part 4) ASSIGNMENT OF EXISTING MUTATIONAL SIGNATURES (assignment.R)

The protocol is publicly available at this link: https://github.com/BIMIB-DISCo/VirMutSig.

KEY RESOURCES TABLE

Reagent or RESOURCE	Source	Identifier
Deposited data		
Public database analyzed [example]: RNA-seq	NCBI	[*example*] PRJNA610428 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA610428)
Public database analyzed [example]: Amplicon	NCBI	[*example*] PRJNA645906 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA645906)
Software and algorithms		
VirMutSig	Graudenzi et al. (2021)	https://github.com/BIMIB-DISCo/VirMutSig
Unix/MacOS operating system	Canonical Ltd. Apple Inc.	Ubuntu 20.04 Focal macOS 10.15 Catalina
Docker	Merkel, (2014).	https://www.docker.com/get-started (v. 20)
Docker Image	Docker Hub	dcblab/virmutsig_img:latest (https://hub.docker.com/r/dcblab/virmutsig_img)
Nextflow	Di Tommaso et al. (2017)	https://www.nextflow.io/ (v. 21)
R	The R Foundation	https://www.r-project.org (v. 4)

MATERIALS AND EQUIPMENT

Computational requirements

To execute the VirMutSig protocol, the user requires a computer with the following software specifications:

Software

- Unix/OS operating system (for details on the usage of the protocol on Windows OS, please refer to the troubleshooting problem 5).
- Nextflow (version = 21) (https://www.nextflow.io/)
- Docker (version = 20) (https://www.docker.com/get-started)
- R (version = 4) with the installed packages listed below (https://www.r-project.org)
- The Docker image of VirMutSig (https://hub.docker.com/r/dcblab/virmutsig_img)
- A working internet connection (*for installation only*).

All the other required tools and libraries will already be installed in the provided Docker image.

Software and R packages		
Part	**Package**	**Version**
1–2) INSTALLATION and DATA PREPROCESSING PIPELINE	Nextflow (Di Tommaso et al., 2017)	version = 21
	Docker (Merkel 2014)	version = 20.10.8
	Docker image: dcblab/virmutsig_img	version = latest
	R	version = 4.0.1
3) DE NOVO DISCOVERY OF MUTATIONAL SIGNATURES (denovo.R)	BiocManager (Morgan 2021)	version = 1.30.16
	Biobase (Huber et al., 2015)	version = 2.46
	data.table (Dowle and Srinivasan, 2021)	version = 1.14.0
	ggplot2 (Wickham 2016)	version = 3.3.5
	gridExtra (Auguie 2017)	version = 2.3
	NMF (Gaujoux and Seoighe, 2020)	version = 0.23.0
	nnls (Mullen and van Stokkum, 2012)	version = 1.4
	Optparse (Davis, 2020)	version = 1.6.6
	Stringi (Gagolewski 2021)	version = 1.7.4
	Yaml (Stephens, 2020)	version = 2.2.1

(*Continued on next page*)

Continued

Software and R packages

Part	Package	Version
4) ASSIGNMENT OF EXISTING MUTATIONAL SIGNATURES (assignment.R)	All the above denovo packages	
	Glmnet (Friedman et al., 2010)	version = 4.1–2
	Lsa (Wild, 2020)	version = 0.73.2
	Matrix (Bates and Maechler, 2021)	version = 1.3.4

Hardware

- Parts 1–2

The resources required for processing a single sample are limited, but we suggest running Part 2 in parallel, if possible. Memory: 8Gb required, processors: 1 required, 8 recommended.

- Parts 3–4

Memory: 4GB required, processors: 1 required, 4 recommended.

Input data

The VirMutSig protocol requires as input either: (*i*) RNA-seq, or (*ii*) Amplicon Illumina raw data, generated from sequencing experiments of viral samples (obtained, e.g., from primary isolates), and typically available as FASTQ files.

Generally, one will have a FASTQ file for each sample of the dataset.

Notice that the protocol works with both (*i*) single- and (*ii*) paired-end sequencing library preparation layout.

FASTQ files

FASTQ files can be collected in different ways.

If one is interested in the analysis of public datasets, e.g., those available on NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra), one can create a file with a list of SRR accession number (one per line) to use as input data.

The VirMutSig will automatically download the proper FASTQ files; see Part 2.

Instead, if one has produced her/his own FASTQ files, or if they are available in other databases, one must aggregate them in one directory and specify its path as input; see Part 2.

Reference genome file

A reference genome is required to perform variant calling. In (Ramazzotti et al., 2021) we proposed the SARS-CoV-2-ANC as reference genome. Such genome is already available in the VirMutSig GitHub repository and is used as default.

However, any viral genome can be used as reference. To do so, the chosen genome must be provided as a FASTA file; see Part 2.

STEP-BY-STEP METHOD DETAILS

Part 1. Installation

Timing: 15 min

1. Downloading VirMutSig
   The installation of VirMutSig is done by downloading the GitHub repository in a local directory, which includes the VirMutSig scripts and the example files.
   After the installation, a new folder named VirMutSig will be generated.
   Please install VirMutSig by moving to your local directory and using GitHub with the following command in the terminal:

```
>git clone https://github.com/BIMIB-DISCo/VirMutSig.git
```

   VirMutSig includes 4 directories with the following names:
   a. "preprocessing"
      This directory contains all the files and directories required to perform Part 2, such as:
      i. SRR_to_SNVlist.nf: the Nextflow pipeline file
      ii. nextflow.config: the config file with the preprocessing settings and parameters
      iii. reference: a directory with the SARS-CoV-2-ANC reference file
      iv. bin: a directory with an R script used to create the SNVs list for Parts 3–4.
   b. "denovo"
      This directory contains the corresponding R script to perform the *de novo* discovery of mutational signatures (Part 3). The files included are:
      i. denovo.R
         This script performs the discovery of viral mutational signatures from the list of selected SNVs taken as input, by employing a NMF approach described in the above section or in (Lal et al., 2021). The user must also specify the number of contexts, i.e., the flanking bases to the genome positions. They may either be 6 or 96, please refer to (Lal et al., 2021) for further details.
      ii. denovo_config.yaml
         This file contains the parameters explained above. It could be modified using any text editor.
      iii. denovo_utils.R
         This R file contains some functions used by the main denovo.R scripts.

   Note: please do not modify this file.
   c. "assignment"
      This directory contains the corresponding R script to perform the *assignment* of the activity of mutational signatures to each sample (Part 4). The files included are:
      i. assignment.R
         This script takes as input the signatures.txt file generated by the denovo.R script or provided by the user, and the list of SNVs to perform the assignment of the signatures to each sample. Bootstrap can be employed to assess the statistical confidence of the signature assignments.
      ii. assignment_config.yaml
         This file contains the parameters explained above. It could be modified using any text editor.
      iii. assignment_utils.R
         This R file contains some functions used by the main assignment.R scripts.

Note: please do not modify this file.

d. "example"

This directory includes an example of the VirMutSig analysis performed on 150 FASTQ files obtained from samples of SARS-CoV-2 via RNA sequencing experiments.

2. Download docker image

The preprocessing pipeline executes all the steps using a Docker image in which all the requested software is already installed to avoid compatibility issues.

Docker can be obtained following the instruction at this link: (https://www.docker.com/get-started).

Once the installation is completed, please get the protocol image called 'dcblab/virmutsig_img' by digiting the following command in a terminal:

```
>docker pull dcblab/virmutsig_img:latest
```

3. Verify docker image installation

To test if the image is correctly installed in your system, please execute the following code:

```
>docker image ls
```

which should display the following lines:

```
REPOSITORY      TAG    IMAGE ID     CREATED    SIZE

dcblab/virmutsig_img  latest    223f27c12b45    4 hours ago  1.56GB
```

Note: in this case, it is possible to proceed to the next steps.

Part 2. Data preprocessing pipeline

Timing: 10 min download + 70 min pipeline execution for each sample (can be parallelized)

The data preprocessing pipeline included in the VirMutSig protocol is provided as a Nextflow pipeline file, named "SRR_to_SNVlist.nf" and included in the "preprocessing" folder of the GitHub repository: https://github.com/BIMIB-DISCo/VirMutSig.

The folder also includes a file named: "nextflow.config", which must be opportunely modified according to the specific experimental settings (see below for the parameter description).

The preprocessing pipeline includes the following processes (detailed in the following): data acquisition, trimming, alignment, remove duplicated reads (optional), get depth information, variant calling, and variant filtering.

As output, the preprocessing pipeline returns a file named: "SNV_list.txt" in which:

rows correspond to all the single nucleotide variants (SNVs) detected in all the samples of the dataset

columns correspond to: sample ID, variant ID, genome position, reference allele, alternative allele, reference three nucleotides (flanking bases), supporting reads, coverage, variant frequency (VF), and p-value (returned by the variant caller).

To run the preprocessing processes of the protocol, simply move to the "preprocessing" folder and execute the following command from the terminal:

```
>nextflow run SRR_to_SNVlist.nf
```

In the following, we describe the preprocessing processes and the related settings in detail.

4. Data acquisition
   From now on, we will consider 'preprocessing' as a working directory from the one specified during the installation: 'user_local_directory/VirMutSig/preprocessing'
   Input data can be either (*i*) downloaded from public repositories, or (*ii*) provided from local folders, if available.
   a. Input data acquisition from public repositories
      Input data can be downloaded from public repositories such as, e.g., the NCBI Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra). In this case, one can use the SRA Run Selector web interface (https://www.ncbi.nlm.nih.gov/Traces/study/) to search for a dataset and download the "Accession List" by using the related button. The obtained list of SRR IDs (one per line) can be passed as input to download the files via SRA-toolkit, by editing the following parameter of the "nextflow.config":

```
params.FASTQ_input = '../example/SRAlist_paired.txt'
```

      Also in this case, the library preparation layout (single-end or paired-end) must be specified by editing the following parameter in "nextflow.config" file:

```
params.library_preparation = 'paired' // or 'single'
```

      By default, the downloaded FASTQ files will be stored in the following path: '/intermediate/FASTQ' (relative to SRR_to_SNVlist.nf file). It possible to change the directory by editing the following parameter in in "nextflow.config" file:

```
params.FASTQdir = 'intermediate/FASTQ'
```

   b. Input data from local folders

In this case, the user must indicate the directory where the FASTQ files are located, editing the following parameter of the "nextflow.config":

```
params.FASTQ_input = '/Path/To/Directory/'
```

The library preparation layout (single-end or paired-end) must be specified by editing the following parameter in "nextflow.config" file:

```
params.library_preparation = 'paired' // or 'single'
```

Notice that with the "paired" configuration two FASTQ files are expected for each sample, formatted as (sampleID_1.fastq.gz and sampleID_2.fastq.gz).

Conversely, with the "single" configuration one sampleID.fastq.gz file is expected for each sample.

5. Trimming

This step aims at removing from the sequence the nucleotides with low sequencing quality. To perform this step, the Nextflow pipeline exploits the tool Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic).

To tune the additional parameters used by Trimmomatic please edit the following parameter in "nextflow.config" file:

```
params.trimmomatic_setting = 'LEADING:20 TRAILING:20
SLIDINGWINDOW:4:20 MINLEN:40'
```

In brief, LEADING and TRAILING parameters cut the bases that display a quality below a certain threshold, respectively at the start and the end of a read (20).

The SLIDINGWINDOW parameter sets the size of a sliding window (4 in our example) and deletes all the bases from the leftmost position of the window to the end of the read, when the average quality detected in the window drops below a given threshold (e.g., 20).

Finally, the MINLEN parameter specifies the minimum length of a read to be kept (40 bases in our example). Please, refer to the Trimmomatic documentation for more details.

6. Alignment

   Reads need to be aligned to a reference genome, which can be selected by the user, e.g.,

   a. SARS-CoV-2-ANC (Ramazzotti et al., 2021),
   b. EPI_ISL_405839 / GeneBank ID: MN975262.1 (https://www.ncbi.nlm.nih.gov/nuccore/MN975262.1) (Bastola et al., 2020),
   c. EPI_ISL_402125 / NCBI ID: NC_045512.2 (https://www.ncbi.nlm.nih.gov/nuccore/1798174254) (Andersen et al., 2020).

In the subfolder "preprocessing/reference" of the GitHub repository, we provide the SARS-CoV-2-ANC genome in FASTA format.

The user can specify the reference genome file by editing following parameter in "nextflow.config" file:

```
params.fasta = 'reference/SARS-CoV-2-ANC.fasta'
```

The alignment is performed with BWA-MEM (https://github.com/lh3/bwa), which generates a SAM file including the aligned reads. All the associated files used by the BWA aligner will be automatically generated and placed in the same reference genome folder.

Each SAM file will be sorted and compressed into a BAM file with Samtools (http://www.htslib.org/).

By default, the BAM files will be stored in the following path: '/intermediate/BAM' (relative to SRR_to_SNVlist.nf file). It is possible to change the directory by editing the following parameter in in "nextflow.config" file:

```
params.BAMdir = 'intermediate/BAM'
```

7. Remove duplicated reads (optional)

Often, after obtaining an aligned BAM file, is useful to mark and remove duplicated reads to reduce the impact of the amplification bias, especially with RNA-seq experiments. Otherwise, when dealing with Amplicon data, several duplicated reads are expected and should not be removed. For this reason, we made this step optional.

To include or skip this step in the preprocessing pipeline, please set accordingly the following parameter in "nextflow.config" file:

```
params.remove_duplicates = ''true'' // or ''false''
```

Note: the tool used to perform this task is Picard (https://broadinstitute.github.io/picard/).

8. Get depth information

The aim of this step is to obtain the depth information for each sample in every genome position (i.e., number of reads mapped on each position of the viral genome).

To perform this task, we used Samtools.

By default, the coverage files will be stored in the following path: '/intermediate/COVERAGE' (relative to SRR_to_SNVlist.nf file). It is possible to change the directory by editing the following parameter in in "nextflow.config" file:

```
params.COVERAGEdir = 'intermediate/COVERAGE'
```

9. Variant calling

Variants can be called comparing the aligned reads with the reference genome. To do so, we used Samtools (http://www.htslib.org/) and VarScan (http://varscan.sourceforge.net/). More in detail, we used the mpileup command included in Samtools which converts the BAM file into the pileup format required by VarScan.

Then we used the VarScan pileup2snp for variant calling to produce a VCF file for each BAM file. The VarScan setting can be adjusted by editing the following parameter included in "nextflow.config" file.

```
params.varscan = '–min-var-freq 0.01 –p-value 1'
```

Note: For more information, please refer to the VarScan documentation.

By default, the VCF files will be stored in the following path: '/intermediate/VCF' (relative to SRR_to_SNVlist.nf file). It is possible to change the directory by editing the following parameter in in "nextflow.config" file:

```
params.VCFdir = 'intermediate/VCF'
```

Note: We suggest performing the alignment step according to the manufacturer's recommendations for all sequencing technologies.

⚠ CRITICAL: All the above steps can be performed also by applying different pipelines for variant calling, such as the one proposed in [https://github.com/andersen-lab/ivar], which was specifically designed for handling viral amplicon data obtained via the *artic protocol*.

10. Variant filtering
    a. In order to reduce the impact of noise in the data (due, e.g., to sequencing issues), we adopted multiple quality control (QC) filters to select only the reliable SNVs. The last step of the preprocessing pipeline applies different filtering criteria on the detected SNVs. The following parameters determine the filtering threshold, and they can be set by changing the corresponding numeric value included into a string assigned to VirMutSig_QCfilter parameter present in the "nextflow.config" file.

```
params.SNV_filters = 'PV_THR:0.01 VAR_FREQ_THR:0.05 MIN_COV:20 ALT_READ_THR:3'
```

    i. p-value on variant calling significance (PV_THR). The filter removes all the variants called with a significance p-value larger than a given threshold (default = 0.01).
    ii. Frequency threshold (VAR_FREQ_THR). The filter keeps only variants observed in the data with a variant frequency that exceed the specified threshold (default = 0.05).
    iii. Minimum coverage (MIN_COV). The filter keeps only variants with the specified minimum coverage (default = 20).
    iv. Minimum alternative read count (ALT_READ_THR). The filter keeps only variants with a minimum number of reads showing the alternative allele equal to the given threshold. (default = 3)

Note: Indels are not considered in the analysis.

    b. By default, the SNV_list.txt file will be stored in the 'example' directory. It is possible to change the directory by editing the following parameter in in "nextflow.config" file:

```
params.SNVlistdir = '../example
```

This step uses a R script included in the "preprocessing/bin" directory of the GitHub repository named makeSNVlist.R. It takes as input different arguments with the following fixed order:

    i. A string with the path of the directory containing the VCF files generated by the variant caller.
    ii. A string with the path of the directory containing the depth files generated by step 8. Such files must end with '.depth.txt'.
    iii. Reference file in fasta format. (e.g., SARS-CoV-2-ANC.fasta)

Note: If another pipeline has been used to perform variant calling, the R script can be used to aggregate the vcf files into a SNV list. Instead, if VirMutSig preprocessing steps have been used, the script is automatically executed via Nextflow.

⚠ CRITICAL: Parameters must be set according to the specific features of the datasets (see, e.g., the guidelines proposed on the website: https://virological.org/).

11. Further information

The Trimming and Alignment step require greater computational resources than the other processes. For this reason, it is possible to specify the maximum number of cores to be used, by changing the cpus value in the following setting of the "nextflow.config" file:

```
process {
     withName: 'Trimming_single' {cpus = 4}
     withName: 'Trimming_paired' {cpus = 4}
     withName: 'Alignment_and_sorting_single' {cpus = 8}
     withName: 'Alignment_and_sorting_paired' {cpus = 8}
     }
```

An increase of the cpus number assigned to the processes reduces the computational time required to trim and align the reads, but it also reduces the number of samples that can be analyzed in parallel. For these reasons, one should select a proper value based on the available computational resources and number of samples.

All parameters ("params.") can be specified by overriding when the pipeline is launched. To do so, please specify the corresponding parameter name (the string following "params.") and specify the opportune argument of the Nextflow command.

In the example, the SRR list file path and the output directory path of the SNVs list file will be specified without editing the "nextflow.config" file.

```
>nextflow run SRR_to_SNVlist.nf \
-FASTQ_input 'SRRfile/custom/path' \
-SNVlistdir 'SNVlist/output/path'
```

For a summary of the settings and processes executed with the example configuration please see Figure 1.

Part 3. *De novo* discovery of mutational signatures (denovo.R)

Timing: ~20 min (approximate time required to perform the step on a dataset with ~100 samples and ~10,000 mutations, with a significant variability related to the signature rank range)

12. Input files and setup
    The denovo.R script requires two input files to be executed.
    a. List of selected SNVs [SNV_list.txt]
       This file is generated following the preprocessing step. It is a semicolon-separated file with a SNV for each line. It includes (at least) the following column headers:
       i. SampleId: The ID of the sample.
       ii. Position: The genome position.
       iii. Reference: the reference allele (A, T, C, G).
       iv. Alternative: the alternative allele (A, T, C, G).
       v. VariantCount: the number of the reads including the alternative allele.
       vi. TotalCount: the total number of reads covering the genome position.
       For the 96-contexts analysis (see below) the following column is also required:
       vii. ReferenceTrinucleotide: the triplet with the reference bases before and after the variant position (e.g., ATC where T is the reference).
       Please see the file SNV_list.txt contained in the "/example" directory for an example of input formatting (see Table 1).
    b. Configuration file [denovo_config.yaml]

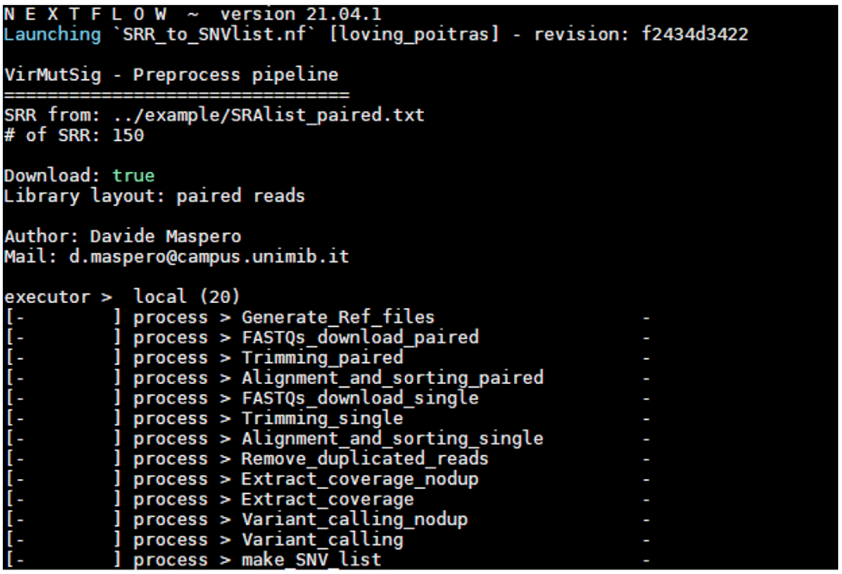


Figure 1. Summary of the settings and processes executed with the example configuration

We analyzed 150 samples with a paired-end library layout, downloaded from the SRA database. The related processes are automatically executed via Nextflow.

The parameters to run denovo.R are set in a text file called as default 'denovo_config.yaml'. The file must contain a parameter in each line formatted as:

```
PARAMETER NAMES: value
```

The parameters that can be modified are the following:

[variant selection]

i. CLONAL_SNV_THR: Double [0,1]. default = 0.9.
   This parameter defines the variant frequency threshold that determines whether a given variant is clonal.

ii. MINOR_SNV_SEL: String {'always', 'all'}. default = 'always'.
   To reduce the bias induced by mutations transmitted and inherited in the population during the epidemic spread, for the signature analysis we select and employ only the SNVs that are never observed with a frequency higher than *CLONAL_SNV_THR* in any sample. The selected SNVs are defined as '*always minor*'. This parameter together with the parameter CLONAL_SNV_THR are critical as they determine the variants that are selected for the signatures analysis; the default parameters that we suggest here (i.e., CLONAL_SNV_THR = 0.90 and MINOR_SNV_SEL = "always") are very conservative and they should be determined based on the aim of the study. For further details, please refer to (Ramazzotti et al., 2021).

iii. MAX_SNV_SAMPLE: Integer [1, Inf]. default = 100.

Table 1. First 5 lines of the SNV_list.txt file imported in Rstudio

SampleID	VariantID	Position	Reference	Alternative	Reference Trinucleotide	VariantCount	TotalCount	Variant frequency	Pvalue
SRR12661198	9996_C_T	9996	C	T	TCA	494	5720	0.0864	2.85E-144
SRR12661096	9965_C_T	9965	C	T	TCT	68	1263	0.0538	4.82E-20
SRR12833654	9960_G_T	9960	G	T	TGT	49	641	0.0764	6.85E-16
SRR12661080	995_C_T	995	C	T	ACG	165	2618	0.063	5.54E-48
SRR12661132	9945_G_T	9945	G	T	AGA	41	389	0.1054	1.50E-13

In order to reduce the impact of highly mutated samples, likely due to degradation of their biological isolation, we suggest removing the samples exhibiting more than a user selected number of variants.

iv. MIN_SNV_SAMPLE: Integer [1, Inf]. default = 6.
To assess the existence of statistically significant mutational signatures, we remove all the samples with number of detected SNVs lower than this value.
[analysis parameters]

v. NUM_CONTEXTS: Integer {6, 96}. default = 6.
This parameter specifies the number of different nucleotide substitution types (based on the flanking bases) that are considered.
6 indicates that no flanking bases are considered. In this case, the possible substitution types are: C>A (or G>T), C>G (or G>C), C>T (or G>A), T>A (or A>T), T>C (or A>G), T>G (or A>C).
96 indicates that the two flanking bases are considered. In this case, the possible substitution types include, e.g.,: ACA>AAA (or AGA>ATA), ACG>AAG (or AGG>ATG), ATA>ACA (or AAA>AGA), etc. For further details, please refer to (Lal et al., 2021).

vi. MIN_NUM_SIG: Integer [1, *NUM_CONTEXTS*]. default = 1.
This parameter sets the lower bound of the range for the number of distinct signatures to be searched.

vii. MAX_NUM_SIG: Integer [*MIN_NUM_SIG*, *NUM_CONTEXTS*]. default = 6
This parameter sets the upper bound of the range for the number of distinct signatures to be searched.

viii. NMF_ITER: Integer [1, inf]. default = 100.
This parameter sets the number of Negative Matrix Factorization iterations performed during the inference of **A** and **B** matrices

ix. SEED: Integer [0, Inf]. default = 0 (with 0 the seed will be randomly set).
This parameter initializes the random number generator. It is used to obtain reproducible results by keeping the same SEED.

x. N_CORE: Integer [0, Inf]. default = 0 (with 0 the number of cores available will automatically be detected).
This parameter indicates the maximum number of computational units (cpus) available for the inference.

⚠ CRITICAL: MAX_NUM_SIG and MIN_SNV_SAMPLE must be set accordingly with the NUM_-CONTEXTS parameter. To obtain robust results, we suggest setting the former equal to half the number of contexts (i.e., 3 or 48) and the latter larger than the number of contexts.

13. Run denovo.R
    a. denovo.R can be run with the following command:

```
>Rscript denovo.R \
-SNV_list path/to/SNV_list.txt \
-config_file path/to/denovo_config.txt \
-output_dir path/to/output/dir
```

The script will generate in the output directory (specified by the user) the following three subdirectories:

i. "Signatures":
This directory will contain a file for each number of searched signatures named 'x_signatures.txt'. Each file will be formatted as a table with the context as header and a row for each signature.

ii. "Graphical":
This directory will contain the graphical representation of each signature set found in the data.
iii. "Rank":
This directory will contain a set of files to determine the optimal number of signatures.

b. Unfortunately, no automated procedures are available to find the correct number of signatures, so we here provide a set of metrics to determine it. The denovo.R script provides four metrics as output and the relative plots:
i. explained variance (Hutchins et al., 2008)
ii. cophenetic coefficient (Brunet et al., 2004),
iii. dispersion coefficient (Kim and Park, 2007),
iv. silhouette consensus coefficient.

The four metrics are shown in Figure 2 and Table 2.

The first one measures how good is the fit of each given number of signatures, considering the observed mutational profiles. This metric is useful to estimate when the number of signatures is too high (i.e., overfitting). To this end, we suggest choosing the optimal number of signatures based on the *bend rule* that selects the number of signatures corresponding to the elbow in the explained variant plot (see the example in Figure 2A).

Often there is uncertainty to select a unique *elbow point* on the plot, so the latter three metrics can be used as additional evaluation of the optimal rank. Roughly, they provide a measure of stability of the NMF solutions over multiple runs (*NMF_ITER* parameter). These coefficients range from 0 to 1 and higher values indicate higher consistency among NMF solutions.

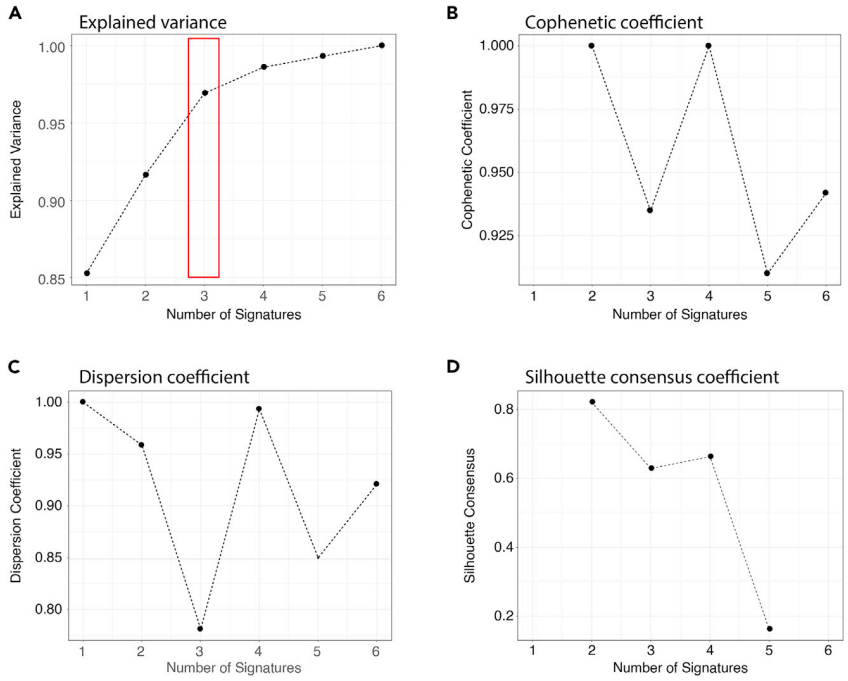


Figure 2. Metrics to evaluate the best number of signatures
(A) shows the explained variance at different ranks; in this case we observed a bend at 3.
(B–D) show stability-based coefficients which report the consistency of NMF solutions across multiple iterations.

Table 2. Metric coefficients for each different signature rank searched via denovo.R script

Rank	Cophenetic coefficient	Dispersion coefficient	Silhouette consensus	Explained variance
1	*NA*	1	*NA*	0.85
2	1	0.96	0.82	0.92
3	0.93	0.78	0.63	0.97
4	1	0.99	0.66	0.99
5	0.91	0.85	0.16	0.99
6	0.94	0.92	*NA*	1

Considering all the above, our suggestion is to select one or few elbow points and among them select the one corresponding to higher cophenetic, dispersion, or silhouette coefficient. If still there is ambiguity it is reasonable to take the lower one.

Considering the example, and the metrics shown in Figure 2, we conservatively selected 3 as the optimal number of signatures, as at this rank we have a first bend in the explained variance and high values of the other metrics.

Note: Please get from the 'Signatures' directory the corresponding file that should be then used in the following assignment step.

Note: Each signature file contains a different number of signatures (identified with SIG_NUM) found in the data. In the 'graphical' directory, denovo.R generates pdf files with the same name as the corresponding signatures set. Those files contain a plot of the categorical distributions.

Part 4. Assignment of existing mutational signatures (assignment.R)

Timing: 15 min (approximate time required to perform the step on a dataset with ~100 samples and ~10,000 mutations, with significant variability related to bootstrap iterations)

14. Input files and setup
    The assignment.R script requires three input files to be executed.
    a. List of selected SNVs [SNV_list.txt]
       This file is the same used in Part 3, please see above.
    b. Signatures file [x_signatures.txt]
       This file contains for each signature the frequency of substitutions. Their values are grouped by context and normalized up to 1. The file must be formatted as a table with the context as header and a row for each signature.
       The header must be formatted as: G>T:C>A;G>C:C>G;G>A:C>T;A>T:T>A;A>G:T>C;A>C:T>G.
       Notice that, for instance, variants from G to T and C to A are aggregated together in the cases of 6-context. For further details, especially for the 96-contexts representation, please refer to (Alexandrov et al., 2013).
       This file can be generated via Part 3 or can be directly passed by the user.
    c. Configuration file [assignment_config.yaml]
       The parameters to run assignment.R are set in a text file called as default 'assignment_config.txt'. The file must contain a parameter in each line formatted as PARAMETER_NAMES: value
       The parameters that can be modified are the following:
       [variant selection]
       i. CLONAL_SNV_THR: Double [0,1]. default = 0.9. This parameter defines whether a given SNV is considered as clonal (default = 0.9).

ii. MINOR_SNV_SEL: String {'always', 'all'}. default = 'always'. To reduce the bias induced by the mutation transmitted and inherited among the population during the epidemic spread we select for the signature analysis only SNVs never observed with a frequency higher than CLONAL_SNV_THR in any sample. The remaining SNVs are defined as 'always minor'. For further details, please refer to (Graudenzi et al., 2021).
iii. MAX_SNV_SAMPLE: Integer [1, Inf]. default = 100. In order to reduce the impact of highly mutated samples likely due to degradation of their biological isolation. We suggest removing the samples exhibiting more than a user selected number of variants.
iv. MIN_SNV_SAMPLE: Integer [1, Inf]. default = 6. In order to assess the existence of statistically significant mutational signatures we have to remove all the samples with less than a given number of detected SNVs.
[analysis parameters]
v. BOOTSTRAP: String {'yes', 'no'}. default = 'yes'.
vi. GOODNESS_FIT_THR: Double [0.5,1]. default = 0.95. This threshold indicates the minimum level of goodness of fit until when keep adding signatures (among the signatures.txt file) to the fit, in a given sample. The goodness of fit is measured with the cosine similarity between observed and predicted counts in each sample.
vii. MIN_SIG_FREQ: Double [0,1]. default = 0.05. Each signature is considered to be present in a given sample if its activity (alpha value) is greater than the value of this parameter.
viii. P_VALUE: threshold to be used when assessing significance of the exposure of samples to signatures. To this extent, Mann–Whitney U test is performed whose results are evaluated with the given P_VALUE threshold. default = 0.05
ix. NUM_ITER: Integer [1, Inf]. default = 100
x. SEED: Integer [0, Inf]. default = 0 (with 0 the seed will be set randomly)
xi. N_CORE: Integer [0, Inf]. default = 0 (with 0 the number of core available will automatically detected)

⚠ CRITICAL: We suggest setting the parameters in accordance with the number of contexts. Similar to Part 3 MIN_SNV_SAMPLE should be larger than the number of contexts to provide reasonable results.

15. Run assignment.R

assignment.R can be run by the following command:

```
>Rscript assignment.R \
-SNV_list path_to_SNV_list.txt \
-signature path_to_signature_x.txt \
-config_file path_to_assignment_config.txt \
-output_file path_to_output_file
```

The script will generate a table in a text file specified by the user (default: assignment_result.txt). This table will have a row for each selected sample and one column for each signature, called "Sn_exposure". This column reports the alpha values of each signature found in each sample (i.e., a numeric value measuring the level of activity of the signature in that sample).

If the bootstrap procedure is performed, another column for each signature is added into file. This column, called 'Sn_pvalue', shows the p-value of significance of observing each signature in a given sample obtained with the harmonic mean of the one-sided (greater) Mann–Whitney U-test.

A

signature	C>A:G>T	C>G:G>C	C>T:G>A	T>A:A>T	T>C:A>G	T>G:A>C
S1	0.0000000	0.03817730	0.8550846	0.02318943	0.06242990	0.02111873
S2	0.6012669	0.06115096	0.1772178	0.06814967	0.04531042	0.04690416
S3	0.0000000	0.06397831	0.0000000	0.23301743	0.61005199	0.09295227

B

Figure 3. Output of the denovo.R script

(A) The table reports the frequency of the nucleotide substitutions for each context.

(B) The same values are represented with bar-plots. The figure is saved as 3_signatures.pdf file located in "VirMutSig/example/denovo_results/signature/figures/".

EXPECTED OUTCOMES

denovo.R

This step provides as output:

The signature matrix (B) for each considered number of signatures and the corresponding graphical visualization (see Figure 3). These files are stored into the "signatures" directory.

The metrics to estimate the optimal number of signatures in a text file ("metric_coefficients.txt") and the corresponding plots (see Figure 2 and Table 2). These files are stored into the "rank" directory.

Among them the user should select the optimal solution ("signatures/files/x_signatures.txt") and use it as input for the following step.

assignment.R

This step returns a matrix of the activity of each signature assigned to each sample, saved as a text file ("assignment_result.txt"). See Table 3.

QUANTIFICATION AND STATISTICAL ANALYSIS

The produced signature-assignment matrix ("assignment_result.txt") can be used as input in downstream analyses.

For example, as shown in (Graudenzi et al., 2021), the following tasks can be performed:

Table 3. First 5 entries of the assignment_results.txt file obtained after assignment.R execution

Sample	S1_exposure	S2_exposure	S3_exposure	S1_pvalue	S2_pvalue	S3_pvalue
SRR12351622	4.63	0.13	1.92	1.98e-18	1	1.98e-18
SRR12351634	1.14	0.44	4.48	1.98e-18	1.98e-18	1.98e-18
SRR12351645	0.81	1.84	3.68	1.98e-18	1.98e-18	1.98e-18
SRR12351651	2.32	0.47	1.05	1.98e-18	1.98e-18	1.98e-18
SRR12351655	4.65	0.26	2.06	1.98e-18	1	1.98e-18

For each sample, an exposure value is assigned for each signature. The corresponding harmonic mean p-value of the one-sided Mann–Whitney U-test is computed with a bootstrap procedure.

1. Stratification of samples into signature-based clusters

Samples can be stratified by applying k-means (or any other clustering methods) on the signature-assignment matrix. Standard heuristics should be employed to determine the optimal number of clusters.

2. Corrected-for-signatures dN/dS analysis

dN/dS analysis is a standard population genetics method to assess the selection pressure acting on the virus. After the identification of viral mutational signatures, it is possible to investigate whether the SNVs generated by the corresponding mutational process are positively or negatively selected in the population. To this end, the dN/dS analysis must be corrected for the expected mutation frequency specific for each signature profiles, as proposed in (Graudenzi et al., 2021).

LIMITATIONS

Reference genome

Any reference genome can be used with VirMutSig. Using different reference genomes may slightly change the results. For SARS-CoV-2 analysis we highlight the presence of two other reference genomes besides SARS-CoV-2-ANC used in this protocol (see "reference genome" in "materials and equipment" section):

EPI_ISL_405839 (Bastola et al., 2020)

EPI_ISL_402125 (Andersen et al., 2020)

Notice that the number of mismatches between those two reference genomes and the one SARS-CoV-2-ANC is only 5 nucleotides.

Dataset quality

In these kinds of analyses using good quality data is mandatory because the level of random mutations (sequencing artifacts or samples biological degradation) could affect the significance of the signature identified and assigned.

However, we suggested parameter settings that can mitigate this issue. Depending on the specific quality of the data, the stringency of such parameters could be increased.

TROUBLESHOOTING

Problem 1

It may be difficult to choose the optimal rank (i.e., the number of mutational signatures) on the basis of the available metrics (see step 13).

Potential solution

The estimation of the optimal number of signatures is a critical task. We have already discussed in the text how one should consider the results of different metrics to estimate the optimal rank. However, such metrics may sometimes lead to ambiguous (or conflicting) indications, making it difficult to select the optimal rank. In this case, one may try to improve the stability of the results by increasing the number of NMF iterations, since a higher number of iterations is expected to deliver more reliable and stable results. To this end, please increase the value of the NMF_ITER parameter in the 'de-novo_config.yalm' file.

As an alternative option, we suggest increasing the number of samples in the datasets, if possible.

Problem 2

The input file for both the *de novo* and the *assignment* analyses is a matrix with samples (rows) x mutations (columns) that is automatically generated from the provided list of SNVs. If the mutations kept after the quality filter are too few, the resulting matrix might be too sparse to obtain robust results (see parts 3 and 4).

Potential solution

The de novo extraction of mutational signatures is harder with very sparse matrices. If the input data are too sparse, one may increase the number of mutations by loosening the quality check filters or lowering the threshold to filter out fixed variants. In both cases, more mutations will be employed in the analysis.

However, one should be cautious when relaxing quality control filters and should aim at preserving a good tradeoff between the number of variants and the quality of the data, because including lower quality mutations may reduce the robustness of the results and the reliability of the inference.

Problem 3

An issue similar to the one presented in Problem 2 may arise when the number of samples that satisfies the quality criteria (e.g., minimum number of mutations) is too low (see parts 3 and 4).

Potential solution

Increasing the number of samples is extremely beneficial for the de novo inference of mutational signatures. We provide a set of quality checks parameters to set the minimum and maximum number of mutations used to filter out samples (i.e., MIN_SNV_SAMPLE, MAX_SNV_SAMPLE). One may set these parameters to increase the sample size of the datasets, but as stated in the previous point, one should keep in mind that including lower quality samples may also reduce the quality of the inference.

Problem 4

After performing the bootstrap (see part 4) no significant signatures were assigned to the samples.

Potential solution

In the bootstrap procedure, the algorithm detects the signatures that are significantly exceeding a given exposure for each sample. The method repeats the fit until a given 'GOODNESS_FIT_THR' threshold is reached (to reduce overfitting) and identifies the signatures significantly exceeding 'MIN_SIG_FREQ' percentage of mutations for each sample. Reducing these two parameters will result in a larger number of significant signatures assigned to each sample. A correct tuning of such parameters depends on the quality of the data and on the minimum number of mutations which are considered to be significant to assess if a signature is active.

Problem 5

The user may face issues with software compatibility during the setup of VirMutSig or one needs to execute it on a computer with a Windows operating system (see parts 1, 2, 3, and 4). .

Potential solution

We provide a Docker image (see part 1) including all the VirMutSig scripts and data already preinstalled and tested. It is also possible to execute all the VirMutSig steps within the Docker image even on a computer with Windows OS.

To do this, please start by selecting a directory (e.g., "C:\user\data") on the local machine and copy all the files requested for the user analysis (e.g., SRA_list or fastq files, reference genome of choice, etc.). This directory will be the only local directory available within the Docker image.

After obtaining the dcblab/virmutsig_img, the user will need to access it by executing the following command in a terminal or in the command prompt window.

```
>docker run -it –mount \

type=bind,source="C:\\user\\data'',target=/VirMutSig/UserData/ \

dcblab/virmutsig_img
```

Please, change "C:\user\data" with the absolute path of the selected directory.

The terminal will changce, and the username will be replaced by something similar to:

```
root@d55f578b7306:/#
```

Now, it is possible to execute all the protocol steps within the new terminal, following the same instructions described above.

The user files are located in the /VirMutSig/UserData directory. To get the VirMutSig result files please copy them into this location, and they will also be available in "C:\user\data\" local directory.

All the files in VirMutSig_img are writable. For example, the config files (i.e., "nextflow.config", "denovo_config.yalm", and "assignment_config.yalm") can be modified using the "nano" editor with the following command:

```
root@d55f578b7306:/# nano /VirMutSig/denovo/denovo_config.yaml
```

Once completed the analyses, please digit exit to close the VirMutSig image. Notice that, only the files modified or created within the /VirMutSig/UserData directory will be permanently available (in C:\user\data), while all other edits will be discarded.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Daniele Ramazzotti (daniele.ramazzotti@unimib.it).

Materials availability

This study did not generate new unique reagents.

Data and code availability

This study did not generate any unique data sets. The FASTQ files used in this protocol are public available at NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) using the following access numbers NCBI: PRJNA610428, PRJNA645906

The VirMutSig protocol and the example files are available at: https://github.com/BIMIB-DISCo/VirMutSig.

ACKNOWLEDGMENTS

This work was partially supported by the Elixir Italian Chapter and the SysBioNet project, a Ministero dell'Istruzione, dell'Università e della Ricerca initiative for the Italian Roadmap of European Strategy Forum on Research Infrastructures, and by the Associazione Italiana per la Ricerca sul Cancro (AIRC)-IG grant 22082. D.R. and F.A. were partially supported by a Bicocca 2020 Starting Grant. D.R. was supported by a Premio Giovani Talenti dell'Università degli Studi di Milano-Bicocca. We thank Marco Antoniotti, Giulio Caravagna, Chiara Damiani, Lucrezia Patruno, and Francesco Craighero for helpful discussions.

AUTHOR CONTRIBUTIONS

Writing – original draft and writing – review & editing, D.M., D.R., and A.G.; software, D.M., D.R., and F.A.; validation, D.M., D.R., F.A., and A.G.; funding acquisition, A.G., D.R., D.P., R.P.; supervision, A.G. and D.R. All authors read, revised, and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V.,., and Stratton, M.R. (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., and Garry, R.F. (2020). The proximal origin of SARS-CoV-2. Nat. Med. *26*, 450–452.

Auguie, B. (2017). gridExtra: Miscellaneous Functions for "Grid" Graph- Ics. R Package Version 2.3. https://CRAN.R-project.org/package=gridExtra.

Bastola, A., Sah, R., Rodriguez-Morales, A.J., Lal, B.K., Jha, R., Ojha, H.C.,., and Pandey, B.D. (2020). The first 2019 novel coronavirus case in Nepal. Lancet Infect. Dis. *20*, 279–280.

Bates, D., and Maechler, M. (2021). Matrix: Sparse and Dense Matrix Classes and Methods. R Package Version 1.3-4. https://CRAN.R-project.org/package=Matrix.

Brunet, J.P., Tamayo, P., Golub, T.R., and Mesirov, J.P. (2004). Metagenes and molecular pattern discovery using matrix factorization. PNAS *101.12*, 4164–4169.

Dowle, M., and Srinivasan, A. (2021). data.table: Extension of 'data.Frame'. R Package Version 1.14.0. https://CRAN.R-project.org/package=data.table.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for Generalized linear models via coordinate descent. J. Stat. Softw. *33*, 1–22.

Gagolewski, M. (2021). Stringi: Fast and Portable Character String Processing in R. R Package Version 1.7.4. https://stringi.gagolewski.com/.

Gaujoux, R., and Seoighe, C. (2020). The Package NMF: Manual Pages. R Package Version 0.23.0. CRAN. https://cran.r-project.org/package=NMF.

Graudenzi, A., Maspero, D., Angaroni, F., Piazza, R., and Ramazzotti, D. (2021). Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity. iScience *24*, 102116.

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S.,., and Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods *12*, 115–121.

Hutchins, L.N., Murphy, S.M.,, Singh, P., and Graber, J.H. (2008). Position-dependent motif characterization using non-negative matrix factorization. Bioinformatics *24*, 2684–2690.

Kim, H., and Park, H. (2007). Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics *23*, 1495–1502.

Lal, A., Liu, K., Tibshirani, R., Sidow, A., and Ramazzotti, D. (2021). De novo mutational signature discovery in tumor genomes using SparseSignatures. PLoS Comput. Biol. *17*, e1009119.

Merkel, D. (2014). Docker: lightweight Linux containers for consistent development and deployment. Linux J. *2014*, 2.

Morgan, M. (2021). BiocManager: Access the Bioconductor Project Package Repository. R Package Version 1.30.16. https://CRAN.R-project.org/ package=BiocManager.

Mullen, K.M., and van Stokkum, I.H.M. (2012). Nnls: The Lawson- Hanson Algorithm for Non-negative Least Squares (NNLS). R Package Version 1.4. https://CRAN.R-project.org/package=nnls.

Ramazzotti, D., Angaroni, F., Maspero, D., Gambacorti-Passerini, C., Antoniotti, M., Graudenzi, A., and Piazza, R. (2021). VERSO: a comprehensive framework for the inference of robust phylogenies and the quantification of intra-host genomic diversity of viral samples. Patterns *2*, 100212.

Stephens, J. (2020). Yaml: Methods to Convert R Data to YAML and Back. R Package Version 2.2.1. https://CRAN.R-project.org/package=yaml.

Davis, T.L. (2020). Optparse: Command Line Option Parser. R Package Version 1.6.6 (https://CRAN.R-project.org/package=optparse).

Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nat. Biotechnol. *35*, 316–319.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag). https://ggplot2.tidyverse.org.

Wild, Fridolin (2020). Lsa: Latent Semantic Analysis. R Package Version 0.73.2. https://CRAN.R-project.org/package=lsa.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 28 May 2020
doi: 10.3389/fbioe.2020.00523

An Optimal Control Framework for the Automated Design of Personalized Cancer Treatments

Fabrizio Angaroni[1†], *Alex Graudenzi*[1,2*†], *Marco Rossignolo*[3,4], *Davide Maspero*[1,2,5], *Tommaso Calarco*[6], *Rocco Piazza*[7,8], *Simone Montangero*[4,9‡] *and Marco Antoniotti*[1,10‡]

[1] Department of Informatics, Systems and Communication, University of Milan-Bicocca, Milan, Italy, [2] Institute of Molecular Bioimaging and Physiology, Consiglio Nazionale delle Ricerche (IBFM-CNR), Segrate, Milan, Italy, [3] Center for Integrated Quantum Science and Technologies, Institute for Quantum Optics, Universitat Ulm, Ulm, Germany, [4] Istituto Nazionale di Fisica Nucleare (INFN), Padova, Italy, [5] Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy, [6] Forschungszentrum Jülich, Institute of Quantum Control (PGI-8), Jülich, Germany, [7] Department of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy, [8] Hematology and Clinical Research Unit, San Gerardo Hospital, Monza, Italy, [9] Department of Physics and Astronomy "G. Galilei", University of Padova, Padova, Italy, [10] Bicocca Bioinformatics Biostatistics and Bioimaging Centre - B4, Milan, Italy

OPEN ACCESS

Edited by:
Paola Lecca,
Free University of Bozen-Bolzano, Italy

Reviewed by:
Bruno Carpentieri,
Free University of Bozen-Bolzano, Italy
Fabio Bagagiolo,
University of Trento, Italy

***Correspondence:**
Alex Graudenzi
alex.graudenzi@unimib.it

[†] *These authors have contributed equally to this work*

[‡] *These authors share senior authorship*

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *06 February 2020*
Accepted: *01 May 2020*
Published: *28 May 2020*

Citation:
Angaroni F, Graudenzi A, Rossignolo M, Maspero D, Calarco T, Piazza R, Montangero S and Antoniotti M (2020) An Optimal Control Framework for the Automated Design of Personalized Cancer Treatments. Front. Bioeng. Biotechnol. 8:523. doi: 10.3389/fbioe.2020.00523

One of the key challenges in current cancer research is the development of computational strategies to support clinicians in the identification of successful personalized treatments. Control theory might be an effective approach to this end, as proven by the long-established application to therapy design and testing. In this respect, we here introduce the Control Theory for Therapy Design (CT4TD) framework, which employs optimal control theory on patient-specific pharmacokinetics (PK) and pharmacodynamics (PD) models, to deliver optimized therapeutic strategies. The definition of personalized PK/PD models allows to explicitly consider the physiological heterogeneity of individuals and to adapt the therapy accordingly, as opposed to standard clinical practices. CT4TD can be used in two distinct scenarios. At the time of the diagnosis, CT4TD allows to set optimized personalized administration strategies, aimed at reaching selected target drug concentrations, while minimizing the costs in terms of toxicity and adverse effects. Moreover, if longitudinal data on patients under treatment are available, our approach allows to adjust the ongoing therapy, by relying on simplified models of cancer population dynamics, with the goal of minimizing or controlling the tumor burden. CT4TD is highly scalable, as it employs the efficient dCRAB/RedCRAB optimization algorithm, and the results are robust, as proven by extensive tests on synthetic data. Furthermore, the theoretical framework is general, and it might be applied to any therapy for which a PK/PD model can be estimated, and for any kind of administration and cost. As a proof of principle, we present the application of CT4TD to Imatinib administration in Chronic Myeloid leukemia, in which we adopt a simplified model of cancer population dynamics. In particular, we show that the optimized therapeutic strategies are diversified among patients, and display improvements with respect to the current standard regime.

Keywords: personalized therapy, optimal control theory, pharmacodynamics, pharmacokinetics, RedCRAB, chronic myeloid leukemia

1. INTRODUCTION

The increasing availability of reliable experimental data on cancer patients and the concurrent decreasing costs of computational power are fueling the development of algorithmic strategies for the automated generation of experimental hypotheses in cancer research. This is particularly relevant in the sphere of precision and personalized medicine, as efficient methods are urgently needed to make sense of available data and support clinicians in delivering patient-specific therapeutic strategies (Salgado et al., 2018). To this end, methods borrowed from optimal control theory (e.g., Bertsekas, 1995; Bailey and Haddad, 2005; Lenhart and Workman, 2007; Aström and Murray, 2010) can be employed in combination with efficient techniques for data analysis (Michor et al., 2005; Tang et al., 2011; Olshen et al., 2014; Rainero et al., 2018), to produce accurate predictive model of the clinical outcome of a given therapy in single cancer patients.

Here, we introduce a theoretical framework named CT4TD (Control Theory for Therapy Design), which employs the RedCRAB optimal control algorithm (Heck et al., 2018b; Omran et al., 2019), on patient-specific pharmacokinetics and pharmacodynamics (PK/PD) models (Welling, 1997), with the goal of delivering an optimized drug administration schedule (see **Figure 1** for a schematic representation of the framework).

In brief, PK models describe the temporal dynamics of the concentration of a given drug in a certain tissue or organ, whereas PD models depict the efficacy of the drug with respect to distinct concentration values. The CT4TD framework first defines patient-specific PK models based on demographic factors, such as, e.g., age, sex, and body weight. Such models are employed to automatically identify optimized therapy dosages and/or schedules to reach given target drug concentrations, as those commonly used in clinical practice, also by respecting any desired constraint such as, e.g., the maximum allowed number of doses per day or the maximum dosage. In this way, our framework can guide clinicians in the setting of optimized regimes at diagnosis, allowing for an either more or less aggressive tuning; this approach mimics the *steady state* optimization commonly proposed in pharmacological studies.

Furthermore, when longitudinal experimental data on tumor burden—e.g., the fraction of tumor cells on the total, in liquid tumors—are available for patients under standard treatment, CT4TD allows to determine optimized personalized strategies to be used in order to minimize or even eradicating the cancer cell subpopulation. In fact, with the CT4TD framework it is possible to fit experimental data with a hierarchical population dynamics model, which describes the temporal evolution of cancer subpopulations in a given tumor (Michor et al., 2005; Stiehl and Marciniak-Czochra, 2012; Werner et al., 2016; Stiehl et al., 2018). Such model allows to measure the impact of a given therapy over the tumor's ability to expand and develop and, accordingly, to estimate patient-specific PD models from clinical data, which are then employed to design optimized therapeutic regimes aimed at reducing the tumor burden.

Therefore, CT4TD can support clinicians in designing personalized therapies both at diagnosis and when longitudinal data on disease progression have become available. In all scenarios, with our approach it is possible to compare the actual therapeutic regime with the optimized one, showing improvements in terms of efficacy, toxicity, and overall costs.

The CT4TD theoretical framework is general and applicable to any kind of drugs, as long as PK/PD models can be retrieved or estimated. Yet, liquid tumors allow to safely adopt several simplifications and define simple and reliable models of population dynamics, avoiding possible complications due to the spatial and morphological properties of solid tumors (Graudenzi et al., 2014, 2018).

For this reason, in this work we apply the CT4TD to the specific case of Imatinib administration in patients with Chronic Myeloid leukemia (CML), and we show the advantages of employing our automated and data-driven framework in terms of increased efficacy of the therapy and reduction of the overall costs and toxicity. In particular, we here present the results of the application of the CT4TD framework in two ideally subsequent scenarios.

In the first case, CT4TD is used to identify the best therapeutic regime to reach selected target drug concentrations, as those commonly used in the clinic (e.g., Gambacorti-Passerini et al., 1997; Peng et al., 2005; Baccarani et al., 2014). This scenario provides indications which can be employed by clinicians at the time of the diagnosis. Importantly, the inclusion of demographic factors within the PK models (Widmer et al., 2006) allows to define personalized drug schedules that are different from standard practice. A *robustness analysis* to assess the impact of intra-patient variability and of possible systematic errors proves the safety of the hypotheses generated with our approach, especially with respect to possible technical or measurement errors.

In the second case, we employ longitudinal data on tumor burden of a selected cohort of CML patients under standard treatment, in order to retrieve personalized PD models and, accordingly, to identify an *adjusted* therapy that is most effective in minimizing cancer subpopulation, once the major molecular response has been observed. In both cases the results allow to explicitly evaluate the advantages in costs and improved efficacy with respect to standard therapies.

2. BACKGROUND

2.1. Pharmacokinetic and Pharmacodynamic Models

Pharmacokinetic (PK) models (Welling, 1997) are mathematical models that describe the temporal evolution of the concentration of a substance in a certain tissue of the body. Commonly used techniques to study such processes are the so-called *compartmental* models (Welling, 1997), i.e., dynamical models based on the law of conservation of mass, and which assume that the body is composed by a certain number of macroscopic coupled subsystem, namely compartments. Such models assume an instantaneous mixing of the drug in a compartment and a perfect transport among them, and are usually defined via systems of differential equations (e.g., Schwilden, 1981).

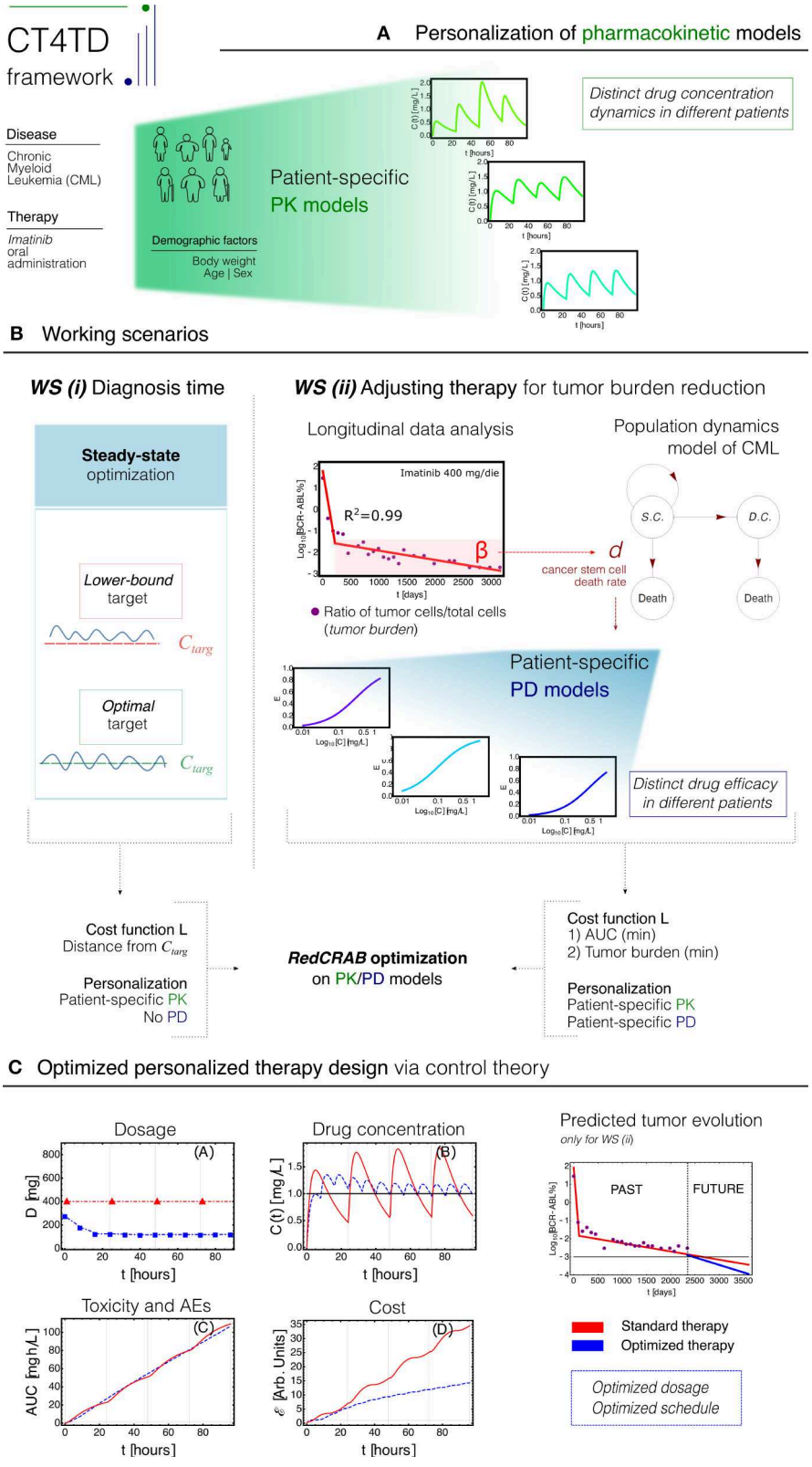


FIGURE 1 | CT4TD pipeline. **(A)** The CT4TD framework employs demographic factors such as body weight, age, and sex to define patient-specific parameters of the pharmacokinetic models. We here focus on the case of Imatinib administration in Chronic Myeloid leukemia (CML). **(B)** CT4TD manages two working scenarios: (i) at

(Continued)

FIGURE 1 | time of diagnosis, CT4TD can be used to reach given optimal/lower-bound drug concentration targets, e.g., from clinical studies (*steady state optimization*); (ii) when longitudinal data on tumor burden variation under standard therapy are available, CT4TD fits the data points with a hierarchical population dynamics model of CML, and this allows to estimate patient-specific pharmacodynamics (PD) parameters, based on the observed cancer cell death rate. In both scenarios, optimization on pharmacokinetics/pharmacodynamics (PK/PD) models is performed via RedCRAB, on distinct cost functions, aimed at: either being close to given target concentrations (and strictly larger in the lower-bound case)—WS (i); minimizing the Area Under the Curve (AUC) *and* the tumor burden—Working Scenario (WS) (ii). **(C)** optimized personalized dosage and schedule are returned, allowing to measure *in silico* the differences with respect to standard administration, in terms of dosage, drug concentration, cost, and AUC. WS (ii) allows to predict the tumor burden evolution in case of an optimized therapy.

The solution of such systems provide predictions about the variation of drug concentration in time, in a certain tissue. A limitation of PK models is the employment of coarse-grained oversimplifications, which require *ad-hoc* assumptions and are valid only for sufficiently long timescales.

Pharmacodynamic (PD) models (Welling, 1997; Rowland et al., 2011) study the relationship between the concentration of a drug and the resulting effect, in terms of efficacy and possible adverse effects (AEs). The effects of a certain substance are estimated by modeling relevant biochemical reactions, usually by exploiting the law of mass action (see, e.g., Goutelle et al., 2008). One of the major limitations of PD models is that it is usually impossible to have all the measurements necessary to determine the kinetic constants of the involved chemical reactions. For this reason, the efficacy of a drug is usually estimated with statistical methods and target concentrations are defined with respect to some arbitrary criteria (Peng et al., 2005; Larson et al., 2008a; Takahashi et al., 2010; von Mehren and Widmer, 2011; Baccarani et al., 2014).

PK/PD models are increasingly used to define new drug dosage guidelines and protocols (e.g., Peng et al., 2005). Nonetheless, standard approaches to this end are affected by several major issues. Usually the optimal dose is identified in phase I dose escalating clinical trials. Moreover, such trials may suffer from possible idiosyncrasies of the study, from the presence of unknown confounding factors and from the often limited sample size. Another problem is that the recommended dosage is often defined as optimal for an ideal—and non existing—*average* patient, because the efficacy is only defined statistically. As a consequence, the same drug dosage/schedule might be either insufficient or exceeding for different patients. In the former case, this might lead to a non-optimal clinical outcome, in terms of lower efficacy of the treatment, whereas in the latter case an excess of drug may raise the probability of AEs, as well as the economic cost of the therapy, an aspect that is particular relevant for oncological therapies (Fojo and Grady, 2009; Himmelstein et al., 2009; Experts in Chronic Myeloid Leukemia, 2013; Gomez-de León et al., 2017; Jabbour et al., 2017).

Therefore, effective strategies for the identification of optimized personalized therapy schedules are needed, in order to possibly reduce the amount of drug and minimize the probability of related adverse effects, while providing the same or an even better efficacy—i.e., clinical outcome—, with respect to the standard administration schedule. As a side effect, an optimized personalized schedule would also deliver a minimal economic cost, i.e., more patients will be able to afford its costs.

In this respect, CT4TD allows to: (i) define patient-specific PK models that depend on a number of demographic factors and biological covariates, such as age, sex, ethnicity, and body weight, as proposed by Widmer et al. (2006); (ii) estimate personalized PD models from longitudinal experimental data on tumor burden (if available). This allows to identify personalized therapeutic strategies, which explicitly account for the expected differences in PK and PD, due to the physiological heterogeneity of the individuals. It is important to stress that population PK/PD models are employed in a wide range of distinct diseases such, e.g., cancer (Yoshitsuga and et al., 2012), HIV (Chan et al., 2011), diabetes (Landersdorfer and Jusko, 2008), as well as in anesthesia administration (Potts et al., 2008).

2.2. Applications of Optimal Control Theory in Medicine

Control theory is an interdisciplinary field bridging engineering and mathematics, whose main objective is to define an opportune control function that modifies the state of a given dynamical system in order to perform a specific task, while minimizing the cost and maximizing the performance (Bertsekas, 1995; Lenhart and Workman, 2007; Aström and Murray, 2010) (see section 3 for a technical description).

Two main classes of controls exist: (i) *open-loop* control, and (ii) *closed-loop* (feedback) control. In the former case, the set and sequence of control actions is chosen a priori, by exploiting theoretical study on the models. In this case, the input is independent with respect to the output (e.g., possible measurements on the system) (Lenhart and Workman, 2007). Closed-loop control, instead, introduces in the procedure one or more feedback loops, which are able to quantify the real response of the system to variations of the control functions, and adjust them according to the differences recorded between the theoretical and real behaviors of the system (Aström and Murray, 2010).

There are several examples of successful applications of control theory in pharmacology (see Bailey and Haddad, 2005; Shi et al., 2014). In this respect, the final goal is to determine the optimal set of therapeutic choices—e.g., dosages and schedules—to obtained a desired efficacy, while minimizing the overall costs. Closed-loop controls are extremely effective in achieving this goal and have been often implemented in real-world health-care settings (Haddad et al., 2006; Steil, 2013; Jayachandran et al., 2014; Shi et al., 2014; Babaei and Salamci, 2015; Fuentes-Garí et al., 2015; Naşcu et al., 2015). Nonetheless, technological problems, such the absence of real-time measurements, as well as possible problems in titrating drugs to the right concentration, are still limiting real-life applications (Bailey and Haddad, 2005; Cunningham et al., 2018). For this reason, open-loop controls are still a viable option, mostly because of the applicability in a

wide range of real-world scenarios for which, for instance, real-time measurements and/or therapy adjustments are unfeasible. Moreover, open-loop controls have proven to identify more effective drug concentration in therapeutic ranges than standard clinical practice (Barbolosi and Iliadis, 2001; Ledzewicz and Schättler, 2006; Zhu and Qian, 2014; Bara et al., 2017; Rocha et al., 2018; Yoon et al., 2018).

However, many approaches in both categories are based on limiting assumptions. Certain techniques, for instance, assume continuous—yet unrealistic—drug infusion procedures (e.g., Pefani et al., 2013). Some methods rely on often speculative mathematical models, which cannot be evaluated due to the lack of opportune experimental data (Yoon et al., 2018).

The CT4TD framework aims at improving the current state-of-the-art, by solving an open-loop control problem on PK/PD models via RedCRAB (Heck et al., 2018b; Omran et al., 2019), a remote suite based on dCRAB (Doria et al., 2011; Rach et al., 2015), an algorithm for optimization and control. The dCRAB algorithm is particularly suitable for complex optimization problems when it is neither possible or efficient to build the gradient from the set of differential equations, defined by the main dynamics. In the aforementioned case, the standard gradient-based methods could not be efficient or failed the gradient calculation. The dCRAB optimal control tool has the peculiarity to avoid local traps by changing the optimization basis and paves also the possibility to perform a closed-loop optimization, using the feedback provided by the patient's response. Extensions in this sense are underway.

2.3. Mathematical Modeling of Cell Population Dynamics

Many healthy and aberrant biological tissues are characterized by a hierarchical organization, constituted by an ordered sequence of discrete maturation states, as driven by differentiation processes. In this respect, a number of mathematical models have been proposed to study the cell population dynamics, both in healthy systems (Marciniak-Czochra and Stiehl, 2013) and in cancer (Michor et al., 2005; Tang et al., 2011; Stiehl and Marciniak-Czochra, 2012; Olshen et al., 2014; Altrock et al., 2015; Werner et al., 2016; Stiehl et al., 2018).

In such models, cells are divided in n non-intersecting compartments, with every ensemble representing a certain stage of cell differentiation. The time ordering of the differentiation stage defines an explicit hierarchy among such ensembles. Accordingly, a *lineage* is defined as a collection of compartments that fully describe all the stages of differentiation of cells within a certain tissue (see **Figure S7**).

Various approaches are employed to model the dynamics of such systems such as, e.g., ordinary differential equations (ODEs), discrete-time and continuous-time Markov chains, master equations, etc. (see Altrock et al., 2015 for a recent review). Each strategy displays a specific trade-off in terms of expressivity and computational complexity. For instance, ODEs are very convenient from the computational perspective, but they are not suitable in certain cases, e.g., when representing low numbers of cells. Conversely, probabilistic models allow for a richer representation of the system, yet at the cost of a higher computational burden and mathematical complexity.

For sake of simplicity, the CT4TD framework employs a ODEs hierarchical model of cell population dynamics to fit longitudinal data on tumor burden (Michor et al., 2005; Tang et al., 2011; Stiehl and Marciniak-Czochra, 2012; Olshen et al., 2014; Altrock et al., 2015; Werner et al., 2016; Stiehl et al., 2018). On the one hand, this model provides a description of cell population dynamics in time for any given patient. On the other hand, it allows to estimate the efficacy of the therapy in each patient, on the basis of the observed cancer subpopulation decay, which is then used to estimate patient-specific PD models (see section 3 for further details).

3. MATERIALS AND METHODS

3.1. Estimation of Patient-Specific PK Models of Imatinib in CML

In order to describe the various steps of the CT4TD pipeline in detail, we here present its application to the specific case of Imatinib administration in CML. Yet, we stress that the theoretical approach is general and could be applied for any therapy for which PK/PD models can be estimated.

We here employ the PK model of oral administration of Imatinib introduced by Widmer et al. (2006) (see **Figure S1**): if $\chi_g(t)$ is the amount of Imatinib in the gastrointestinal tract, k_a is the first order absorption rate, f is the bioavailability, i.e., the fraction of an administered dose of unchanged drug that reaches the circulatory system, and D the ingested dose, then:

$$\frac{d\chi_g(t)}{dt} = -k_a\chi_g(t), \quad \chi_g(0) = Df, \tag{1}$$

so if v is the volume of the distribution, i.e., the theoretical volume needed to account for the overall amount of drug in the body in case the drug was evenly distributed throughout the body, CL the clearance, i.e., the volume of plasma cleared of the drug per unit time, $C(t)$ the concentration in the blood, $\chi_b(t)$ the amount of Imatinib in the blood ($C(t) = \frac{\chi_b(t)}{v}$), then:

$$\frac{d\chi_b(t)}{dt} = +k_a\chi_g(t) - CL \cdot C(t), \quad C(0) = 0. \tag{2}$$

An example of the solution of Equation (2) can be found in **Figure S2**.

Both equations can be tuned to consider demographic factors as body weight, age and sex, thus providing patient-specific PK models. More in detail, such demographic factors can be incorporated in the clearance CL and in the volume of the distribution v, as initially proposed by Widmer et al. (2006):

$$CL = \theta_a + \theta_1\frac{BW - \overline{BW}}{\overline{BW}} + \theta_2 q - \theta_2(1-q) + \theta_3\frac{AGE - \overline{AGE}}{\overline{AGE}}, \tag{3}$$

$$v = \theta_b + \theta_4 q - \theta_4(1-q), \tag{4}$$

where θ_i, for $i = a, b, 1, 2, 3, 4$ are constants, BW is the body weight of the patient and $\overline{BW}$ is its population-average, AGE is the age of the patient and $\overline{AGE}$ its population-average and q is a binary variable which takes value 1 for male and 0 for female. Estimation of such parameters is provided by Widmer et al. (2006) and shown in **Tables S1**, **S2**. As in the dataset used in the case study, only the information about age and sex was available, we estimated the corresponding BW in each patient on the basis of average measures provided by McDowell et al. (2005)[1].

3.1.1. Therapy Simulation

In order to simulate a therapy, we need to model a multi-dose oral administration. Let t_{in} and t_{fin} be the initial and the final time of the therapy, we suppose to give $n+1$ doses at time $t_0, t_1, t_2, \ldots, t_n$, with a dose amount D_i ($i = 0, 1, 2, 3, \ldots n$), respectively. Thus, we have n first order differential equations like Equation (1); by using the superposition principle the solution will be:

$$\begin{array}{ll} t_0 \le t < t_1 & \chi_g(t) = \chi_{g0}(t), \\ t_1 \le t < t_2 & \chi_g(t) = \chi_{g0}(t) + \chi_{g1}(t), \\ t_2 \le t < t_3 & \chi_g(t) = \chi_{g0}(t) + \chi_{g1}(t) + \chi_{g2}(t), \\ \ldots & \ldots \\ t_{in} \le t < t_{fin} & \chi_g(t) = \sum_{i=0}^{n} \chi_{gi}(t), \end{array} \tag{5}$$

where $\chi_{gi}(t) = \psi_{gi}(t_i)e^{-k_a(t-t_i)}$ is a solution of Equation (1), with $\psi_{gi}(t_i) = fD_i$. Then, substituting $\chi_g(t)$ into the Equation (2) it is possible to study the dynamics of blood concentration of a certain drug for a multi-dose oral administration (see **Figure S2** for an example). Notice that $\frac{\chi_g(t)}{v} = C(t)$.

3.2. Estimation of Patient-Specific PD Models From Experimental Data

CT4TD includes a data analysis module, aimed at identifying patient-specific parameters of the PD model from experimental data, and which relies on a widely-used model of cancer population dynamics. The following subsection include details on each pipeline step.

3.2.1. Population Dynamics Model of Leukemia

In CT4TD we use the simplest compartmental model of population dynamics for which it is possible to estimate the parameters from available experimental data.

In detail, the organization of leukemic systems is characterized by a hierarchy that is analogous to the healthy hematopoietic counterpart, and which can be modeled in the simplest case with two compartments: (i) cancer stem cells (CSC) and (ii) progressively differentiated cancer cells (Michor et al., 2005; Tang et al., 2011; Stiehl and Marciniak-Czochra, 2012; Olshen et al., 2014; Wodarz et al., 2014; Werner et al., 2016; Stiehl et al., 2018).

[1]It is known that other factors can change the value of the clearance and the volume of the distribution. For instance, the concentration of the α_1-acid glycoprotein affects the volume of the distribution, whereas the MDR1 genotype, the CYP3A4 activity and the creatinine clearance affect both CL and v (Widmer et al., 2006). However, we here limit to consider the aforementioned demographic measurements, because of their availability in our and in most datasets.

Note that the model could be generalized to account for m lineages, in order to represent the possible presence of subpopulations of tumor cells with distinct properties (e.g., therapy resistant phenotypes), and to account for complex interaction phenomena (e.g., between lymph nodes/bone-marrow and blood-stream). However, in order to allow for an accurate and robust parameter estimation, more complex models of population dynamics would typically require—among other things—a much higher number of data points than those usually available. In addition, limitations regarding parameter identification of ODE models with inadequate data (i.e., identifiability of a model) were described in Saccomani et al. (2010) and Hong et al. (2019), and justify our choice of adopting highly simplified models, at least until new suitable experimental data will become available.

In this case, first order differential equations are suitable to describe the population dynamics, because experimental evidences show that the proliferation of healthy cells display an exponential increase (Marciniak-Czochra and Stiehl, 2013), whereas cancer cells under therapy display an exponential decay (Michor et al., 2005; Tang et al., 2011; Olshen et al., 2014; Rainero et al., 2018). Thus, we analyse the fluxes between compartments, by defining the following constants:

- $p_{i,k}$ is the division rate of the cells in the i ensemble of the k lineage.
- $d_{i,k}$ is the death rate of the cells in the i ensemble of the k lineage—this rate will be estimated from experimental data.
- $a_{i,k} \in [0, 1]$ is the probability that, when a cell undergoes mitosis, both of its daughters belong to the i ensemble in the k lineage; therefore, $1 - a_{i,k}$ is the probability of belonging to the $i+1$ ensemble. With respect to CSCs (or SC) this quantifies the self-renewal process.

Note that, we consider a symmetric differentiation scheme, according to which after mitosis both cells are of the same type, either stem or differentiated.

In a single individual, the dynamics of the healthy system—which includes stem cells, progenitors and differentiated cells—and that of the leukemic system coexist. Yet, we here assume that CML cells are cytokine-independent (as formulated by Werner et al., 2016), so the equations describing the dynamics of leukemic subpopulation do not include terms related to the healthy counterpart, (i.e., the ODE system of healthy and leukemic subpopulations becomes uncoupled). As a consequence, the dynamics of the leukemic system can be defined as follows:

$$\begin{array}{l} \frac{dl_1(t)}{dt} = \lambda l_1(t), \\ \frac{dl_2(t)}{dt} = \gamma l_1(t) + \tau l_2(t), \end{array} \tag{6}$$

where:

$$\begin{array}{l} \lambda = (2a_{1,l} - 1)p_{1,l} - d_{1,l},\ \gamma = 2(1 - a_{1,l})p_{1,l}, \\ \tau = -d_{2,l}. \end{array} \tag{7}$$

This is a typical example of a linear autonomous system, and the solution could be obtained analytically in a recursive way:

$$\begin{aligned} l_1(t) &= l_1(0)e^{\lambda t}, \\ l_2(t) &= \frac{e^{\tau t}(\gamma l_1(0) - \lambda l_2(0) + \tau l_2(0)) - l_1(0)\gamma e^{-\lambda t}}{\tau - \lambda}. \end{aligned} \tag{8}$$

3.2.2. Experimental Data Fitting

Once the leukemia 2-compartments model has been defined, it is possible to estimate its parameters from experimental data. In particular, we here focus on the specific case of CML. As every cancer cell in CML is characterized by the BCR-ABL mutation, it is possible to distinguish healthy from cancer cells with Q-PCR measurement, and this allows to have longitudinal experimental data returning the fraction of cancer cells over the total, i.e., the tumor burden (Michor et al., 2005; Tang et al., 2011; Olshen et al., 2014; Rainero et al., 2018).

The CT4TD fits the longitudinal data on tumor burden in each patient with a biphasic exponential, which in log-scale describes two distinct and intersecting lines, as proposed in Michor et al. (2005), Tang et al. (2011), and Olshen et al. (2014). In particular, we selected the combination of straight lines minimizing the value of R^2 (i.e., the standard coefficient of the goodness of a linear regression, which quantifies the portion of the response that is explained by a linear model), by scanning all the points of intersection (with a step of 1 day) and fitting data with two lines with distinct slopes, via a standard non-linear fit (see the **Table S3** for all parameter estimation). We also tried to fit data with either one or three distinct lines, yet in our case study the best fit was obtained in the two-lines case.

Once the two best fitting curves have been obtained for each patient, we adopt a simplifying assumption that allows us to estimate the parameters of the compartmental model from data. In Marciniak-Czochra et al. (2009), it is shown that the leftmost curve (with higher slope) is likely to represent the overall population dynamics involving cancer stem cells, cancer progenitors and cancer differentiated cells (decreasing in population size), together with that of healthy blood cells (increasing in population size). Considering that the Q-PCR measurements of the BCR-ABL fusion gene return the ratio between cancer cells and the total number of cells in the system, it would be impossible to disentangle the contribution of each subpopulation to the overall dynamics, and to reliably estimate the values of γ and τ in Equation (7), without *ad-hoc* assumptions and/or further opportune experiments.

Instead, it is possible to hypothesize that the rightmost curve (i.e., the second exponential decay, with lower slope) accounts for the dynamics involving a completely recovered healthy cell subpopulation—thus, healthy cells can be considered as constant in number—and a decaying cancer stem cell subpopulation, with no progenitors and differentiated cancer cells left in the system, as a consequence of the therapy (Michor et al., 2005; Tang et al., 2011; Olshen et al., 2014; Rainero et al., 2018). With this assumption, it is possible to estimate the parameters of the first compartment, and in particular, the cancer stem cell death rate $d_{1,l}$ in Equation (9), from experimental data. This also allows us not to explicit consider a model for the healthy hematopoietic system.

In detail, we assume that the exponential decay of the rightmost curve (i.e. the exponential decay of CSC, given by the first equation in Equation 8) accounts for the dynamics of the CSC subpopulation only. In this way, it is possible to evaluate the effect of a standard Imatinib therapy—400 mg per day—directly on the CSC decay, as estimated from any patient's data.

In fact, β_j, i.e., the measured slope accounting for the decay of CSCs, will be given by:

$$\beta_j = Log_{10}[e][(2a_{1,l,j} - 1)p_{1,l,j} - d_{1,l,j}] = Log_{10}[e]\lambda_j, \tag{9}$$

where j is the patient's index.

3.2.3. Identification of Patient-Specific PD Models

Various PD models can be employed to estimate the efficacy of Imatinib in CML. In our case, we use a PD model based on the maximum-inhibition effect (E_{max}) (Peng et al., 2005) (see **Figure S3**):

$$E(C(t)) = \frac{E_{max} \cdot C^n(t)}{EC_{50}^n + C^n(t)}, \tag{10}$$

where $E(t)$ is the effect, E_{max} is the maximum effect, $C(t)$ is the concentration of the drug, EC_{50} the concentration of the drug that produces half of maximal effect, and n is a shape factor.

In order to identify patient-specific PD models from the parameters of the leukemia model estimated from data, we can safely suppose a linear relation between the population-average of the efficacy $\langle E \rangle$ and the population-average of $\langle d_{1,l} \rangle$ (Gambacorti-Passerini et al., 1997). Hence, the relation is the following:

$$\langle d_{1,l} \rangle = K \langle E \rangle, \tag{11}$$

where K is a conversion constant. In this case, the efficacy of a certain concentration of a drug is directly proportional to the increase of the cancer cell death rate.

It is possible to estimate K from the available dataset, by employing patient-specific PK models and an average benchmark PD model ($n = 1, EC_{50} = 0.123 \quad [mg/L]$ and $E_{max} = 1$). To do this, we first compute the time-average of the concentration $\bar{C}_j(t)$ for each patient, with respect to a 40-days standard therapy—400 mg Imatib per day—, which we then use to compute the time-average of the efficacy $\bar{E}_j$ as per Equation (10), by considering unique average PD parameters for all patients. Finally, we consider the population-average the efficacy $\langle E \rangle$, as computed on all patients. $\langle d_{1,l,j} \rangle$ is then obtained by using formula (Equation 9), setting $a_{1,l} = 0.87$ and $p_{1,l} = 0.45 \quad [days^{-1}]$ as proposed (Stiehl et al., 2018), and by taking the mean over all patients. As a result, the conversion factor for this dataset is $K \approx 0.377 \pm 0.0007 \quad [days^{-1}]$. Since we suppose a linear relation between K, $\langle d_{1,l} \rangle$ and $\langle E \rangle$ (see Equation 11), the confidence interval of K is determined via standard (linear) error propagation procedure.

At this point, we are able to estimate the personalized parameters of the PD model, and in particular of EC_{50}, by supposing that the maximum efficacy is $E_{max} = 1$ and the shape factor is $n = 1$, for all patients (Peng et al., 2005; Weigel et al., 2010). Therefore, the relation is the following:

$$EC_{50,j} = \bar{C}_j \left[\frac{KE_{max}}{d_{1,l,j}} - 1 \right]^{1/n}. \tag{12}$$

With this procedure, we can estimate the value of $EC_{50,j}$ for each patient from individual longitudinal data on tumor burden. This result leads to the definition of PK/PD personalized models, which integrate demographic factors and Q-PCR data that measure the response of each patient to the therapy, and represents one of the major novelties of our approach.

The results for all patients are presented in **Table S4** and in **Figure 4**. Notice that, if longitudinal data on single patients are not available, the CT4TD allows to employ a unique (average) PD model for all patients, as estimated from experimental studies (see e.g., Gambacorti-Passerini et al., 1997; Peng et al., 2005; Picard et al., 2007; Baccarani et al., 2014).

3.3. Definition of the PK/PD Control Problem

We formally define the *PK/PD control problem* for the administration of discrete doses as follows. Let be t_{in} and t_{fin} the initial and the final time of the therapy. Here we aim at finding: (i) the optimal doses $D_0^*, D_1^*, \dots D_n^*$, and (ii) the optimal schedule of administration $t_0^*, t_1^*, \dots, t_n^*$, such that a functional that represents the *cost* $\mathcal{L}(C(t, \{(D_0, t_0), (D_1, t_1), \dots, (D_n, t_n)\}))$ is minimized. Notice that the set $\{(D_0^*, t_0^*), (D_1^*, t_1^*), \dots, (D_n^*, t_n^*)\}$ are the control functions and C^* is the optimal unknown drug concentration, described in a general setting with the ODE in Equation (18) (the solution of the particular case of multi-dose oral administration is shown in Equation 5). To simplify the notation, in the following we will refer to $\mathcal{L}(C(t, \{(D_0, t_0), (D_1, t_1), \dots, (D_n, t_n)\}))$ as $\mathcal{L}(C(t, \{(D_i, t_i)\}))$.

The definition of the *cost* functional $\mathcal{L}$ is the *core* of the PK/PD control problem and can include various weighed terms, which one should wisely select with respect the specific problem and goals. In particular, $\mathcal{L}$ may (or may not) include distinct terms accounting for: (i) the efficacy E of the therapy, as derived via PD models, such as the Hill equation (Goutelle et al., 2008) or the E_{max} model (Peng et al., 2005) (if average or patient-specific parameters can be estimated); (ii) the toxicity of the therapy and/or the possible AEs as measured, e.g., via the Area Under the Curve (AUC); (iii) in case the PD model is unknown or indefinable—a distance between the optimized concentration $C^*(t)$ and a target concentration C_{tg} as estimated, for instance, from clinical trials (Baccarani et al., 2014); (iv) the economic cost of the therapy; (v) the properties and the temporal evolution of the disease, as in the case of the tumor burden estimation from longitudinal experimental data (Michor et al., 2005; Stiehl and Marciniak-Czochra, 2012; Altrock et al., 2015; Werner et al., 2016; Stiehl et al., 2018) (in this case the goal will be the optimization of the performance of the therapy with respect to the minimization of cancer subpopulations; (vi) the probability of developing resistance to the therapy (Michor et al., 2005; Tang et al., 2011), etc. Obviously, some of these terms are highly correlated, as for example the AUC and the economic cost. Notice also that the choice of opportune weights is crucial in defining an effective control, if more than one term is used, and that it is necessary to fix *n a priori* when minimizing $\mathcal{L}$. The latter choice is related to the applicability to current real-world scenarios, in which practical limitations usually prevent to exceed a certain amount of doses per day, as well as to administer continuous dosages, especially with respect to cancer therapies. In detail, we here define cost functions $\mathcal{L}$ with respect to two distinct scenarios: (i) optimization of therapy for fixed target concentrations, (ii) optimization of therapy for tumor burden reduction or stabilization (as proposed by West et al., 2018).

3.3.1. Working Scenario (i): Optimal Control With Fixed Target Concentration at Diagnosis Time (Patient-Specific PK Models—No PD Models)

In many real-world scenarios it is not possible to retrieve or estimate the parameters of the PD models, for instance at the time of diagnosis. In this case, CT4TD can be used to find the best personalized therapeutic strategy to either: (i) be as close as possible to a given *optimal* target drug concentration, or (ii) be close to, but strictly larger than a given *lower-bound* target (i.e., *steady state optimization*; Shargel et al., 1999). In the first case—i.e., *optimal* target concentration—CT4TD employs a simple Euclidean distance between two concentrations:

$$\mathcal{E}_{(C_1,C_2)} = \left|C_2(t) - C_1(t)\right|. \tag{13}$$

C_{tg} is the target drug concentration, necessary to have a major molecular response, estimated, e.g., via clinical trials. For example, in clinical studies, Imatinib concentration in blood is required to be above 0.57 $[mg/L]$ and with a time average of 1 $[mg/L]$ (Peng et al., 2005). $C^*(t)$ is the unknown concentration of drug in blood, which will be identified by solving the control problem.

Then, in this case the *cost* is defined as follows:

$$\mathcal{L}(C^*(t, \{(D_i^*, t_i)\})) = \int_{t_{in}}^{t_{fin}} dt \mathcal{E}_{(C^*, C_{tg})}. \tag{14}$$

This *cost* favors the solutions which are close to the target, with no preference between above or under the target. In the second case—i.e., *lower-bound* target concentration—CT4TD uses a *step* distance in the space of concentrations. In this case the distance between two concentrations becomes:

$$\mathcal{S}_{(C_1,C_2)} = \begin{cases} \mathcal{E}_{(C_1,C_2)}, & C_1 \geq C_2, \\ G, & C_1 < C_2, \end{cases} \tag{15}$$

where G is a constant. It is then possible to define the *cost* as follows:

$$\mathcal{L}(C^*(t, \{(D_i^*, t_i)\})) = \int_{t_{in}}^{t_{fin}} dt \mathcal{S}_{(C^*, C_{tg})}. \tag{16}$$

In this case, CT4TD will give solutions that display concentrations above the lower-bound target. We stress that working scenario (i) is general, as target drug concentration can be derived from any given clinical trial or practice. In such case, the target concentration is a parameter of the cost functional, which can be opportunely modified according to the considered therapy, both in the optimal target (Equation 14) and the lower-bound target cases (Equation 16).

3.3.2. Working Scenario (ii): Optimal Control for Tumor Burden Reduction (Patient-Specific PK Models—Patient-Specific PD Models)

When it is possible to estimate patient-specific PD parameters from longitudinal data on tumor burden variation under standard treatment, CT4TD can be used to identify an adjusted optimized therapy to reduce such burden in each patient. In particular, our approach can use a *cost* function with the aim of: (i) minimizing the number of cancer stem cells (e.g., at the end of the treatment, (ii) minimizing the AUC of the therapy, which accounts for toxicity and possible AEs. To do this, we need to introduce two arbitrary weights W_1 and W_2, which account for the relative relevance of the two distinct components. Notice that we cannot consider the exponent of Equation (8) only, as the dynamics of $l_1(t)$ is a monotone function and the minimization of cancer stem cells is reached for $D_i \to \infty$, which implicates that if the maximum amount of drug is bounded (i.e., D_{max}) the optimal solution is trivially reached for $D_i = D_{max}, \forall i$. Therefore, we have:

$$\begin{aligned}\mathcal{L}(C^*(t,\{(D_i^*,t_i)\})) &= \int_{t_{in}}^{t_{fin}} dt[W_1(Log_{10}[e]\lambda(E_j(C^*(t)))) \\ &\quad + W_2 C^*(t)] \\ &= \int_{t_{in}}^{t_{fin}} dt[W_1(Log_{10}[e]((2a_{1,l}-1)p_{1,l} \\ &\quad - KE_j(C^*(t)))) + W_2 C^*(t)],\end{aligned} \tag{17}$$

Notice that cost functional in Equation (17) includes the net growth rate of the CSCs, i.e., λ in Equation (8). This choice allows us not to know or estimate the initial number of CSCs $l_1(0)$, since is not possible to infer this quantity from tumor burden data only. The ratio $\phi = \frac{W_1}{W_2}$ determines the overall of the optimal solution and should be wisely chosen. In order to provide some indications on this modeling choice, we performed an extensive scan of ϕ (in the range $\phi \in [10, 100]$), with respect to all patients and we analyzed the variation of the time-average concentration $\bar{C}$. The results are shown in **Figure 5**. From this analysis, one can see that a sound choice for ϕ might be in the range [60–65] for males and in [70–75] for females (note this choice depends on the units of measurement), meaning that the weight corresponding to the time evolution of CSCs is relatively more relevant than that corresponding to the toxic effects.

We finally specify that working scenario (ii) was originally designed for liquid tumor therapies, as it requires the definition and the measurement of the tumor burden, which is used to estimate the CSCs net growth rate. Whether measurements on tumor burden and an appropriate model for cancer population dynamics would be available for distinct cancer types, our framework might be applied without any significant theoretical modification.

3.4. Resolution of Control Problem via RedCRAB

In CT4TD the control problem is heuristically solved by using RedCRAB, the remote version of the dressed Chopped RAndom Basis (dCRAB) optimal control via a cloud server (Caneva and et al., 2011; Doria et al., 2011; Rach et al., 2015; Heck et al., 2018a; Omran et al., 2019). Optimal control theory has been used for decades to optimize classical processes, and its quantum counterpart has been increasingly exploited in the last years (Khaneja and et al., 2005; Spörl et al., 2007; Caneva and et al., 2009; Brif et al., 2010; Lloyd and Montangero, 2014; Koch, 2016; Pichler et al., 2016; Sørensen and et al., 2016; van Frank and et al., 2016; Deffner and Campbell, 2017; Goerz et al., 2017). In its simplest version, optimal control drives the state of the system to a goal one, characterized by some desired properties, by using a set of time-dependent controls.

Here, the dynamics of the system is identified by the concentration of the drug in a certain compartment $C(t, D(t))$, which obeys the time evolution equation:

$$\frac{\partial C(t,D(t)))}{\partial t} = f(t, D(t), C(t, D(t))), \tag{18}$$

where $D(t)$ is the time-dependent control function, i.e., the doses function defined in section (see the section describing the *patient-specific PK models of Imatinib in CML*). The goal here is to optimize the drug administration schedule (see section 3.4.1) while minimize the cost functional as defined in subsections describing *working scenario* (i) and *working scenario* (ii) (see above).

Starting from the standard administration schedule $D_0(t)$, the optimization proceeds by looking for the optimal correction $g(t)$ such the optimal administration schedule will be $D(t) = D_0(t) + g(t)$. Following (Rach et al., 2015), the correction $g(t)$ is expanded in a truncated function space, specifically in random Fourier components as:

$$g(t) = \Gamma(t) \sum_{k=1}^{n_c} [A_k \sin(\omega_k t) + B_k \cos(\omega_k t)] \tag{19}$$

where $\omega_k = 2\pi(k + r_k)/T$ and $r_k \in [-0.5, 0.5]$, n_c is the total number of frequency used, T is the final time, and $\Gamma(t)$ is a fixed scaling function to keep the values at initial and final times unchanged. In conclusion, the control problem is reformulated as maximization of a multivariable function $\mathcal{L}(A_k, B_k)$ with fixed ω_k, and can be efficiently solved numerically by searching the best combination of $\{A_k, B_k\}$ with the preferred method of choice, here a direct-search method (Nelder and Mead, 1965). Notice that, each frequency ω_k is independently optimized: indeed, after a certain number of iterations, we move to the next ω_{k+1}, by introducing an external loop on the frequencies, i.e., super-iterations. This allows the algorithm to include a high number of Fourier components and efficiently find the optimal solution, by avoiding local traps that can stick the optimization into not the global minimum (Rach et al., 2015).

In the RedCRAB optimization, the server generates and transmits a set of controls to the CT4TD, which evaluates the cost function, by interfacing with MATLAB and communicates it to the server completing one iteration. The optimization continues iteratively by providing the optimal set of controls as well as giving back the figure of merit, until the convergence is reached.

We specify that, as for any heuristic method, the solution provided by our approach might be sub-optimal. This depends on the complexity of the search space and by the computational resources available. Yet, as proven in several real-world applications (Doria et al., 2011; Rach et al., 2015; Hoeb and et al., 2017; Omran et al., 2019) RedCRAB was proven to be a computationally efficient and robust technique.

3.4.1. Optimal Dosage

To solve the optimization problem with respect to dosages, we optimize a control field $D(t)$ defined between ($t_0 \leq t \leq t_f$) via RedCRAB. Then, we proceed by mapping the $D(t)$ doses function into $(n+1)$-integer values which correspond to $(n+1)$-doses D_j^* (j = 0,..., n), where n is the number of total doses given to the patient; t_f is the final time of the therapy and $t_0 = 0$ the initial time. Accordingly, we can define the schedule of administration as:

$$(t_0, t_1, ..., t_i, ..., t_n, t_{n+1}) = \left(0, \frac{t_f}{n+1}, ..., i \cdot \frac{t_f}{n+1}, ..., n \cdot \frac{t_f}{n+1}, t_f\right) \quad (20)$$

i.e., $t_i = i \cdot \frac{t_f}{n+1}$ for $i = 0, 1, ..., n+1$ Indeed, the n-doses D_j^* are obtained by integrating the doses function $D(t)$ between adjacent times in the schedule administration as follows:

$$D_j^* = \int_{t_j}^{t_{j+1}} ds\, D(s) \quad (21)$$

with $j = 0, ..., n$. The more general case where also the time schedule of the administration is optimized (i.e., *optimal schedule* case) is described in the SM.

4. RESULTS

4.1. Imatinib Administration in Chronic Myeloid Leukemia—CML

We here show the application of CT4TD to the specific case of Imatinib mesylate administration in patients with CML. The final goal is to determine the drug optimized dosage and schedule in two distinct scenarios.

1. In the first case the goal is to optimize personalized therapeutic strategies to reach given target concentrations, as those commonly used in clinical protocols, and by assuming to be at diagnosis time.
2. In the second case, we employ the population dynamics models, as retrieved by fitting longitudinal data on single patients under standard treatment, to deliver patient-specific therapies that are most effective in reducing/eradicating the tumor subpopulation after the major molecular response, on the basis of PK/PD personalized models.

Imatinib is an inhibitor of the BCR-ABL tyrosine kinase, which is known to bind to the inactive form of BCR-ABL at nanomolar concentration, competing with the ATP for its binding pocket and hindering the switch of the fusion kinase to the active form, therefore impairing the catalytic activity of the enzyme (Gambacorti-Passerini et al., 2003). The therapy is in most cases long-life (Michor et al., 2005; Tang et al., 2011; Branford et al., 2013; Olshen et al., 2014; Rainero et al., 2018) and the treatment is expensive ($\approx$ 30,000 US$ per year; Cole and Dusetzina, 2018). Therefore, the impact of an optimized and personalized administration would be two-fold: on the one hand, it could be effective in optimizing the performance, while reducing the toxicity and minimizing the adverse effects for long-term therapies (Larson et al., 2008b; Mughal and Schrieber, 2010; Hu et al., 2012); on the other hand, it could help in reducing the overall economical costs, which currently limit the access to therapy, hence making long-term health care more sustainable (Fojo and Grady, 2009; Himmelstein et al., 2009)[2].

4.2. Datasets

We applied the CT4TD framework to a longitudinal dataset from Michor et al. (2005), in which 29 CML patients have been monitored with a peripheral blood draw taken every 90 days, from the time of diagnosis up to a maximum time of about 3, 500 days (average time $\approx 2,659 \pm 938$ days) For these patients, the administration schedule has been 400 mg Imatinib/day for the whole considered period.

In particular the fraction of cancer cells in blood—that will be referred to as *tumor burden* from now on—can be reliably estimated by analyzing the expression level of the fusion gene BCR-ABL, thus providing an easy way to monitor the disease progression, as well the response to therapy. As BCR-ABL transcript is solely expressed by the leukemic cells, its measurement by mean of quantitative PCR (Q-PCR) is considered one of the most sensitive and specific techniques to indirectly assess the tumor burden, and is the standard de facto for monitoring minimal residual disease in CML.

More in detail, we selected a subset of the dataset provided in Michor et al. (2005), by removing all the patients that displayed too few data points (i.e., < 3), or that were characterized by resistant mutations, i.e., specific DNA alterations that render the therapy via Imatinib ineffective, usually due to steric impediments (Shah et al., 2004). In such cases, it is common practice to employ an alternative therapy, based either on Dasatinib, Nilotinib, Ponatinib, or Bosutinib (Shah et al., 2004). We decided to leave resistant patients out of the analysis for two distinct technical reasons. First, this scenario would require a more complex population dynamics model—i.e., with more subpopulations—, characterized by many more parameters, often impossible to estimate. Second, in this case the identification of an optimized therapy should involve two distinct controls, and even if theoretically possible, this would require to obtain data concerning the effect of Dasatinib/Nilotinib/Ponatinib/Bosutinib on tumor burden, which are not present in the used dataset.

We eventually selected 22 (out of 29) patients, for which the therapy led to a successful major molecular response (MMR), i.e., the ratio of cells with BCR-ABL mutation is ≤ 0.1 on the international scale (Griffiths et al., 2014).

[2]Note that it was recently hypothesized that CML CSC could be resistant to the effects of Imatinib and persist in all patients on long-term therapy. (Holyoake and Vetrie, 2017). However, only further experimental studies could unravel this point.

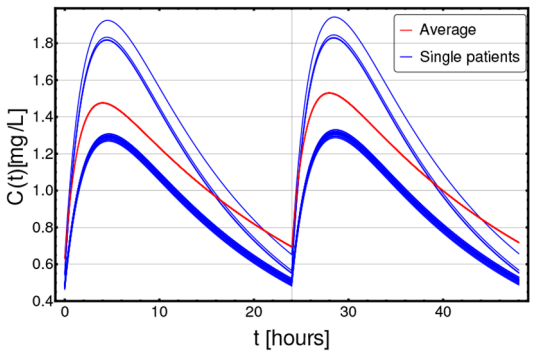


FIGURE 2 | Patient-specific PK models. Personalized pharmacokinetic curves (blue solid lines), as estimated from demographic factors, such as age, sex, and body weight, as per Equations (3) and (4), in the range $t \in [0, 48]$ h (x axis); y axis describes the drug concentration $C(t)$ in $[mg/L]$. The blue curves correspond to the 22 distinct patients included in the dataset, whereas the dashed red curve represents the population-average pharmacokinetics.

4.3. Patient-Specific PK Models

We first use patient-specific PK models by incorporating demographic factors—i.e., body weight, age, and sex—in the clearance and in the volume of the distribution, as per Equations (3) and (4) (Widmer et al., 2006)(see section 3). The parameter settings of RedCRAB optmization for this case study are shown in **Table S6**.

In **Figure 2**, the PK curves corresponding to the 22 patients (in blue) and the average PK model (in red) over a selected time window ($[0, 48]$ h) are displayed.

4.4. Defining Personalized Optimized Administration at Diagnosis Time

CT4TD can be employed when CML is diagnosed, in order to identify optimized therapeutic strategies that lead to drug concentrations as close as possible to given targets. We here present the application of CT4TD to two distinct targets.

The first target concentration is $C_{targ}(t) = 0.57 \quad [mg/L]$, which is currently the most widely employed in the clinic (Peng et al., 2005). It is hypothesized that any effective therapy should ensure a drug concentration close to, but strictly larger than this value, in order to lead to a good performance, while minimizing the AEs (Graham et al., 2002; Faber et al., 2016). In this case, we consider this concentration as a *lower-bound* target, and the goal of the CT4TD framework will be to design an optimized therapy to be close to, but strictly larger than this concentration value, by employing an opportune distance notion (see section 3).

The second target concentration is $C_{targ}(t) = 1 \quad [mg/L]$ and is supposed to provide a more effective therapy, but at the cost of an increased likelihood of AEs and toxicity (Gambacorti-Passerini et al., 1997; Picard et al., 2007; Baccarani et al., 2014). From common practice, an effective therapy is that leading to values of drug concentrations *around* this target. For this reason, CT4TD will return an optimized therapy ensuring a drug concentration *as close as possible* to this *optimal* target (see section 3). Note that one could select any arbitrary target concentration, or even combinations of targets, and this would not affect the validity of our approach. For both targets we tested distinct settings, in which we considered, respectively, 1 and 3 doses per day at fixed times (i.e., 1 dose each 24 and 8 h, respectively)[3].

In **Figure 3**, we present the application of CT4TD to a selected patient (n. 0001 00004 AJR, male), with respect to the *lower-bound* target $C_{targ}(t) = 0.57 \quad [mg/L]$ (left panels), and the *optimal* target $C_{targ}(t) = 1 \quad [mg/L]$ (right panels). In particular, we compared the standard administration (red), the 1-dose optimized therapy (blue) and the 3-doses optimized therapy (green), on a temporal window of 14 days, with respect to drug dosage (**Figures 3A–E**),drug concentration in blood (**Figures 3B–F**), cumulative (Euclidean) distance with respect to the target concentration (**Figures 3C–E**), and AUC (**Figures 3D–H**).

When assessing the goodness of a therapy in the lower-bound scenario—i.e., $C_{targ} = 0.57 \; [mg/L]$—it is important to look at both the distance to the target *and* the overall time in which the drug concentration is above such target. In **Figures 3A–D**, one can see that the optimized 1-dose strategy displays higher cumulative distance and area under the curve—AUC—with respect to the standard schedule, due to the fact that drug concentration is always strictly larger than the lower-bound target. This is proven by the proportion of time spent above the target (computed on the whole period), which is 100, 100, and 88.6%, for the 1-dose, the 3-doses, and the standard administrations, respectively. The 3-doses optimized strategy displays a remarkable improvement also with respect to cumulative distance and AUC, proving to be an effective therapeutic choice for this specific patient. This expected result shows the effectiveness of our methodological approach in producing biologically-plausible experimental hypotheses.

With respect to the optimal target—i.e., $C_{targ}(t) = 1 \; [mg/L]$—, an effective therapy should ensure a drug concentration as close as possible to the target, thus reducing the drug surpluses, while minimizing the cases of insufficient dosage. In this case, the 1-dose optimized scenario almost overlaps with the standard administration (yet, this is not always the case as one can see, for example, with respect to 0006 00007 RJW in **Figure S18**), whereas the 3-doses optimized strategy displays an improvement in terms of cumulative distance, as the drug concentration is constantly kept much closer to the desired target, thus importantly reducing under- and over-dosing (**Figures 3E–H**). In **Figures S8–S28**, one can find the results of the analyses on the other 21 patients included the dataset.

4.5. Adjusting Treatment for Tumor Burden Reduction

The CT4TD framework can be employed in order to identify optimized therapeutic strategies for patients that are currently treated with a standard regime, and for which longitudinal data on tumor burden variation are available. In this case, in order to estimate personalized PD models from experimental data—which describe the individual therapeutic response to identical drug concentrations—, CT4TD employs a module which fits

[3]It would be possible to use our theoretical framework to define a free-time schedule optimization procedure. Yet, we believe that current practices in Imatinib oral administration would make a free-time schedule scarcely usable.

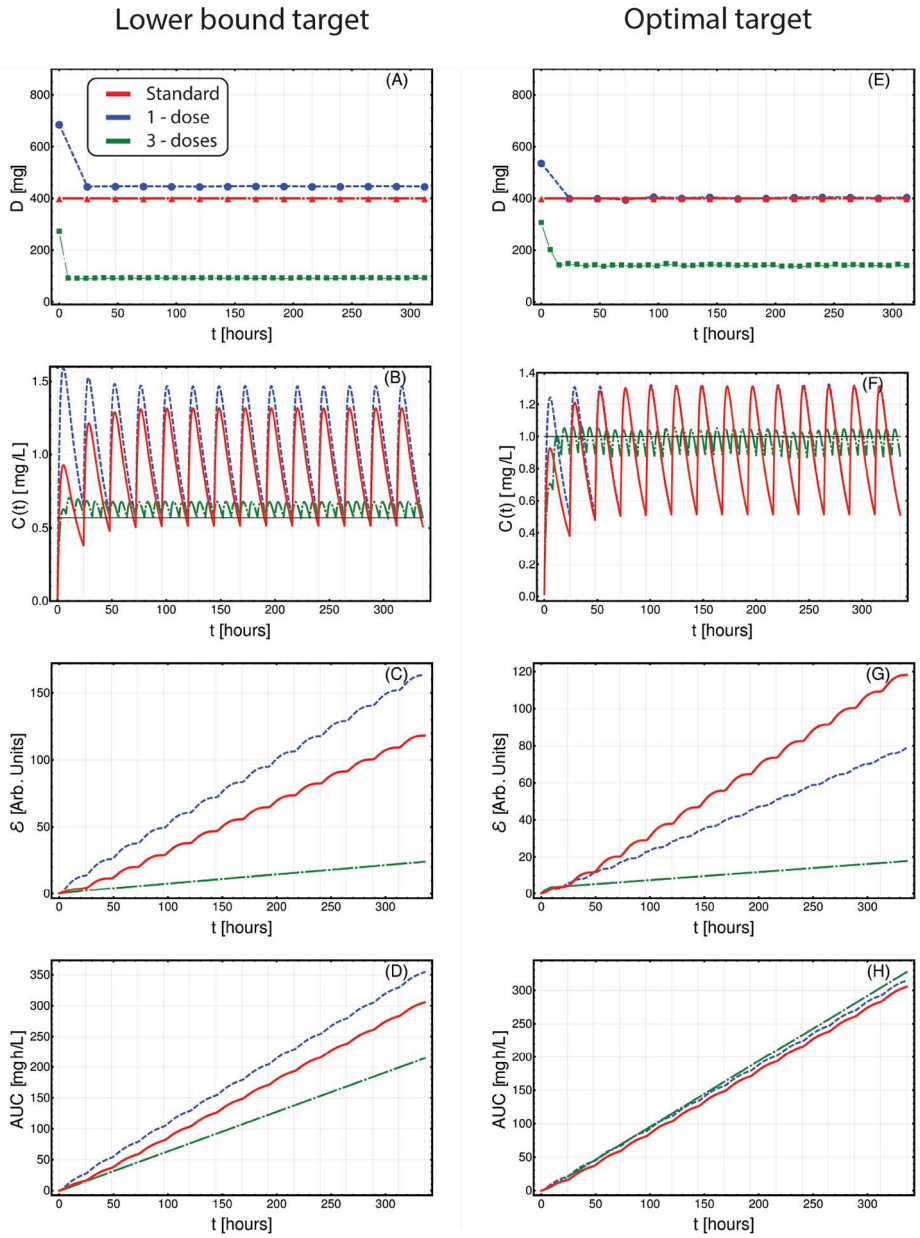


FIGURE 3 | Patient-specific optimized therapy with fixed target drug concentrations. Optimized Imatinib administration returned by CT4TD for patient 0001 00004 AJR from Michor et al. (2005), in the cases of: 1-dose/day (blue) and 3-doses/day (green), with respect to: *lower-bound* target concentration $C_{targ} = 0.57$ $[mg/L]$ **(A–D)**, and *optimal* target concentration $C_{targ} = 1$ $[mg/L]$ **(E–H)**. Standard administration—i.e., 400 mg Imatinib/day—is shown with a red dashed line. In this case, the optimization is obtained on patient-specific PK parameters, without considering the PD models. **(A,E)** Imatinib scheduled dosage in *mg* (y axis), displayed on 14 days (x axis). **(B–F)** Imatinib concentration in blood in $[mg/L]$ (y axis). **(C–G)** Variation of the cumulative distances between the observed concentration and the selected target in time. **(D–H)** Temporal variation of the AUC in $[mg \cdot h/L]$.

longitudinal data on tumor burden with a hierarchical model of cancer population dynamics[4].

[4]Notice that any arbitrary ODE mathematical model could be employed, as long it is effective in representing the phenomenological properties of the disease (e.g., multi-stable states), and that sufficient and adequate data are available to estimate its parameters.

In particular, we fitted each patient's data with a biphasic exponential, which in log-scale describes the presence of two straight lines with distinct slopes, as proposed by Michor et al. (2005), Tang et al. (2011), and Olshen et al. (2014) (see section 3 for further details). With a few assumptions, the slope of such lines can be used to estimate the parameters of a 2-compartment population dynamics model of CML and, in particular, the

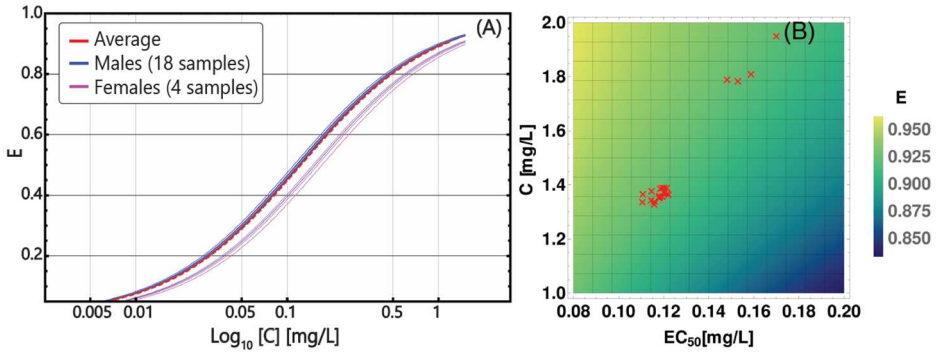


FIGURE 4 | Patient-specific PD models. **(A)** personalized PD curves obtained from Equation (10) by using $E_{max} = 1$ and $n = 1$ for all patients, and distinct values of EC_{50}, based on the death rate of cancer stem cells, as estimated from longitudinal data on tumor burden; x axis (Log_{10} scale) describes the concentration in the range $C \in [0, 1.2]$ $[mg/L]$, on y axis the efficiency E is displayed. The solid blue curves correspond to the 18 distinct male patients included in the dataset and the solid pink curves correspond to the 4 distinct female patients and the dashed red curve represents the population-average pharmacodynamics. **(B)** Heat-map returning the variation of efficiency E, computed via Equation (10), with respect to distinct parameters of the PK model—i.e., patient-specific time-average concentration $\bar{C}$ (y axis)—and of the PD model—i.e., patient-specific EC_{50} (x axis). Red triangles represent the 22 patients in the dataset.

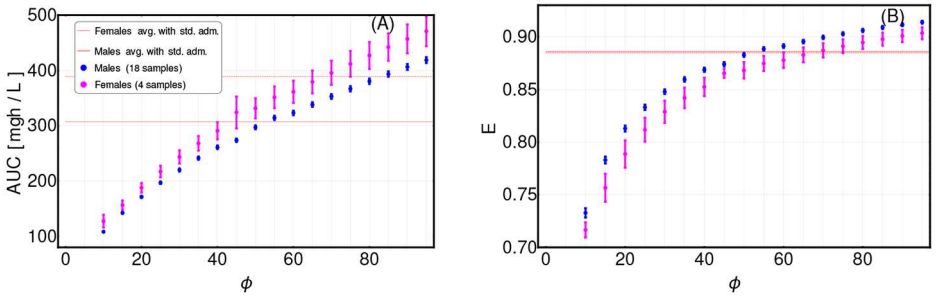


FIGURE 5 | Assessment of term weights in cost function definition. The definition of the cost function for the adjusting treatment scenario requires to set the weights of the different terms. We here considered two terms, in order to: (i) minimize the tumor burden (weight W_1), and (ii) minimize the AUC (weight W_2) (see section 3 for further details). We scanned the values of $\phi = \frac{W_1}{W_2}$ in the range [10, 100], by repeatedly applying the CT4TD framework to the 22-patients CML dataset from Michor et al. (2005). **(A)** Distribution of the value of the AUC after 14-days of the optimized therapy retrieved by CT4TD (1-dose case), for distinct values of ϕ, with respect to the 22 samples in the datasets, divided in males (blue) and females (pink), and compared to the average AUC values returned by standard administration (400 mg Imatinib/day) in males (red solid line) and females (red dashed line). **(B)** Distribution of efficiency computed via Equation (10) on the time-average concentration over 14 days of the optimized therapy retrieved by CT4TD (1-dose case), for distinct values of ϕ, and compared to the average efficiency in the standard administration scenario (solid and dashed red lines overlap).

(stem) cancer subpopulation death rate in presence of a standard Imatinib therapy—i.e., 400 *mg* per day—in each patient. This allows to estimate the patient-specific parameters of the PD model. The results of the data analysis on all patients are presented in **Table S3** and in **Figures S4–S6**. In **Figure 4A**, one can see the personalized PD curves for the 22 patients, computed via Equation (10), as compared to the average one.

We also assessed the relative relevance of the personalized parameters of the PD and PK models with respect to the efficacy of the therapy. The heat-map in **Figure 4B** returns the variation of the efficacy with respect to combination of time-average concentration $\bar{C}$ and EC_{50}, highlighting the personalized parameters of the 22 patients. As a first result, one can see that much of the variance in our dataset is due to differences in PK, rather than to PD, which however is still relevant. Notice also that the two visible clusters basically overlap with the male and the female groups, providing a possible explanation of the distinct therapeutic response observed in clinical studies (Branford et al., 2013). We stress that the estimation of personalized PD models from experimental data of patients under treatment is one of the major novelties of our approach, and, in combination with the demographics-based PK models, allow to identify patient-specific therapeutic regimes that are optimized to minimize the tumor burden.

In order to identify personalized optimized therapies, we finally defined a cost function with the goals of: (i) maximizing the reduction of the tumor burden, and (ii) minimizing the toxicity and possible AEs, in terms of AUC (see section 3 for further details). Such cost function requires to set opportune weights W_1 and W_2 for the two terms, respectively. In particular, a parameter $\phi = \frac{W_1}{W_2}$ is defined, which can be opportunely tuned to favor either the first or the second term. However, the choice

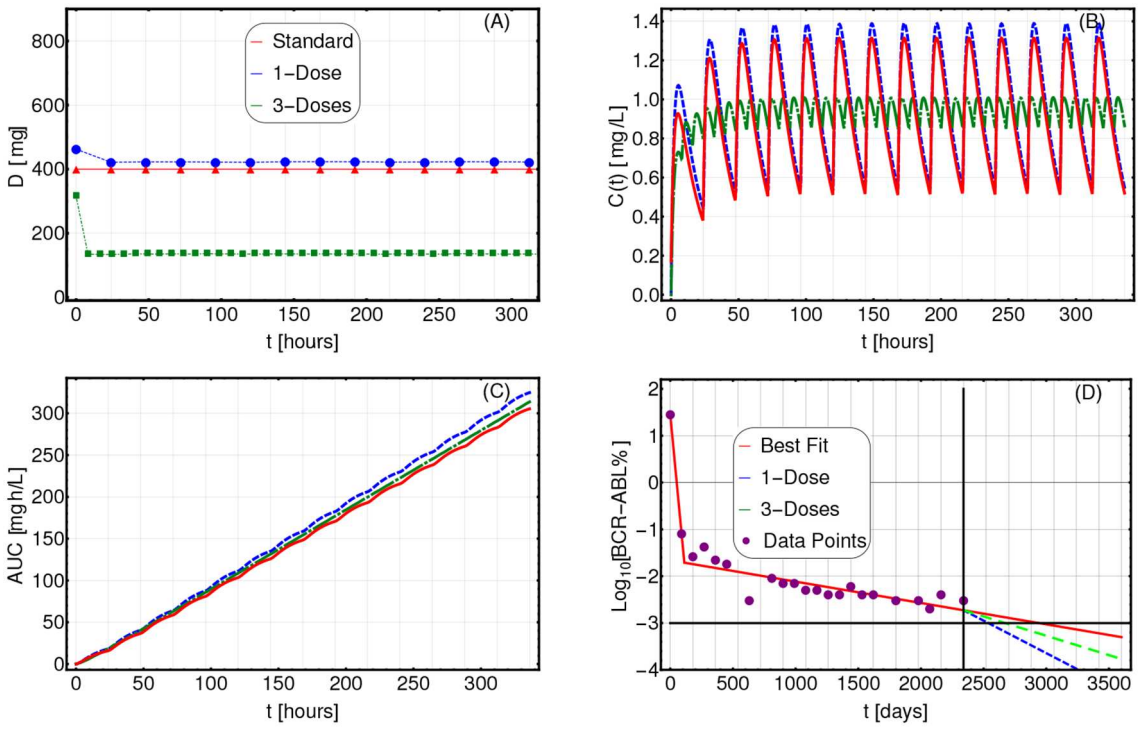


FIGURE 6 | Adjusting therapy for tumor burden minimization. Imatinib administration optimized for tumor burden minimization in patient 0001 00004 AJR (male) from Michor et al. (2005), in the cases of: 1-dose/day (purple) and 3-doses/day (green), with $\phi = 60$, as compared to standard administration (red). In this case, the optimization is obtained on patient-specific PK and PD models. **(A)** Imatinib scheduled dosage in *mg* (y axis), displayed on 14 days (x axis). **(B)** Imatinib concentration in blood in $[mg/L]$ (y axis). **(C)** Temporal variation of the AUC in $[mg \cdot h/L]$ (y axis). **(D)** Longitudinal data points on tumor burden recorded in the interval $t \in [0, 2340]$ days (purple). The best fit is shown with red lines. The slope of the right-most line is used to determine the cancer stem cell death rate and, in turn, the patient-specific PD parameters. The blue (green) line represents the predicted cancer subpopulation decay in case the 1-dose (3-doses) optimized therapy was adopted from day 2, 340 to day 3, 500.

of a specific value for ϕ is arbitrary and depends on subjective research and clinical criteria.

To investigate the sensitivity of our framework to the variation of this parameter, we repeatedly applied the CT4TD framework to the CML dataset, by scanning various values of ϕ, and eventually assessed the differences in: (i) the time-average AUC as computed on a 14-days temporal window, and (ii) the efficiency computed on the time-average concentration in the same period, with respect to a 1-dose optimization scenario (the 3-doses scenario can be found in **Figure S50**). In **Figure 5**, one can see the distribution of both quantities with respect to the 22 patients in the dataset, divided in males (blue) and females (pink), as compared to the average AUC and efficiency for the standard administration case (red).

A first important thing to notice is that the results are highly sensitive with respect to the choice of ϕ, and can either display improvements (e.g., higher AUC and/or lower efficiency) or worsening with respect to the standard case in distinct cases. Moreover, male and female groups show significantly different distributions, thus pointing at physiological differences that should be considered in therapy design. As a rule-of-thumb, we suggest to select a value of ϕ for which a slightly larger value of efficiency is observed, while not inducing a too high increase in AUC. In our case, we selected a value of ϕ equal to 60 for men and of 75 for women.

In **Figures 6A–D**, we show the comparison among the actual therapeutic regime administered to a selected patient (n.0001 00004 AJR, male—code ID in **Table S3**) and the optimized therapies identified via CT4TD by setting $\phi = 60$, in both 1-dose (blue) and 3-doses (green) scenarios, in terms of: (i) drug dosage, (ii) drug concentration, (iii) AUC, and (iv) variation of the tumor burden in time. In particular, the temporal evolution of the tumor burden from diagnosis to the present is displayed by showing the experimental data points (purple) and the best fit (red), whereas the predicted future evolution is shown with respect to the 1-dose (blue) and the 3-doses (green) optimized strategies.

A result is that, given similar AUC curves (i.e., similar toxicity and AEs), both the 1-dose and the 3-doses optimized strategies lead to a significantly faster predicted tumor burden decay. In particular, the tumor burden decay is, respectively 3.07 and 1.82 times faster for the 1-dose and the 3-doses regimes, with respect to standard administration. This result paves the way for an automated strategy for therapy adjustment design, which might be further developed by employing closed-loop controllers.

In **Figures S29–S49**, one can find the results of the analyses on the other 21 patients included the dataset.

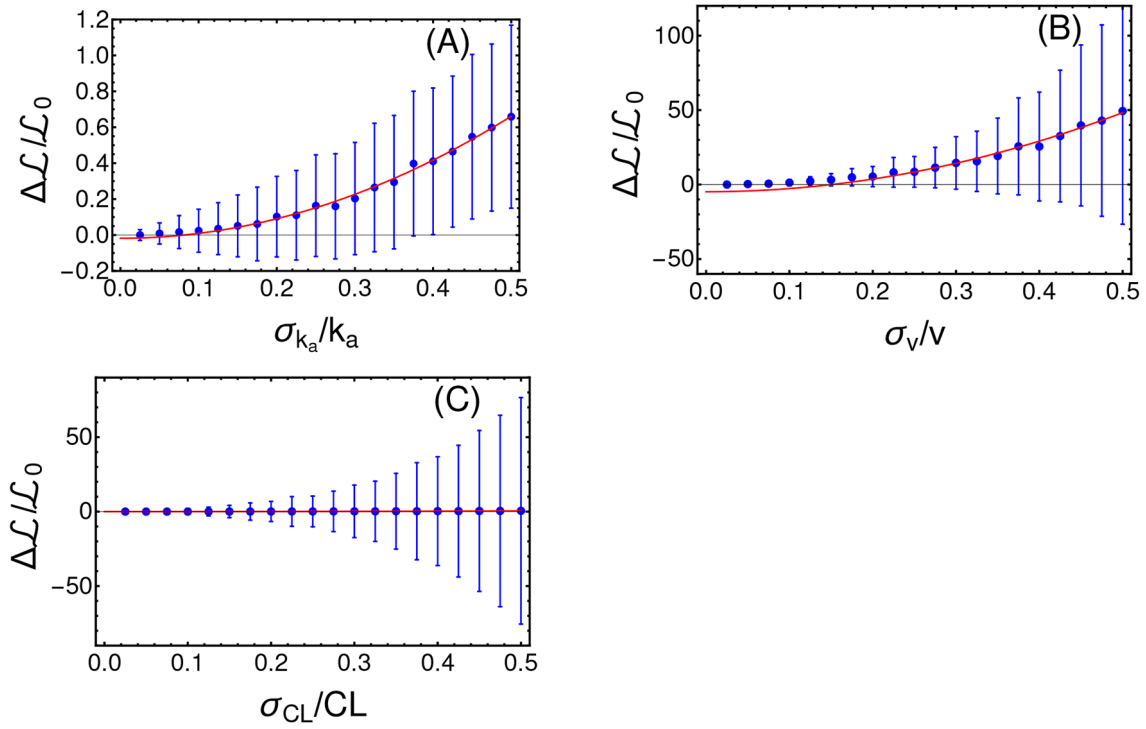


FIGURE 7 | Robustness analysis with respect to intra-patient variability. Intra-patient variability is here defined as stochastic noise randomly picked in the uniform distribution $\chi_r = [-\sigma_r, \sigma_r]$ and applied to distinct PK parameters. In particular, we here show the average and the standard deviation of the relative variation of the *cost* $\Delta\mathcal{L}/\mathcal{L}_0$, with respect to distinct levels of noise, in the cases: $r = k_a$ **(A)**, $r = v$ in **(B)**, and $r = CL$ **(C)**. Values of the fit are provided in **Table S5**.

4.5.1. Robustness Analysis

In order to assess the reliability of the results produced by CT4TD, we tested its robustness with respect to intra-patient variability and to possible systematic errors. To account for intra-patient variability, we introduced a stochastic and uniformly distributed noise, i.e., $\chi_r = [-\sigma_r, \sigma_r]$ with $r = k_a, CL, v$, to the following parameters of the PK model: k_a, CL, and v, for every time point in the analysis. We performed 700 distinct PK simulations, on the average patient, in the specific scenario of a target concentration $C_{targ}(t) = 0.57 \quad [mg/L]$ and 1 dose per day. We then analyze the relative variation of the average *cost* $\Delta\mathcal{L}$, as compared to the noise-free case, with respect to the width of the distribution of noise σ_r. In **Figure 7**, one can notice that $\Delta\mathcal{L}$ variation with respect to the noise level follows an approximately quadratic trend, which is proven by fitting the data points with a curve with equation $b + a\sigma_r^2$ (the complete results of the fit are provided in **Table S5**). Note that such results are in agreement with other works that use quantum optimal control (Montangero et al., 2007; Kallush et al., 2014; Hoeb and et al., 2017). We performed a further robustness analysis, to assess the impact of systematic errors, as those possibly due to scarce reliability of the demographic study and/or to small or imbalanced datasets, and which may result in errors in the estimation of the PK parameters. To this end, we generated an optimized schedule for a set of PK parameters $\{k_a, CL, v\}$ and then we applied such schedule to a simulated patient where a parameter at time is different, e.g., $\{k'_a, CL, v\}$ with $k'_a = k_a + \delta_{k_a}$. We finally measured the difference of $\Delta\mathcal{L}$, as a function of δ_{k_a}. Also in this case, we show in **Figure 8** that the results produced by CT4TD are robust with respect to possible technical or measurement errors. In fact, with respect to an error of $\approx \pm 30\%$, we observe a maximum difference $\approx 10\%$ in performance, as compared to the noise-free case.

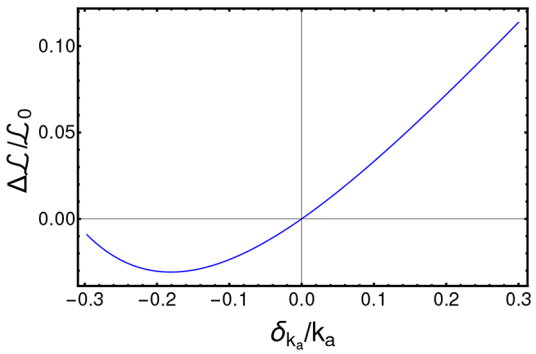


FIGURE 8 | Robustness analysis with respect to systematic errors. Relative variation of the *cost* $\Delta\mathcal{L}/\mathcal{L}_0$ with respect to the relative value of the systematic error δ_{k_a}/k_a.

5. DISCUSSION

The introduction of the CT4TD framework aims at providing an automated and data-driven procedure for decision support in health care and personalized therapy design in cancer, especially by exploiting the increasing available computational power, which allows one to perform large-scale simulations and efficient search in the parameter space, and to deal with noisy and imperfect data.

In particular, CT4TD aims at overcoming the limitations of current control-based methods for therapeutic hypothesis generation. First, its completely general theoretical approach allows to consider: (i) any disease for which a PK/PD model can be derived and its parameters measured, (ii) any kind of administration, e.g., continuous drug infusion or discrete doses, (iii) any measurable term that is considered as relevant in the definition of a therapeutic *cost*. CT4TD eventually allows to evaluate *in silico* the outcome of the designed therapy.

Furthermore, CT4TD introduces the possibility of designing optimized therapeutic strategies based on experimental data concerning the disease progression. The identification of data-based patient-specific PK/PD models is one of the major novelties of CT4TD and has a profound impact on the characterization of tumor heterogeneity and, accordingly, on the customization of cancer therapies.

One of the main limitations of CT4TD derives from the adoption of highly simplified models of cancer population dynamics. Unfortunately, the shortage of adequate longitudinal data on tumor dynamics prevents to estimate the parameters of more sophisticate and biologically realistic models, which may take into account, for instance, the existence of various competing cancer subpopulations, or the complex interplay occurring within the tumor microenvironment. However, we claim that our theoretical approach is completely general and it will hold whether and when higher-resolution longitudinal data on disease progression would become available, allowing for instance to measure the (sub)clonal prevalence variation in time (as proposed, e.g., by Acar et al., 2019).

Several developments of CT4TD are underway. In particular, the possibility of tuning the PK/PD models to include information on the somatic evolutionary history of the tumors (Ramazzotti et al., 2015; Caravagna et al., 2016, 2018) will be essential in delivering more effective personalized therapeutic strategies. This is especially important for tumors displaying high levels of intra-tumor heterogeneity, which is known to be responsible for drug resistance, therapy failure and relapse (McGranahan and Swanton, 2015).

As CT4TD relies on the RedCRAB optimization framework (Heck et al., 2018b; Omran et al., 2019), the overall procedure could be implemented in remote, paving the way for a wireless decision support system for therapy design, to be used directly by clinicians (Jeong et al., 2015). In this respect, as a future development, an open-source computational tool will be made available to the scientific community, allowing to perform individual-specific analysis for a wide range of disease.

DATA AVAILABILITY STATEMENT

All data used in this paper are available from the supplementary material of Michor et al. (2005). We provide the source code and the input data to reproduce the case studies at: https://github.com/BIMIB-DISCo/CT4TD.

AUTHOR CONTRIBUTIONS

FA, AG, SM, and MA designed the research. FA and MR implemented the method. FA, AG, and SM performed the research. FA, AG, DM, TC, and RP analyzed the data. All authors wrote and revised the paper.

ACKNOWLEDGMENTS

This work was partially supported by the Elixir Italian Chapter and the SysBioNet project, a Ministero dell'Istruzione, dell'Universitá e della Ricerca initiative for the Italian Roadmap of European Strategy Forum on Research Infrastructures. Furthermore we acknowledge financial support from IQST alliance. Partial support was also given by the CRUK/AIRC Accelerator Award #22790, Single-cell Cancer Evolution in the Clinic. We thank Giulio Caravagna for helpful discussions. This manuscript has been released as a pre-print at bioRxiv (Angaroni et al., 2019).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00523/full#supplementary-material

REFERENCES

Acar, A., Nichol, D., Fernandez, J., Cresswell, G. D., Barozzi, I., Hong, S. P., et al. (2019). Exploiting evolutionary herding to control drug resistance in cancer. *BioRxiv* 566950. doi: 10.1101/566950

Altrock, P. M., Liu, L. L., and Michor, F. (2015). The mathematics of cancer: integrating quantitative models. *Nat. Rev. Cancer* 15:730. doi: 10.1038/nrc4029

Angaroni, F., Graudenzi, A., Rossignolo, M., Maspero, D., Calarco, T., Piazza, R., et al. (2019). Personalized therapy design for liquid tumors via optimal control theory. *bioRxiv* 662858. doi: 10.1101/662858

Aström, K. J., and Murray, R. M. (2010). *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton: Princeton University Press.

Babaei, N., and Salamci, M. U. (2015). Personalized drug administration for cancer treatment using model reference adaptive control. *J. Theor. Biol.* 371, 24–44. doi: 10.1016/j.jtbi.2015.01.038

Baccarani, M., Druker, B. J., Branford, S., Kim, D., Pane, F., Mongay, L., et al. (2014). Long-term response to imatinib is not affected by the initial dose in patients with Philadelphia chromosome-positive chronic myeloid leukemia in chronic phase: final update from the tyrosine Kinase inhibitor optimization and selectivity (tops) study. *Int. J. Hematol.* 99, 616–624. doi: 10.1007/s12185-014-1566-2

Bailey, J. M., and Haddad, W. M. (2005). Drug dosing control in clinical pharmacology. *IEEE Control Syst.* 25, 35–51. doi: 10.1109/MCS.2005.1411383

Bara, O., Djouadi, S., Day, J., and Lenhart, S. (2017). Immune therapeutic strategies using optimal controls with L1 and L2 type objectives. *Math. Biosci.* 290, 9–21. doi: 10.1016/j.mbs.2017.05.010

Barbolosi, D., and Iliadis, A. (2001). Optimizing drug regimens in cancer chemotherapy: a simulation study using a PK/PD model. *Comput. Biol. Med.* 31, 157–172. doi: 10.1016/S0010-4825(00)00032-9

Bertsekas, D. P. (1995). *Dynamic Programming and Optimal Control, Vol. 1*. Belmont, MA: Athena Scientific.

Branford, S., Yeung, D. T., Ross, D. M., Prime, J. A., Field, C. R., Altamura, H. K., et al. (2013). Early molecular response and female sex strongly predict stable undetectable BCR-ABL1, the criteria for imatinib discontinuation in patients with CML. *Blood* 121, 3818–3824. doi: 10.1182/blood-2012-10-462291

Brif, C., Chakrabarti, R., and Rabitz, H. (2010). Control of quantum phenomena: past, present and future. *New J. Phys.* 12:075008. doi: 10.1088/1367-2630/12/7/075008

Caneva, T., Calarco, T., Fazio, R., Santoro, G. E., and Montangero, S. (2011). Speeding up critical system dynamics through optimized evolution. *Phys. Rev. A* 84:012312. doi: 10.1103/PhysRevA.84.012312

Caneva, T., Murphy, M., Calarco, T., Fazio, R., Montangero, S., Giovannetti, V., et al. (2009). Optimal control at the quantum speed limit. *Phys. Rev. Lett.* 103:240501. doi: 10.1103/PhysRevLett.103.240501

Caravagna, G., Giarratano, Y., Ramazzotti, D., Tomlinson, I., Graham, T. A., Sanguinetti, G., et al. (2018). Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nat. Methods* 15:707. doi: 10.1038/s41592-018-0108-x

Caravagna, G., Graudenzi, A., Ramazzotti, D., Sanz-Pamplona, R., De Sano, L., Mauri, G., et al. (2016). Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proc. Natl. Acad. Sci. U.S.A.* 113, E4025-E4034. doi: 10.1073/pnas.1520213113

Chan, P. L., Jacqmin, P., Lavielle, M., McFadyen, L., and Weatherley, B. (2011). The use of the saem algorithm in monolix software for estimation of population pharmacokinetic-pharmacodynamic-viral dynamics parameters of maraviroc in asymptomatic HIV subjects. *J. Pharmacokinet. Pharmacodyn.* 38, 41–61. doi: 10.1007/s10928-010-9175-z

Cole, A. L., and Dusetzina, S. B. (2018). Generic price competition for specialty drugs: too little, too late? *Health Affairs* 37, 738–742. doi: 10.1377/hlthaff.2017.1684

Cunningham, J. J., Brown, J. S., Gatenby, R. A., and Staňková, K. (2018). Optimal control to develop therapeutic strategies for metastatic castrate resistant prostate cancer. *J. Theor. Biol.* 459, 67–78. doi: 10.1016/j.jtbi.2018.09.022

Deffner, S., and Campbell, S. (2017). Quantum speed limits: from Heisenberg's uncertainty principle to optimal quantum control. *J. Phys. A Math. Theor.* 50:453001. doi: 10.1088/1751-8121/aa86c6

Doria, P., Calarco, T., and Montangero, S. (2011). Optimal control technique for many-body quantum dynamics. *Phys. Rev. Lett.* 106:190501. doi: 10.1103/PhysRevLett.106.190501

Experts in Chronic Myeloid Leukemia (2013). The price of drugs for chronic myeloid leukemia (CML) is a reflection of the unsustainable prices of cancer drugs: from the perspective of a large group of CML experts. *Blood* 121, 4439–4442. doi: 10.1182/blood-2013-03-490003

Faber, E., Divoká, M., Skoumalová, I., Novák, M., Marešová, I., Mičová, K., et al. (2016). A lower dosage of imatinib is sufficient to maintain undetectable disease in patients with chronic myeloid leukemia with long-term low-grade toxicity of the treatment. *Leukemia Lymphoma* 57, 370–375. doi: 10.3109/10428194.2015.1056184

Fojo, T., and Grady, C. (2009). How much is life worth: cetuximab, non-small cell lung cancer, and the 440 billion question. *J. Natl. Cancer Instit.* 101, 1044–1048. doi: 10.1093/jnci/djp177

Fuentes-Garí, M., Velliou, E., Misener, R., Pefani, E., Rende, M., Panoskaltsis, N., et al. (2015). A systematic framework for the design, simulation and optimization of personalized healthcare: making and healing blood. *Comput. Chem. Eng.* 81, 80–93. doi: 10.1016/j.compchemeng.2015.03.008

Gambacorti-Passerini, C., Le Coutre, P., Mologni, L., Fanelli, M., Bertazzoli, C., Marchesi, E., et al. (1997). Inhibition of the ABL kinase activity blocks the proliferation of BCR/ABL+ leukemic cells and induces apoptosis. *Blood Cells Mol. Dis.* 23, 380–394. doi: 10.1006/bcmd.1997.0155

Gambacorti-Passerini, C. B., Gunby, R. H., Piazza, R., Galietta, A., Rostagno, R., and Scapozza, L. (2003). Molecular mechanisms of resistance to imatinib in Philadelphia-chromosome-positive leukaemias. *Lancet Oncol.* 4, 75–85. doi: 10.1016/S1470-2045(03)00979-3

Goerz, M. H., Motzoi, F., Whaley, K. B., and Koch, C. P. (2017). Charting the circuit QED design landscape using optimal control theory. *NPJ Quantum Inf.* 3:37. doi: 10.1038/s41534-017-0036-0

Gomez-de León, A., Gómez-Almaguer, D., Ruiz-Delgado, G. J., and Ruiz-Arguelles, G. J. (2017). Insights into the management of chronic myeloid leukemia in resource-poor settings: a Mexican perspective. *Expert Rev. Hematol.* 10, 809–819. doi: 10.1080/17474086.2017.1360180

Goutelle, S., Maurin, M., Rougier, F., Barbaut, X., Bourguignon, L., Ducher, M., et al. (2008). The hill equation: a review of its capabilities in pharmacological modelling. *Fund. Clin. Pharmacol.* 22, 633–648. doi: 10.1111/j.1472-8206.2008.00633.x

Graham, S. M., Jørgensen, H. G., Allan, E., Pearson, C., Alcorn, M. J., Richmond, L., et al. (2002). Primitive, quiescent, Philadelphia-positive stem cells from patients with chronic myeloid leukemia are insensitive to STI571 *in vitro*. *Blood* 99, 319–325. doi: 10.1182/blood.V99.1.319

Graudenzi, A., Caravagna, G., De Matteis, G., and Antoniotti, M. (2014). Investigating the relation between stochastic differentiation, homeostasis and clonal expansion in intestinal crypts via multiscale modeling. *PLoS ONE* 9:e97272. doi: 10.1371/journal.pone.0097272

Graudenzi, A., Maspero, D., and Damiani, C. (2018). "Modeling spatio-temporal dynamics of metabolic networks with cellular automata and constraint-based methods," in *International Conference on Cellular Automata* (Cham: Springer), 16–29. doi: 10.1007/978-3-319-99813-8_2

Griffiths, M., Patton, S. J., Grossi, A., Clark, J., Paz, M. F., and Labourier, E. (2014). Conversion, correction, and international scale standardization: results from a multicenter external quality assessment study for bcr-abl1 testing. *Arch. Pathol. Lab. Med.* 139, 522–529. doi: 10.5858/arpa.2013-0754-OA

Haddad, W. M., Hayakawa, T., and Bailey, J. M. (2006). Adaptive control for nonlinear compartmental dynamical systems with applications to clinical pharmacology. *Syst. Control Lett.* 55, 62–70. doi: 10.1016/j.sysconle.2005.05.002

Heck, R., Vuculescu, O., Jakob Srensen, J., Zoller, J., Andreasen, M. G., Bason, M. G., et al. (2018a). Remote optimization of an ultracold atoms experiment by experts and citizen scientists. *Proc. Natl. Acad. Sci. U.S.A.* 115, E11231-E11237.

Heck, R., Vuculescu, O., Sørensen, J. J., Zoller, J., Andreasen, M. G., Bason, M. G., et al. (2018b). Remote optimization of an ultracold atoms experiment by experts and citizen scientists. *Proc. Natl. Acad. Sci. U.S.A.* 115, E11231-E11237. doi: 10.1073/pnas.1716869115

Himmelstein, D. U., Thorne, D., Warren, E., and Woolhandler, S. (2009). Medical bankruptcy in the United States, 2007: results of a national study. *Am. J. Med.* 122, 741–746. doi: 10.1016/j.amjmed.2009.04.012

Hoeb, F., Angaroni, F., Zoller, J., Calarco, T., Strini, G., Montangero, S., et al. (2017). Amplification of the parametric dynamical casimir effect via optimal control. *Phys. Rev. A* 96:033851. doi: 10.1103/PhysRevA.96.033851

Holyoake, T. L., and Vetrie, D. (2017). The chronic myeloid leukemia stem cell: stemming the tide of persistence. *Blood* 129, 1595–1606. doi: 10.1182/blood-2016-09-696013

Hong, H., Ovchinnikov, A., Pogudin, G., and Yap, C. (2019). Sian: software for structural identifiability analysis of ode models. *Bioinformatics* 35, 2873–2874. doi: 10.1093/bioinformatics/bty1069

Hu, W., Lu, S., McAlpine, I., Jamieson, J. D., Lee, D. U., D. Marroquin, L., et al. (2012). Mechanistic investigation of imatinib-induced cardiac toxicity and the involvement of C-ABL kinase. *Toxicol. Sci.* 129, 188–199. doi: 10.1093/toxsci/kfs192

Jabbour, E. J., Lin, J., Siegartel, L. R., Lingohr-Smith, M., Menges, B., and Makenbaeva, D. (2017). Evaluation of healthcare resource utilization and incremental economic burden of patients with chronic myeloid leukemia after disease progression to blast phase. *J. Med. Econ.* 20, 1007–1012. doi: 10.1080/13696998.2017.1345750

Jayachandran, D., Rundell, A. E., Hannemann, R. E., Vik, T. A., and Ramkrishna, D. (2014). Optimal chemotherapy for leukemia: a model-based strategy for individualized treatment. *PLoS ONE* 9:e109623. doi: 10.1371/journal.pone.0109623

Jeong, J.-W., McCall, J. G., Shin, G., Zhang, Y., Al-Hasani, R., Kim, M., et al. (2015). Wireless optofluidic systems for programmable in vivo pharmacology and optogenetics. *Cell* 162, 662–674. doi: 10.1016/j.cell.2015.06.058

Kallush, S., Khasin, M., and Kosloff, R. (2014). Quantum control with noisy fields: computational complexity versus sensitivity to noise. *New J. Phys*. 16:015008. doi: 10.1088/1367-2630/16/1/015008

Khaneja, N., Reiss, T., Kehlet, C., Schulte-Herbrggen, T., Glaser, S. J. (2005). Optimal control of coupled spin dynamics: Design of NMR pulse sequences by gradient ascent algorithms. *J. Magn. Reson*. 172, 296–305. doi: 10.1016/j.jmr.2004.11.004

Koch, C. P. (2016). Controlling open quantum systems: tools, achievements, and limitations. *J. Phys. Condens. Matter* 28:213001. doi: 10.1088/0953-8984/28/21/213001

Landersdorfer, C. B., and Jusko, W. J. (2008). Pharmacokinetic/pharmacodynamic modelling in diabetes mellitus. *Clin. Pharmacokinet*. 47, 417–448. doi: 10.2165/00003088-200847070-00001

Larson, R. A., Druker, B. J., Guilhot, F., O'Brien, S. G., Riviere, G. J., Krahnke, T., et al. (2008a). Imatinib pharmacokinetics and its correlation with response and safety in chronic-phase chronic myeloid leukemia: a subanalysis of the iris study. *Blood* 111, 4022–4028.

Larson, R. A., Druker, B. J., Guilhot, F., O'Brien, S. G., Riviere, G. J., Krahnke, T., et al. (2008b). Imatinib pharmacokinetics and its correlation with response and safety in chronic-phase chronic myeloid leukemia: a subanalysis of the iris study. *Blood* 111, 4022–4028. doi: 10.1182/blood-2007-10-116475

Ledzewicz, U., and Schättler, H. M. (2006). Drug resistance in cancer chemotherapy as an optimal control problem. *Discrete Cont. Dyn. Syst. Ser. B* 6:129. doi: 10.3934/dcdsb.2006.6.129

Lenhart, S., and Workman, J. T. (2007). *Optimal Control Applied to Biological Models*. London: CRC Press. doi: 10.1201/9781420011418

Lloyd, S., and Montangero, S. (2014). Information theoretical analysis of quantum optimal control. *Phys. Rev. Lett*. 113:010502. doi: 10.1103/PhysRevLett.113.010502

Marciniak-Czochra, A., and Stiehl, T. (2013). "Mathematical models of hematopoietic reconstitution after stem cell transplantation," in *Model Based Parameter Estimation*, eds H. Bock, T. Carraro, W. Jäger, S. Körkel, R. Rannacher, and J. Schlöder (Berlin; Heidelberg: Springer), 191–206. doi: 10.1007/978-3-642-30367-8_9

Marciniak-Czochra, A., Stiehl, T., Ho, A. D., Jäger, W., and Wagner, W. (2009). Modeling of asymmetric cell division in hematopoietic stem cells-regulation of self-renewal is essential for efficient repopulation. *Stem Cells Dev*. 18, 377–386. doi: 10.1089/scd.2008.0143

McDowell, A. M., Fryar, C., Hirsch, R., and Ogden, C. (2005). Anthropometric reference data for children and adults: U.S. population, 1999-2002. *Adv. Data* 361, 1–5. Available online at: https://stacks.cdc.gov/view/cdc/5630

McGranahan, N., and Swanton, C. (2015). Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell* 27, 15–26. doi: 10.1016/j.ccell.2014.12.001

Michor, F., Hughes, T. P., Iwasa, Y., Branford, S., Shah, N. P., Sawyers, C. L., et al. (2005). Dynamics of chronic myeloid leukaemia. *Nature* 435:1267. doi: 10.1038/nature03669

Montangero, S., Calarco, T., and Fazio, R. (2007). Robust optimal quantum gates for josephson charge qubits. *Phys. Rev. Lett*. 99:170501. doi: 10.1103/PhysRevLett.99.170501

Mughal, T. I., and Schrieber, A. (2010). Principal long-term adverse effects of imatinib in patients with chronic myeloid leukemia in chronic phase. *Biologics* 4:315. doi: 10.2147/BTT.S5775

Naşcu, I., Krieger, A., Ionescu, C. M., and Pistikopoulos, E. N. (2015). Advanced model-based control studies for the induction and maintenance of intravenous anaesthesia. *IEEE Trans. Biomed. Eng*. 62, 832–841. doi: 10.1109/TBME.2014.2365726

Nelder, J. A., and Mead, R. (1965). A simplex method for function minimization. *Comput. J*. 7, 308–313. doi: 10.1093/comjnl/7.4.308

Olshen, A., Tang, M., Cortes, J., Gonen, M., Hughes, T., Branford, S., et al. (2014). Dynamics of chronic myeloid leukemia response to dasatinib, nilotinib, and high-dose imatinib. *Haematologica*. 99, 1701–1709. doi: 10.3324/haematol.2013.085977

Omran, A., Levine, H., Keesling, A., Semeghini, G., Wang, T. T., Ebadi, S., et al. (2019). Generation and manipulation of Schrödinger cat states in Rydberg atom arrays. *Science* 365, 570–574. doi: 10.1126/science.aax9743

Pefani, E., Panoskaltsis, N., Mantalaris, A., Georgiadis, M. C., and Pistikopoulos, E. N. (2013). Design of optimal patient-specific chemotherapy protocols for the treatment of acute myeloid leukemia (AML). *Comput. Chem. Eng*. 57, 187–195. doi: 10.1016/j.compchemeng.2013.02.003

Peng, B., Lloyd, P., and Schran, H. (2005). Clinical pharmacokinetics of imatinib. *Clin. Pharmacokinet*. 44, 879–894. doi: 10.2165/00003088-200544090-00001

Picard, S., Titier, K., Etienne, G., Teilhet, E., Ducint, D., Bernard, M. A., et al. (2007). Trough imatinib plasma levels are associated with both cytogenetic and molecular responses to standard-dose imatinib in chronic myeloid leukemia. *Blood* 109, 3496–3499. doi: 10.1182/blood-2006-07-036012

Pichler, T., Caneva, T., Montangero, S., Lukin, M. D., and Calarco, T. (2016). Noise-resistant optimal spin squeezing via quantum control. *Phys. Rev. A* 93:013851. doi: 10.1103/PhysRevA.93.013851

Potts, A. L., Warman, G. R., and Anderson, B. J. (2008). Dexmedetomidine disposition in children: a population analysis. *Pediatr. Anesth*. 18, 722–730. doi: 10.1111/j.1460-9592.2008.02653.x

Rach, N., Müller, M. M., Calarco, T., and Montangero, S. (2015). Dressing the chopped-random-basis optimization: a bandwidth-limited access to the trap-free landscape. *Phys. Rev. A* 92:062343. doi: 10.1103/PhysRevA.92.062343

Rainero, A., Angaroni, F., Conti, A., Pirrone, C., Micheloni, G., Tarará, L., et al. (2018). gDNA qPCR is statistically more reliable than mRNA analysis in detecting leukemic cells to monitor cml. *Cell Death Dis*. 9:349. doi: 10.1038/s41419-018-0387-2

Ramazzotti, D., Caravagna, G., Olde, L. L., Graudenzi, A., Korsunsky, I., Mauri, G., et al. (2015). Capri: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics* 31, 3016–3026. doi: 10.1093/bioinformatics/btv296

Rocha, D., Silva, C. J., and Torres, D. F. M. (2018). Stability and optimal control of a delayed HIV model. *Math. Methods Appl. Sci*. 41, 2251–2260. doi: 10.1002/mma.4207

Rowland, M., Tozer, T. N., Derendorf, H., and Hochhaus, G. (2011). *Clinical Pharmacokinetics and Pharmacodynamics: Concepts and Applications*. Philadelphia, PA: Wolters Kluwer Health; Lippincott William & Wilkins.

Saccomani, M. P., Audoly, S., Bellu, G., and DAngió, L. (2010). Examples of testing global identifiability of biological and biomedical models with the daisy software. *Comput. Biol. Med*. 40, 402–407. doi: 10.1016/j.compbiomed.2010.02.004

Salgado, R., Moore, H., Martens, J. W., Lively, T., Malik, S., McDermott, U., et al. (2018). Steps forward for cancer precision medicine. *Nat. Rev. Drug Discov*. 17, 1–2. doi: 10.1038/nrd.2017.218

Schwilden, H. (1981). A general method for calculating the dosage scheme in linear pharmacokinetics. *Eur. J. Clin. Pharmacol*. 20, 379–386. doi: 10.1007/BF00615409

Shah, N. P., Tran, C., Lee, F. Y., Chen, P., Norris, D., and S, C. L. (2004). Overriding imatinib resistance with a novel abl kinase inhibitor. *Science* 305, 399–401. doi: 10.1126/science.1099480

Shargel, L., Andrew, B., and Wu-Pong, S. (1999). *Applied Biopharmaceutics and Pharmacokinetics*. Stamford: Appleton & Lange Stamford.

Shi, J., Alagoz, O., Erenay, F. S., and Su, Q. (2014). A survey of optimization models on cancer chemotherapy treatment planning. *Ann. Oper. Res*. 221, 331–356. doi: 10.1007/s10479-011-0869-4

Sørensen, J. J. W. H., Pedersen, M. K., Munch, M., Haikka, P., Jensen, J. H., Planke, T., et al. (2016). Exploring the quantum speed limit with computer games. *Nature* 532, 210–213. doi: 10.1038/nature17620

Spörl, A., Schulte-Herbrüggen, T., Glaser, S. J., Bergholm, V., Storcz, M. J., Ferber, J., et al. (2007). Optimal control of coupled Josephson qubits. *Phys. Rev. A* 75:012302. doi: 10.1103/PhysRevA.75.012302

Steil, G. M. (2013). Algorithms for a closed-loop artificial pancreas: the case for proportional-integral-derivative control. *J. Diab. Sci. Technol*. 7, 1621–1631. doi: 10.1177/193229681300700623

Stiehl, T., Ho, A. D., and Marciniak-Czochra, A. (2018). Mathematical modeling of the impact of cytokine response of acute myeloid leukemia cells on patient prognosis. *Sci. Rep*. 8:2809. doi: 10.1038/s41598-018-21115-4

Stiehl, T., and Marciniak-Czochra, A. (2012). Mathematical modeling of leukemogenesis and cancer stem cell dynamics. *Math. Model. Nat. Phenomena* 7, 166–202. doi: 10.1051/mmnp/20127199

Takahashi, N., Wakita, H., Miura, M., Scott, S., Nishii, K., Masuko, M., et al. (2010). Correlation between imatinib pharmacokinetics and clinical response in Japanese patients with chronic-phase chronic myeloid leukemia. *Clin. Pharmacol. Ther*. 88, 809–813. doi: 10.1038/clpt.2010.186

Tang, M., Gonen, M., Quintas-Cardama, A., Cortes, J., Kantarjian, H., Field, C., et al. (2011). Dynamics of chronic myeloid leukemia response to long-term targeted therapy reveal treatment effects on leukemic stem cells. *Blood* 118, 1622–1631. doi: 10.1182/blood-2011-02-339267

van Frank, S., Bonneau, M., Schmiedmayer, J., Hild, S., Gross, C., Cheneau, M., et al. (2016). Optimal control of complex atomic quantum systems. *Sci. Rep.* 6:34187. doi: 10.1038/srep34187

von Mehren, M., and Widmer, N. (2011). Correlations between imatinib pharmacokinetics, pharmacodynamics, adherence, and clinical response in advanced metastatic gastrointestinal stromal tumor (GIST): an emerging role for drug blood level testing? *Cancer Treat. Rev.* 37, 291–299. doi: 10.1016/j.ctrv.2010.10.001

Weigel, M. T., Dahmke, L., Schem, C., Bauerschlag, D. O., Weber, K., Niehoff, P., et al. (2010). *In vitro* effects of imatinib mesylate on radiosensitivity and chemosensitivity of breast cancer cells. *BMC Cancer* 10:412. doi: 10.1186/1471-2407-10-412

Welling, P. G. (1997). *Pharmacokinetics: Processes, Mathematics, and Applications.* Washington, DC: American Chemical Society (ACS).

Werner, B., Scott, J. G., Sottoriva, A., Anderson, A. R., Traulsen, A., and Altrock, P. M. (2016). The cancer stem cell fraction in hierarchically organized tumors can be estimated using mathematical modeling and patient-specific treatment trajectories. *Cancer Res.* 76, 1705–1713. doi: 10.1158/0008-5472.CAN-15-2069

West, J., You, L., Brown, J., Newton, P. K., and Anderson, A. R. (2018). Towards multi-drug adaptive therapy. *bioRxiv*. 476507. doi: 10.1101/476507

Widmer, N., Decosterd, L., Csajka, C., Leyvraz, S., Duchosal, M., Rosselet, A., et al. (2006). Population pharmacokinetics of imatinib and the role of α1-acid glycoprotein. *Br. J. Clin. Pharmacol.* 62, 97–112. doi: 10.1111/j.1365-2125.2006.02719.x

Wodarz, D., Garg, N., Komarova, N. L., Benjamini, O., Keating, M. J., Wierda, W. G., et al. (2014). Kinetics of CLL cells in tissues and blood during therapy with the BTK inhibitor ibrutinib. *Blood* 123, 4132–4135. doi: 10.1182/blood-2014-02-554220

Yoon, N., Vander V, R., Marusyk, A., and Scott, J. G. (2018). Optimal therapy scheduling based on a pair of collaterally sensitive drugs. *Bull. Math. Biol.* 80, 1776–1809. doi: 10.1007/s11538-018-0434-2

Yoshitsuga, H., Imai, Y., Seriu, T., and Hiraoka, M. (2012). Markov chain Monte Carlo bayesian analysis for population pharmacokinetics of dasatinib in Japanese adult subjects with chronic myeloid leukemia and Philadelphia chromosome positive acute lymphoblastic leukemia. *J. Clin. Pharmacol. Therap*. 43, 29–41. doi: 10.3999/jscpt.43.29

Zhu, Y., and Qian, S.-X. (2014). Clinical efficacy and safety of imatinib in the management of ph+ chronic myeloid or acute lymphoblastic leukemia in Chinese patients. *OncoTargets Ther.* 7:395. doi: 10.2147/OTT.S38846

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Angaroni, Graudenzi, Rossignolo, Maspero, Calarco, Piazza, Montangero and Antoniotti. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

B

Appendix: Code repositories

Variant calling from scRNA-seq — Paper P#2

Bash and R scripts to perform the variant calling pipeline from single-cell RNA-seq data, and reproduce the results presented in the article.

github.com/BIMIB-DISCo/oral_sqamous_longitudinal

Ramazzotti, D., Angaroni, F., Maspero, D., Ascolani, G., Castiglioni, I., Piazza, R., Antoniotti, M., & Graudenzi, A.

VirMutSig — Papers P#7, P#A1

Protocol to execute the Nextflow variant calling pipeline and the R scripts for signature discovery and assignment from viral samples.

github.com/BIMIB-DISCo/VirMutSig

Maspero, D., Angaroni, F., Porro, D., Piazza, R., Graudenzi, A., & Ramazzotti, D.

MaREA4Galaxy — Paper P#3

Galaxy tool for: (*i*) projection of gene expression data onto metabolic models, (*ii*) cluster analysis, (*iii*) visualization of metabolic maps and clusters

galaxyproject.org/use/marea4galaxy/

ELIXIR server Code

Damiani, C., Rovida, L., Maspero, D., Sala, I., Rosato, L., Di Filippo, M., Pescini, D., Graudenzi, A.,Antoniotti, M., & Mauri, G.

MetNet Classification — Paper P#4

Jupyter notebook to execute the classification of cancer samples from the topological features of weighted metabolic networks, via machine learning approaches.

github.com/BIMIB-DISCo/MET-NET-CLASSIFICATION

Machicao, J.,Craighero, F., Maspero, D., Angaroni, F., Damiani, C., Graudenzi, A., Antoniotti, M. & Bruno, O. M.

LACE — Paper P#5

R package of the LACE framework for the reconstruction of models of single-tumor evolution from longitudinal single-cell data (available on Bioconductor), and repository to reproduce the analyses presented in the article.

bioconductor.org/packages/release/bioc/html/LACE.html

github.com/BIMIB-DISCo/LACE-UTILITIES

Ramazzotti, D., Angaroni, F., Maspero, D., Ascolani, G., Castiglioni, I., Piazza, R., Antoniotti, M., & Graudenzi, A.

VERSO — Paper P#6

R package of the VERSO framework for the characterization of viral evulution from deep sequencing data of viral samples (available on Bioconductor), and repository to reproduce the analyses presented in the article.

bioconductor.org/packages/release/bioc/html/VERSO.html

github.com/BIMIB-DISCo/VERSO-UTILITIES

Ramazzotti, D., Angaroni, F., Maspero, D., Gambacorti-Passerini, C., Antoniotti, M., Graudenzi, A., & Piazza, R.

COB tree — Section 3.2.1.1

R script to execute the COB tree algorithm for phylogenetic inference, and to reproduce the preliminary results on simulations.

github.com/DavideMaspero/COBtree

Maspero, D., Angaroni, F., Patruno, L., Ramazzotti, D., Graudenzi, A., & Posada, D.

FBCA — Papers P#8, P#9, P#10

Matlab script for the simulation of the Flux Balance Cellular Automata framework.

github.com/DavideMaspero/FBCA

Maspero, D., Angaroni, F., Patruno, L., Ramazzotti, D., Graudenzi, A., & Posada, D.

Bibliography

[1] G. W. Brier. "VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY". In: *Monthly Weather Review* 78.1 (Jan. 1, 1950), pp. 1–3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[2] P. a. P. Moran. "Random Processes in Genetics". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 54.1 (Jan. 1958), pp. 60–71. DOI: 10.1017/S0305004100033193.

[3] Y.-J. Chu. "On the Shortest Arborescence of a Directed Graph". In: *Scientia Sinica* 14 (1965), pp. 1396–1400.

[4] J. C. Gower. "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis". In: *Biometrika* 53.3-4 (Dec. 1, 1966), pp. 325–338. DOI: 10.1093/biomet/53.3-4.325.

[5] J. Edmonds. "Optimum Branchings". In: *Journal of Research of the National Bureau of Standards, B* 71 (1967), pp. 233–240.

[6] J. C. Gower. "Adding a Point to Vector Diagrams in Multivariate Analysis". In: *Biometrika* 55.3 (Nov. 1, 1968), pp. 582–585. DOI: 10.1093/biomet/55.3.582.

[7] S. A. Kauffman. "Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets". In: *Journal of Theoretical Biology* 22.3 (Mar. 1, 1969), pp. 437–467. DOI: 10.1016/0022-5193(69)90015-0.

[8] S. Kauffman. "Homeostasis and Differentiation in Random Genetic Control Networks". In: *Nature* 224.5215 (5215 Oct. 1969), pp. 177–178. DOI: 10.1038/224177a0.

[9] M. Kimura. "The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations". In: *Genetics* 61.4 (Apr. 1, 1969), pp. 893–903. DOI: 10.1093/genetics/61.4.893.

[10] P. C. Nowell. "The Clonal Evolution of Tumor Cell Populations". In: *Science* (Oct. 1, 1976). DOI: 10.1126/science.959840.

[11] R. E. Tarjan. "Finding Optimum Branchings". In: *Networks* 7.1 (1977), pp. 25–35. DOI: 10.1002/net.3230070103.

[12] J. Felsenstein. “Alternative Methods of Phylogenetic Inference and Their Interrelationship”. In: *Systematic Biology* 28.1 (Mar. 1, 1979), pp. 49–62. DOI: 10.1093/sysbio/28.1.49.

[13] J. Felsenstein. “Phylogenies from Molecular Sequences: Inference and Reliability”. In: *Annual Review of Genetics* 22.1 (1988), pp. 521–565. DOI: 10.1146/annurev.ge.22.120188.002513. pmid: 3071258.

[14] P. Alberch. “From Genes to Phenotype: Dynamical Systems and Evolvability”. In: *Genetica* 84.1 (May 1, 1991), pp. 5–11. DOI: 10.1007/BF00123979.

[15] A. P. Bradley. “The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms”. In: *Pattern Recognition* 30.7 (July 1, 1997), pp. 1145–1159. DOI: 10.1016/S0031-3203(96)00142-2.

[16] H. J. Bandelt, P. Forster, and A. Röhl. “Median-Joining Networks for Inferring Intraspecific Phylogenies.” In: *Molecular Biology and Evolution* 16.1 (Jan. 1, 1999), pp. 37–48. DOI: 10.1093/oxfordjournals.molbev.a026036.

[17] M. Pagel. “Inferring the Historical Patterns of Biological Evolution”. In: *Nature* 401.6756 (6756 Oct. 1999), pp. 877–884. DOI: 10.1038/44766.

[18] J. P. Huelsenbeck et al. “Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology”. In: *Science* 294.5550 (Dec. 14, 2001), pp. 2310–2314. DOI: 10.1126/science.1065889.

[19] L. A. Liotta and E. C. Kohn. “The Microenvironment of the Tumour-Host Interface”. In: *Nature* 411.6835 (May 17, 2001), pp. 375–379. DOI: 10.1038/35077241. pmid: 11357145.

[20] S. T. Sherry et al. “dbSNP: The NCBI Database of Genetic Variation”. In: *Nucleic Acids Research* 29.1 (Jan. 1, 2001), pp. 308–311. DOI: 10.1093/nar/29.1.308.

[21] G. Orphanides and D. Reinberg. “A Unified Theory of Gene Expression”. In: *Cell* 108.4 (Feb. 22, 2002), pp. 439–451. DOI: 10.1016/S0092-8674(02)00655-4. pmid: 11909516.

[22] K. R. Swanson, E. C. Alvord, and J. D. Murray. “Virtual Brain Tumours (Gliomas) Enhance the Reality of Medical Imaging and Highlight Inadequacies of Current Therapy”. In: *British Journal of Cancer* 86.1 (1 Jan. 2002), pp. 14–18. DOI: 10.1038/sj.bjc.6600021.

[23] D. Bryant. “A Classification of Consensus Methods for Phylogenetics”. In: *DIMACS series in discrete mathematics and theoretical computer science* 61 (2003), pp. 163–184.

[24] H. S. Lo et al. "Allelic Variation in Gene Expression Is Common in the Human Genome". In: *Genome Research* 13.8 (Jan. 8, 2003), pp. 1855–1862. DOI: 10.1101/gr.1006603. pmid: 12902379.

[25] H.-W. Ma and A.-P. Zeng. "The Connectivity Structure, Giant Strong Component and Centrality of Metabolic Networks". In: *Bioinformatics* 19.11 (July 22, 2003), pp. 1423–1430. DOI: 10.1093/bioinformatics/btg177.

[26] "Reductionism and Complexity in Molecular Biology". In: *EMBO reports* 5.11 (Nov. 1, 2004), pp. 1016–1020. DOI: 10.1038/sj.embor.7400284.

[27] D. Dan et al. "Solving the Advection-Diffusion Equations in Biological Contexts Using the Cellular Potts Model". In: *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* 72 (4 Pt 1 Oct. 2005), p. 041909. DOI: 10.1103/PhysRevE.72.041909. pmid: 16383422.

[28] T. Flatt. "The Evolutionary Genetics of Canalization". In: *The Quarterly Review of Biology* 80.3 (Sept. 2005), pp. 287–316. DOI: 10.1086/432265. pmid: 16250465.

[29] P. O. Lewis, M. T. Holder, and K. E. Holsinger. "Polytomies and Bayesian Phylogenetic Inference". In: *Systematic Biology* 54.2 (Apr. 1, 2005), pp. 241–253. DOI: 10.1080/10635150590924208.

[30] M. Samoilov, S. Plyasunov, and A. P. Arkin. "Stochastic Amplification and Signaling in Enzymatic Futile Cycles through Noise-Induced Bistability with Oscillations". In: *Proceedings of the National Academy of Sciences* 102.7 (Feb. 15, 2005), pp. 2310–2315. DOI: 10.1073/pnas.0406841102. pmid: 15701703.

[31] K. Kaneko. *Life: An Introduction to Complex Systems Biology*. Springer, 2006.

[32] I. A. Darby and T. D. Hewitson. "Fibroblast Differentiation in Wound Healing and Fibrosis". In: *International Review of Cytology*. Vol. 257. Academic Press, Jan. 1, 2007, pp. 143–179. DOI: 10.1016/S0074-7696(07)57004-X.

[33] M. V. Rockman. "Reverse Engineering the Genotype–Phenotype Map with Natural Genetic Variation". In: *Nature* 456.7223 (7223 Dec. 2008), pp. 738–744. DOI: 10.1038/nature07633.

[34] M. Gerstung et al. "Quantifying Cancer Progression with Conjunctive Bayesian Networks". In: *Bioinformatics* 25.21 (Nov. 1, 2009), pp. 2809–2815. DOI: 10.1093/bioinformatics/btp505.

[35] S. Huang, I. Ernberg, and S. Kauffman. "Cancer Attractors: A Systems View of Tumors from a Gene Network Dynamics and Developmental Perspective". In: *Seminars in Cell & Developmental Biology*. Structure and Function of the Golgi Apparatus and Systems Approaches to Cell and Developmental Biology 20.7 (Sept. 1, 2009), pp. 869–876. DOI: 10.1016/j.semcdb.2009.07.003.

[36] H. Li and R. Durbin. "Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform". In: *Bioinformatics* 25.14 (July 15, 2009), pp. 1754–1760. DOI: 10.1093/bioinformatics/btp324.

[37] H. Li et al. "The Sequence Alignment/Map Format and SAMtools". In: *Bioinformatics* 25.16 (Aug. 15, 2009), pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352.

[38] P. J. Park. "ChIP-seq: Advantages and Challenges of a Maturing Technology". In: *Nature Reviews. Genetics* 10.10 (Oct. 2009), pp. 669–680. DOI: 10.1038/nrg2641. pmid: 19736561.

[39] Z. Wang, M. Gerstein, and M. Snyder. "RNA-Seq: A Revolutionary Tool for Transcriptomics". In: *Nature reviews. Genetics* 10.1 (Jan. 2009), pp. 57–63. DOI: 10.1038/nrg2484. pmid: 19015660.

[40] I. Bozic et al. "Accumulation of Driver and Passenger Mutations during Tumor Progression". In: *Proceedings of the National Academy of Sciences* 107.43 (Oct. 26, 2010), pp. 18545–18550. DOI: 10.1073/pnas.1010978107. pmid: 20876136.

[41] P. J. A. Cock et al. "The Sanger FASTQ File Format for Sequences with Quality Scores, and the Solexa/Illumina FASTQ Variants". In: *Nucleic Acids Research* 38.6 (Apr. 2010), pp. 1767–1771. DOI: 10.1093/nar/gkp1137. pmid: 20015970.

[42] K. Csilléry et al. "Approximate Bayesian Computation (ABC) in Practice". In: *Trends in Ecology & Evolution* 25.7 (July 1, 2010), pp. 410–418. DOI: 10.1016/j.tree.2010.04.001. pmid: 20488578.

[43] G. Fusco and A. Minelli. "Phenotypic Plasticity in Development and Evolution: Facts and Concepts. Introduction". In: *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 365.1540 (Feb. 27, 2010), pp. 547–556. DOI: 10.1098/rstb.2009.0267. pmid: 20083631.

[44] J. D. Orth, I. Thiele, and B. Ø. Palsson. "What Is Flux Balance Analysis?" In: *Nature Biotechnology* 28.3 (3 Mar. 2010), pp. 245–248. DOI: 10.1038/nbt.1614.

[45] P. Danecek et al. "The Variant Call Format and VCFtools". In: *Bioinformatics* 27.15 (Aug. 1, 2011), pp. 2156–2158. DOI: 10.1093/bioinformatics/btr330.

[46] D. Hanahan and R. A. Weinberg. "Hallmarks of Cancer: The Next Generation". In: *Cell* 144.5 (Mar. 4, 2011), pp. 646–674. DOI: 10.1016/j.cell.2011.02.013. pmid: 21376230.

[47] A. Sottoriva, L. Vermeulen, and S. Tavaré. "Modeling Evolutionary Dynamics of Epigenetic Mutations in Hierarchically Organized Tumors". In: *PLOS Computational Biology* 7.5 (May 5, 2011), e1001132. DOI: 10.1371/journal.pcbi.1001132.

[48] M. Ala-Korpela, A. J. Kangas, and P. Soininen. "Quantitative High-Throughput Metabolomics: A New Era in Epidemiology and Genetics". In: *Genome Medicine* 4.4 (Apr. 30, 2012), p. 36. DOI: 10.1186/gm335.

[49] B. R. Angermann et al. "Computational Modeling of Cellular Signaling Processes Embedded into Dynamic Spatial Contexts". In: *Nature Methods* 9.3 (3 Mar. 2012), pp. 283–289. DOI: 10.1038/nmeth.1861.

[50] E. Domingo, J. Sheldon, and C. Perales. "Viral Quasispecies Evolution". In: *Microbiology and molecular biology reviews: MMBR* 76.2 (June 2012), pp. 159–216. DOI: 10.1128/MMBR.05023-11. pmid: 22688811.

[51] T. Hashimshony et al. "CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification". In: *Cell Reports* 2.3 (Sept. 27, 2012), pp. 666–673. DOI: 10.1016/j.celrep.2012.08.003. pmid: 22939981.

[52] J. R. Karr et al. "A Whole-Cell Computational Model Predicts Phenotype from Genotype". In: *Cell* 150.2 (July 20, 2012), pp. 389–401. DOI: 10.1016/j.cell.2012.05.044. pmid: 22817898.

[53] C. L. Kleinman, V. Adoue, and J. Majewski. "RNA Editing of Protein Sequences: A Rare Event in Human Transcriptomes". In: *RNA* 18.9 (Jan. 9, 2012), pp. 1586–1596. DOI: 10.1261/rna.033233.112. pmid: 22832026.

[54] D. C. Koboldt et al. "VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing". In: *Genome Research* 22.3 (Jan. 3, 2012), pp. 568–576. DOI: 10.1101/gr.129684.111. pmid: 22300766.

[55] B. Langmead and S. L. Salzberg. "Fast Gapped-Read Alignment with Bowtie 2". In: *Nature Methods* 9.4 (Mar. 4, 2012), pp. 357–359. DOI: 10.1038/nmeth.1923. pmid: 22388286.

[56] Y. Ni et al. "Simultaneous SNP Identification and Assessment of Allele-Specific Bias from ChIP-seq Data". In: *BMC Genetics* 13.1 (Sept. 5, 2012), p. 46. DOI: 10.1186/1471-2156-13-46.

[57] T. Sakoparnig and N. Beerenwinkel. "Efficient Sampling for Bayesian Inference of Conjunctive Bayesian Networks". In: *Bioinformatics* 28.18 (Sept. 15, 2012), pp. 2318–2324. DOI: 10.1093/bioinformatics/bts433.

[58] Z. Yang and B. Rannala. "Molecular Phylogenetics: Principles and Practice". In: *Nature Reviews Genetics* 13.5 (5 May 2012), pp. 303–314. DOI: 10.1038/nrg3186.

[59] L. B. Alexandrov et al. "Deciphering Signatures of Mutational Processes Operative in Human Cancer". In: *Cell Reports* 3.1 (Jan. 31, 2013), pp. 246–259. DOI: 10.1016/j.celrep.2012.12.008. pmid: 23318258.

[60] A.-L. Barabási. "Network Science". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371.1987 (2013), p. 20120375.

[61] J. Claus et al. "Spatial Aspects in the SMAD Signaling Pathway". In: *Journal of Mathematical Biology* 67.5 (Nov. 1, 2013), pp. 1171–1197. DOI: 10.1007/s00285-012-0574-1.

[62] G. De Matteis, A. Graudenzi, and M. Antoniotti. "A Review of Spatial Computational Models for Multi-Cellular Systems, with Regard to Intestinal Crypts and Colorectal Cancer Development". In: *Journal of Mathematical Biology* 66.7 (June 1, 2013), pp. 1409–1462. DOI: 10.1007/s00285-012-0539-4.

[63] A. Dobin et al. "STAR: Ultrafast Universal RNA-seq Aligner". In: *Bioinformatics* 29.1 (Jan. 1, 2013), pp. 15–21. DOI: 10.1093/bioinformatics/bts635.

[64] S. Huang and S. Kauffman. "How to Escape the Cancer Attractor: Rationale and Limitations of Multi-Target Drugs". In: *Seminars in Cancer Biology*. Cancer-Related Networks: A Help to Understand, Predict and Change Malignant Transformation 23.4 (Aug. 1, 2013), pp. 270–278. DOI: 10.1016/j.semcancer.2013.06.003.

[65] J. Ladyman, J. Lambert, and K. Wiesner. "What Is a Complex System?" In: *European Journal for Philosophy of Science* 3.1 (Jan. 1, 2013), pp. 33–67. DOI: 10.1007/s13194-012-0056-8.

[66] C. E. Meacham and S. J. Morrison. "Tumour Heterogeneity and Cancer Cell Plasticity". In: *Nature* 501.7467 (Sept. 19, 2013), pp. 328–337. DOI: 10.1038/nature12624. pmid: 24048065.

[67] H. L. Rehm. "Disease-Targeted Sequencing: A Cornerstone in the Clinic". In: *Nature Reviews Genetics* 14.4 (4 Apr. 2013), pp. 295–300. DOI: 10.1038/nrg3463.

[68] E. Shapiro, T. Biezuner, and S. Linnarsson. "Single-Cell Sequencing-Based Technologies Will Revolutionize Whole-Organism Science". In: *Nature Reviews Genetics* 14.9 (9 Sept. 2013), pp. 618–630. DOI: 10.1038/nrg3542.

[69] K. Smallbone et al. "A Model of Yeast Glycolysis Based on a Consistent Kinetic Characterisation of All Its Enzymes". In: *FEBS Letters* 587.17 (2013), pp. 2832–2841. DOI: 10.1016/j.febslet.2013.06.043.

[70] A. Sottoriva et al. "Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics". In: *Proceedings of the National Academy of Sciences* 110.10 (2013), pp. 4009–4014.

[71] X. Sun et al. "Systems Modeling of Anti-apoptotic Pathways in Prostate Cancer: Psychological Stress Triggers a Synergism Pattern Switch in Drug Combination Therapy". In: *PLOS Computational Biology* 9.12 (Dec. 5, 2013), e1003358. DOI: 10.1371/journal.pcbi.1003358.

[72] B. Vogelstein et al. "Cancer Genome Landscapes". In: *Science (New York, N.Y.)* 339.6127 (Mar. 29, 2013), pp. 1546–1558. DOI: 10.1126/science.1235122. pmid: 23539594.

[73] J. Walpole, J. A. Papin, and S. M. Peirce. "Multiscale Computational Models of Complex Biological Systems". In: *Annual Review of Biomedical Engineering* 15.1 (2013), pp. 137–154. DOI: 10.1146/annurev-bioeng-071811-150104. pmid: 23642247.

[74] K. A. Wetterstrand. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. Genome.gov. 2013.

[75] A. M. Bolger, M. Lohse, and B. Usadel. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data". In: *Bioinformatics* 30.15 (Aug. 1, 2014), pp. 2114–2120. DOI: 10.1093/bioinformatics/btu170.

[76] A. Bordbar et al. "Constraint-Based Models Predict Metabolic and Associated Cellular Functions". In: *Nature Reviews Genetics* 15.2 (2 Feb. 2014), pp. 107–120. DOI: 10.1038/nrg3643.

[77] R. Bouckaert et al. "BEAST 2: A Software Platform for Bayesian Evolutionary Analysis". In: *PLOS Computational Biology* 10.4 (Apr. 10, 2014), e1003537. DOI: 10.1371/journal.pcbi.1003537.

[78] I. Bozic and M. A. Nowak. "Timing and Heterogeneity of Mutations Associated with Drug Resistance in Metastatic Cancers". In: *Proceedings of the National Academy of Sciences* 111.45 (Nov. 11, 2014), pp. 15964–15968. DOI: 10.1073/pnas.1412075111. pmid: 25349424.

[79] J. Foo and F. Michor. "Evolution of Acquired Resistance to Anti-Cancer Therapy". In: *Journal of Theoretical Biology* 355 (Aug. 21, 2014), pp. 10–20. DOI: 10.1016/j.jtbi.2014.02.025.

[80] K. Leder et al. "Mathematical Modeling of PDGF-Driven Glioblastoma Reveals Optimized Radiation Dosing Schedules". In: *Cell* 156.3 (Jan. 30, 2014), pp. 603–616. DOI: 10.1016/j.cell.2013.12.029. pmid: 24485463.

[81] W. W. Lytton, S. A. Neymotin, and C. C. Kerr. "Multiscale Modeling for Clinical Translation in Neuropsychiatric Disease". In: *Journal of Computational Surgery* 1.1 (Mar. 3, 2014), p. 7. DOI: 10.1186/2194-3990-1-7.

[82] A. Mardinoglu et al. "Genome-Scale Metabolic Modelling of Hepatocytes Reveals Serine Deficiency in Patients with Non-Alcoholic Fatty Liver Disease". In: *Nature Communications* 5.1 (1 Jan. 14, 2014), p. 3083. DOI: 10.1038/ncomms4083.

[83] D. Merkel. "Docker: Lightweight Linux Containers for Consistent Development and Deployment". In: *Linux journal* 2014.239 (2014), p. 2.

[84] S. Picelli et al. "Full-Length RNA-seq from Single Cells Using Smart-seq2". In: *Nature Protocols* 9.1 (Jan. 2014), pp. 171–181. DOI: 10.1038/nprot.2014.006. pmid: 24385147.

[85] J. Podnar et al. "Next-Generation Sequencing RNA-Seq Library Construction". In: *Current Protocols in Molecular Biology* 106.1 (2014), pp. 4.21.1–4.21.19. DOI: 10.1002/0471142727.mb0421s106.

[86] A. Roth et al. "PyClone: Statistical Inference of Clonal Population Structure in Cancer". In: *Nature Methods* 11.4 (4 Apr. 2014), pp. 396–398. DOI: 10.1038/nmeth.2883.

[87] D. Sims et al. "Sequencing Depth and Coverage: Key Considerations in Genomic Analyses". In: *Nature Reviews Genetics* 15.2 (2 Feb. 2014), pp. 121–132. DOI: 10.1038/nrg3642.

[88] M. P. H. Stumpf. "Approximate Bayesian Inference for Complex Ecosystems". In: *F1000prime Reports* 6 (2014), p. 60. DOI: 10.12703/P6-60. pmid: 25152812.

[89] E. L. van Dijk et al. "Ten Years of Next-Generation Sequencing Technology". In: *Trends in Genetics* 30.9 (Sept. 1, 2014), pp. 418–426. DOI: 10.1016/j.tig.2014.07.001.

[90] P. M. Altrock, L. L. Liu, and F. Michor. "The Mathematics of Cancer: Integrating Quantitative Models". In: *Nature Reviews Cancer* 15.12 (12 Dec. 2015), pp. 730–745. DOI: 10.1038/nrc4029.

[91] N. Beerenwinkel et al. "Cancer Evolution: Mathematical Models and Computational Inference". In: *Systematic Biology* 64.1 (Jan. 1, 2015), e1–e25. DOI: 10.1093/sysbio/syu081.

[92] J. D. Buenrostro et al. "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide". In: *Current Protocols in Molecular Biology* 109.1 (2015), pp. 21.29.1–21.29.9. DOI: 10.1002/0471142727.mb2129s109.

[93] J. Carrera and M. W. Covert. "Why Build Whole-Cell Models?" In: *Trends in Cell Biology* 25.12 (Dec. 1, 2015), pp. 719–722. DOI: 10.1016/j.tcb.2015.09.004. pmid: 26471224.

[94] F. Danielsson et al. "Assessing the Consistency of Public Human Tissue RNA-seq Data Sets". In: *Briefings in Bioinformatics* 16.6 (Nov. 1, 2015), pp. 941–949. DOI: 10.1093/bib/bbv017.

[95] F. A. Di Lello, A. C. A. Culasso, and R. H. Campos. "Inter and Intrapatient Evolution of Hepatitis C Virus". In: *Annals of Hepatology* 14.4 (2015), pp. 442–449. pmid: 26019029.

[96] M. El-Kebir et al. "Reconstruction of Clonal Trees and Tumor Composition from Multi-Sample Sequencing Data". In: *Bioinformatics* 31.12 (June 15, 2015), pp. i62–i70. DOI: 10.1093/bioinformatics/btv261.

[97] J. Kuipers and G. Moffa. "Uniform Random Generation of Large Acyclic Digraphs". In: *Statistics and Computing* 25.2 (Mar. 1, 2015), pp. 227–242. DOI: 10.1007/s11222-013-9428-y.

[98] I. C. Macaulay et al. "G&T-seq: Parallel Sequencing of Single-Cell Genomes and Transcriptomes". In: *Nature Methods* 12.6 (6 June 2015), pp. 519–522. DOI: 10.1038/nmeth.3370.

[99] E. K. Markert and A. Vazquez. "Mathematical Models of Cancer Metabolism". In: *Cancer & Metabolism* 3.1 (Dec. 22, 2015), p. 14. DOI: 10.1186/s40170-015-0140-6.

[100] M. Melé et al. "The Human Transcriptome across Tissues and Individuals". In: *Science* (May 8, 2015). DOI: 10.1126/science.aaa0355.

[101] L.-T. Nguyen et al. "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies". In: *Molecular Biology and Evolution* 32.1 (Jan. 1, 2015), pp. 268–274. DOI: 10.1093/molbev/msu300.

[102] J. M. Osborne. "Multiscale Model of Colorectal Cancer Using the Cellular Potts Framework". In: *Cancer Informatics* 14s4 (Jan. 1, 2015), CIN.S19332. DOI: 10.4137/CIN.S19332.

[103] D. Ramazzotti et al. "CAPRI: Efficient Inference of Cancer Progression Models from Cross-Sectional Data". In: *Bioinformatics* 31.18 (Sept. 15, 2015), pp. 3016–3026. DOI: 10.1093/bioinformatics/btv296.

[104] A. Sottoriva et al. "A Big Bang Model of Human Colorectal Tumor Growth". In: *Nature Genetics* 47.3 (3 Mar. 2015), pp. 209–216. DOI: 10.1038/ng.3214.

[105] Y. Wang and N. E. Navin. "Advances and Applications of Single-Cell Sequencing Technologies". In: *Molecular Cell* 58.4 (May 21, 2015), pp. 598–609. DOI: 10.1016/j.molcel.2015.05.005.

[106] P. J. Albert and U. S. Schwarz. "Dynamics of Cell Ensembles on Adhesive Micropatterns: Bridging the Gap between Single Cell Spreading and Collective Cell Migration". In: *PLOS Computational Biology* 12.4 (Apr. 7, 2016), e1004863. DOI: 10.1371/journal.pcbi.1004863.

[107] L. B. Alexandrov et al. "Mutational Signatures Associated with Tobacco Smoking in Human Cancer". In: *Science* (Nov. 4, 2016). DOI: 10.1126/science.aag0299.

[108] D. Arendt et al. "The Origin and Evolution of Cell Types". In: *Nature Reviews Genetics* 17.12 (12 Dec. 2016), pp. 744–757. DOI: 10.1038/nrg.2016.127.

[109] W. F. Doolittle and T. D. P. Brunet. "What Is the Tree of Life?" In: *PLOS Genetics* 12.4 (Apr. 14, 2016), e1005912. DOI: 10.1371/journal.pgen.1005912.

[110] T. Hashimshony et al. "CEL-Seq2: Sensitive Highly-Multiplexed Single-Cell RNA-Seq". In: *Genome Biology* 17.1 (Apr. 28, 2016), p. 77. DOI: 10.1186/s13059-016-0938-8.

[111] Y. Hu et al. "Simultaneous Profiling of Transcriptome and DNA Methylome from a Single Cell". In: *Genome Biology* 17.1 (May 5, 2016), p. 88. DOI: 10.1186/s13059-016-0950-z.

[112] L. A. Hug et al. "A New View of the Tree of Life". In: *Nature Microbiology* 1.5 (5 Apr. 11, 2016), pp. 1–6. DOI: 10.1038/nmicrobiol.2016.48.

[113] K. Jahn, J. Kuipers, and N. Beerenwinkel. "Tree Inference for Single-Cell Data". In: *Genome Biology* 17.1 (May 5, 2016), p. 86. DOI: 10.1186/s13059-016-0936-x.

[114] M. El-Kebir et al. "Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures". In: *Cell Systems* 3.1 (July 27, 2016), pp. 43–53. DOI: 10.1016/j.cels.2016.07.004. pmid: 27467246.

[115] M. Mojtahedi et al. "Cell Fate Decision as High-Dimensional Critical State Transition". In: *PLOS Biology* 14.12 (Dec. 27, 2016), e2000640. DOI: 10.1371/journal.pbio.2000640.

[116] D. Aran, Z. Hu, and A. J. Butte. "xCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape". In: *Genome Biology* 18.1 (2017). DOI: 10.1186/s13059-017-1349-1.

[117] E. Borgiani, G. N. Duda, and S. Checa. "Multiscale Modeling of Bone Healing: Toward a Systems Biology Approach". In: *Frontiers in Physiology* 8 (2017), p. 287. DOI: 10.3389/fphys.2017.00287.

[118] C. Damiani et al. "A Metabolic Core Model Elucidates How Enhanced Utilization of Glucose and Glutamine, with Enhanced Glutamine-Dependent Lactate Production, Promotes Cancer Cell Growth: The WarburQ Effect". In: *PLOS Computational Biology* 13.9 (Sept. 28, 2017), e1005758. DOI: 10.1371/journal.pcbi.1005758.

[119] C. Damiani et al. "popFBA: Tackling Intratumour Heterogeneity with Flux Balance Analysis". In: *Bioinformatics* 33.14 (July 15, 2017), pp. i311–i318. DOI: 10.1093/bioinformatics/btx251.

[120] A. Davis, R. Gao, and N. Navin. "Tumor Evolution: Linear, Branching, Neutral or Punctuated?" In: *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*. Evolutionary Principles - Heterogeneity in Cancer? 1867.2 (Apr. 1, 2017), pp. 151–161. DOI: 10.1016/j.bbcan.2017.01.003.

[121] J. B. Dayton and S. R. Piccolo. "Classifying Cancer Genome Aberrations by Their Mutually Exclusive Effects on Transcription". In: *BMC Medical Genomics* 10 (Suppl 4 Dec. 21, 2017), p. 66. DOI: 10.1186/s12920-017-0303-0. pmid: 29322935.

[122] P. Di Tommaso et al. "Nextflow Enables Reproducible Computational Workflows". In: *Nature Biotechnology* 35.4 (4 Apr. 2017), pp. 316–319. DOI: 10.1038/nbt.3820.

[123] M. G. V. Heiden and R. J. DeBerardinis. "Understanding the Intersections between Metabolism and Cancer Biology". In: *Cell* 168.4 (Feb. 9, 2017), pp. 657–669. DOI: 10.1016/j.cell.2016.12.039. pmid: 28187287.

[124] T. Hirashima, E. G. Rens, and R. M. H. Merks. "Cellular Potts Modeling of Complex Multicellular Behaviors in Tissue Morphogenesis". In: *Development, Growth & Differentiation* 59.5 (2017), pp. 329–339. DOI: 10.1111/dgd.12358.

[125] C. Isella et al. "Selective Analysis of Cancer-Cell Intrinsic Transcriptional Traits Defines Novel Clinically Relevant Subtypes of Colorectal Cancer". In: *Nature Communications* 8.1 (1 May 31, 2017), p. 15107. DOI: 10.1038/ncomms15107.

[126] N. B. Jamieson and A. V. Maker. "Gene-Expression Profiling to Predict Responsiveness to Immunotherapy". In: *Cancer Gene Therapy* 24.3 (3 Mar. 2017), pp. 134–140. DOI: 10.1038/cgt.2016.63.

[127] R. Kamps et al. "Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification". In: *International Journal of Molecular Sciences* 18.2 (2 Feb. 2017), p. 308. DOI: 10.3390/ijms18020308.

[128] E. Lein, L. E. Borm, and S. Linnarsson. "The Promise of Spatial Transcriptomics for Neuroscience in the Era of Molecular Cell Typing". In: *Science* (Oct. 6, 2017). DOI: 10.1126/science.aan6827.

[129] A. Ma'ayan. "Complex Systems Biology". In: *Journal of The Royal Society Interface* 14.134 (Sept. 30, 2017), p. 20170391. DOI: 10.1098/rsif.2017.0391.

[130] A. Malik et al. "Parallel Embryonic Transcriptional Programs Evolve under Distinct Constraints and May Enable Morphological Conservation amidst Adaptation". In: *Developmental Biology* 430.1 (Oct. 1, 2017), pp. 202–213. DOI: 10.1016/j.ydbio.2017.07.019.

[131] A. E. Moor and S. Itzkovitz. "Spatial Transcriptomics: Paving the Way for Tissue-Level Systems Biology". In: *Current Opinion in Biotechnology*. Systems Biology • Nanobiotechnology 46 (Aug. 1, 2017), pp. 126–133. DOI: 10.1016/j.copbio.2017.02.004.

[132] A. Nilsson and J. Nielsen. "Genome Scale Metabolic Modeling of Cancer". In: *Metabolic Engineering* 43 (Pt B Sept. 2017), pp. 103–112. DOI: 10.1016/j.ymben.2016.10.022. pmid: 27825806.

[133] J. Quick et al. "Multiplex PCR Method for MinION and Illumina Sequencing of Zika and Other Virus Genomes Directly from Clinical Samples". In: *Nature Protocols* 12.6 (6 June 2017), pp. 1261–1276. DOI: 10.1038/nprot.2017.066.

[134] E. G. Rens and R. M. H. Merks. "Cell Contractility Facilitates Alignment of Cells and Tissues to Static Uniaxial Stretch". In: *Biophysical Journal* 112.4 (Feb. 28, 2017), pp. 755–766. DOI: 10.1016/j.bpj.2016.12.012. pmid: 28256235.

[135] S. Salehi et al. "ddClone: Joint Statistical Inference of Clonal Populations from Single Cell and Bulk Tumour Sequencing Data". In: *Genome Biology* 18.1 (Mar. 1, 2017), p. 44. DOI: 10.1186/s13059-017-1169-3.

[136] D. Schnoerr, G. Sanguinetti, and R. Grima. "Approximation and Inference Methods for Stochastic Biochemical Kinetics—a Tutorial Review". In: *Journal of Physics A: Mathematical and Theoretical* 50.9 (2017), p. 093001. DOI: 10.1007/s11538-018-0443-1.

[137] R. Schwartz and A. A. Schäffer. "The Evolution of Tumour Phylogenetics: Principles and Practice". In: *Nature Reviews Genetics* 18.4 (4 Apr. 2017), pp. 213–229. DOI: 10.1038/nrg.2016.170.

[138] Y. Shu and J. McCauley. "GISAID: Global Initiative on Sharing All Influenza Data – from Vision to Reality". In: *Eurosurveillance* 22.13 (Mar. 30, 2017), p. 30494. DOI: 10.2807/1560-7917.ES.2017.22.13.30494.

[139] M. Stoeckius et al. "Simultaneous Epitope and Transcriptome Measurement in Single Cells". In: *Nature Methods* 14.9 (9 Sept. 2017), pp. 865–868. DOI: 10.1038/nmeth.4380.

[140] L. Zappia, B. Phipson, and A. Oshlack. "Splatter: Simulation of Single-Cell RNA Sequencing Data". In: *Genome Biology* 18.1 (Sept. 12, 2017), p. 174. DOI: 10.1186/s13059-017-1305-0.

[141] C. Ziegenhain et al. "Comparative Analysis of Single-Cell RNA Sequencing Methods". In: *Molecular Cell* 65.4 (Feb. 16, 2017), 631–643.e4. DOI: 10.1016/j.molcel.2017.01.023. pmid: 28212749.

[142] E. Afgan et al. "The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2018 Update". In: *Nucleic Acids Research* 46.W1 (July 2, 2018), W537–W544. DOI: 10.1093/nar/gky379.

[143] E. Brunk et al. "Recon3D Enables a Three-Dimensional View of Gene Variation in Human Metabolism". In: *Nature Biotechnology* 36.3 (3 Mar. 2018), pp. 272–281. DOI: 10.1038/nbt.4072.

[144] G. Caravagna et al. "Detecting Repeated Cancer Evolution from Multi-Region Tumor Sequencing Data". In: *Nature Methods* 15.9 (9 Sept. 2018), pp. 707–714. DOI: 10.1038/s41592-018-0108-x.

[145] J. Cheng et al. "Metabolomics: A High-Throughput Platform for Metabolite Profile Exploration". In: *Computational Systems Biology: Methods and Protocols*. Ed. by T. Huang. Methods in Molecular Biology. New York, NY: Springer, 2018, pp. 265–292. ISBN: 978-1-4939-7717-8. DOI: 10.1007/978-1-4939-7717-8_16.

[146] S. J. Clark et al. "scNMT-seq Enables Joint Profiling of Chromatin Accessibility DNA Methylation and Transcription in Single Cells". In: *Nature Communications* 9.1 (1 Feb. 22, 2018), p. 781. DOI: 10.1038/s41467-018-03149-4.

[147] A. H. Corbett. "Post-Transcriptional Regulation of Gene Expression and Human Disease". In: *Current Opinion in Cell Biology*. Cell Nucleus 52 (June 1, 2018), pp. 96–104. DOI: 10.1016/j.ceb.2018.02.011.

[148] I. Dagogo-Jack and A. T. Shaw. "Tumour Heterogeneity and Resistance to Cancer Therapies". In: *Nature Reviews Clinical Oncology* 15.2 (2 Feb. 2018), pp. 81–94. DOI: 10.1038/nrclinonc.2017.166.

[149] F. de Velde et al. "Clinical Applications of Population Pharmacokinetic Models of Antibiotics: Challenges and Perspectives". In: *Pharmacological Research* 134 (Aug. 2018), pp. 280–288. DOI: 10.1016/j.phrs.2018.07.005. pmid: 30033398.

[150] C. Evans, J. Hardin, and D. M. Stoebel. "Selecting Between-Sample RNA-Seq Normalization Methods from the Perspective of Their Assumptions". In: *Briefings in Bioinformatics* 19.5 (Sept. 28, 2018), pp. 776–792. DOI: 10.1093/bib/bbx008.

[151] A. Ghaffarizadeh et al. "PhysiCell: An Open Source Physics-Based Cell Simulator for 3-D Multicellular Systems". In: *PLOS Computational Biology* 14.2 (Feb. 23, 2018), e1005991. DOI: 10.1371/journal.pcbi.1005991.

[152] H. Gong, D. Do, and R. Ramakrishnan. "Single-Cell mRNA-Seq Using the Fluidigm C1 System and Integrated Fluidics Circuits". In: *Gene Expression Analysis: Methods and Protocols*. Ed. by N. Raghavachari and N. Garcia-Reyero. Methods in Molecular Biology. New York, NY: Springer, 2018, pp. 193–207. ISBN: 978-1-4939-7834-2. DOI: 10.1007/978-1-4939-7834-2_10.

[153] K. Govek, C. Sikes, and L. Oesper. "A Consensus Approach to Infer Tumor Evolutionary Histories". In: *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. BCB '18. New York, NY, USA: Association for Computing Machinery, Aug. 15, 2018, pp. 63–72. ISBN: 978-1-4503-5794-4. DOI: 10.1145/3233547.3233584.

[154] A. Graudenzi et al. "Integration of Transcriptomic Data and Metabolic Networks in Cancer Samples Reveals Highly Significant Prognostic Power". In: *Journal of Biomedical Informatics* 87 (Nov. 1, 2018), pp. 37–49. DOI: 10.1016/j.jbi.2018.09.010.

[155] J. Hadfield et al. "Nextstrain: Real-Time Tracking of Pathogen Evolution". In: *Bioinformatics* 34.23 (Dec. 1, 2018), pp. 4121–4123. DOI: 10.1093/bioinformatics/bty407.

[156] Y.-J. Ho et al. "Single-Cell RNA-seq Analysis Identifies Markers of Resistance to Targeted BRAF Inhibitors in Melanoma Cell Populations". In: *Genome Research* 28.9 (Jan. 9, 2018), pp. 1353–1363. DOI: 10.1101/gr.234062.117. pmid: 30061114.

[157] T. Höfer and H.-R. Rodewald. "Differentiation-Based Model of Hematopoietic Stem Cell Functions and Lineage Pathways". In: *Blood* 132.11 (Sept. 13, 2018), pp. 1106–1113. DOI: 10.1182/blood-2018-03-791517.

[158] G. La Manno et al. "RNA Velocity of Single Cells". In: *Nature* 560.7719 (7719 Aug. 2018), pp. 494–498. DOI: 10.1038/s41586-018-0414-6.

[159] L. McInnes et al. "UMAP: Uniform Manifold Approximation and Projection". In: *Journal of Open Source Software* 3.29 (Sept. 2, 2018), p. 861. DOI: 10.21105/joss.00861.

[160] NCBI Resource Coordinators. "Database Resources of the National Center for Biotechnology Information". In: *Nucleic Acids Research* 46.D1 (Jan. 4, 2018), pp. D8–D13. DOI: 10.1093/nar/gkx1095.

[161] J. E. O'Reilly and P. C. J. Donoghue. "The Efficacy of Consensus Tree Methods for Summarizing Phylogenetic Relationships from a Posterior Sample of Trees Estimated from Morphological Data". In: *Systematic Biology* 67.2 (Mar. 1, 2018), pp. 354–362. DOI: 10.1093/sysbio/syx086.

[162] E. Papalexi and R. Satija. "Single-Cell RNA Sequencing to Explore Immune Cell Heterogeneity". In: *Nature Reviews Immunology* 18.1 (1 Jan. 2018), pp. 35–45. DOI: 10.1038/nri.2017.76.

[163] F. Rambow et al. "Toward Minimal Residual Disease-Directed Therapy in Melanoma". In: *Cell* 174.4 (Aug. 9, 2018), 843–855.e19. DOI: 10.1016/j.cell.2018.06.025.

[164] S. L. Salzberg. "Open Questions: How Many Genes Do We Have?" In: *BMC Biology* 16.1 (Aug. 20, 2018), p. 94. DOI: 10.1186/s12915-018-0564-x.

[165] A. Sharma et al. "Longitudinal Single-Cell RNA Sequencing of Patient-Derived Primary Cells Reveals Drug-Induced Infidelity in Stem Cell Hierarchy". In: *Nature Communications* 9.1 (1 Nov. 22, 2018), p. 4931. DOI: 10.1038/s41467-018-07261-3.

[166] B. E. Slatko, A. F. Gardner, and F. M. Ausubel. "Overview of Next-Generation Sequencing Technologies". In: *Current Protocols in Molecular Biology* 122.1 (2018), e59. DOI: 10.1002/cpmb.59.

[167] L. Valihrach, P. Androvic, and M. Kubista. "Platforms for Single-Cell Collection and Analysis". In: *International Journal of Molecular Sciences* 19.3 (3 Mar. 2018), p. 807. DOI: 10.3390/ijms19030807.

[168] F. A. Wolf, P. Angerer, and F. J. Theis. "SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis". In: *Genome Biology* 19.1 (Feb. 6, 2018), p. 15. DOI: 10.1186/s13059-017-1382-0.

[169] N. Aguse, Y. Qi, and M. El-Kebir. "Summarizing the Solution Space in Tumor Phylogeny Inference by Multiple Consensus Trees". In: *Bioinformatics* 35.14 (July 15, 2019), pp. i408–i416. DOI: 10.1093/bioinformatics/btz312.

[170] L. Allen, A. O'Connell, and V. Kiermer. "How Can We Ensure Visibility and Diversity in Research Contributions? How the Contributor Role Taxonomy (CRediT) Is Helping the Shift from Authorship to Contributorship". In: *Learned Publishing* 32.1 (Jan. 2019), pp. 71–74. DOI: 10.1002/leap.1210.

[171] D. Aran et al. "Reference-Based Analysis of Lung Single-Cell Sequencing Reveals a Transitional Profibrotic Macrophage". In: *Nature Immunology* 20.2 (2 Feb. 2019), pp. 163–172. DOI: 10.1038/s41590-018-0276-y.

[172] I. Arozarena and C. Wellbrock. "Phenotype Plasticity as Enabler of Melanoma Progression and Therapy Resistance". In: *Nature Reviews Cancer* 19.7 (7 July 2019), pp. 377–391. DOI: 10.1038/s41568-019-0154-4.

[173] J. Bageritz and G. Raddi. "Single-Cell RNA Sequencing with Drop-Seq". In: *Single Cell Methods: Sequencing and Proteomics*. Ed. by V. Proserpio. Methods in Molecular Biology. New York, NY: Springer, 2019, pp. 73–85. ISBN: 978-1-4939-9240-9. DOI: 10.1007/978-1-4939-9240-9_6.

[174] K. Chkhaidze et al. "Spatially Constrained Tumour Growth Affects the Patterns of Clonal Selection and Neutral Drift in Cancer Genomic Data". In: *PLOS Computational Biology* 15.7 (July 29, 2019), e1007243. DOI: 10.1371/journal.pcbi.1007243.

[175] C. Damiani et al. "Integration of Single-Cell RNA-seq Data into Population Models to Characterize Cancer Metabolism". In: *PLOS Computational Biology* 15.2 (Feb. 28, 2019), e1006733. DOI: 10.1371/journal.pcbi.1006733.

[176] R. Diaz-Uriarte and C. Vasallo. "Every Which Way? On Predicting Tumor Evolution Using Cancer Progression Models". In: *PLOS Computational Biology* 15.8 (Aug. 2, 2019), e1007246. DOI: 10.1371/journal.pcbi.1007246.

[177] P. B. Gupta et al. "Phenotypic Plasticity: Driver of Cancer Initiation, Progression, and Therapy Resistance". In: *Cell stem cell* 24.1 (2019), pp. 65–78.

[178] D. C. Hinshaw and L. A. Shevde. "The Tumor Microenvironment Innately Modulates Cancer Progression". In: *Cancer Research* 79.18 (Sept. 15, 2019), pp. 4557–4566. DOI: 10.1158/0008-5472.CAN-18-3962. pmid: 31350295.

[179] S.-R. Hosseini et al. "Estimating the Predictability of Cancer Evolution". In: *Bioinformatics* 35.14 (July 15, 2019), pp. i389–i397. DOI: 10.1093/bioinformatics/btz332.

[180] J. E. Jansen et al. "Combining Mathematical Models With Experimentation to Drive Novel Mechanistic Insights Into Macrophage Function". In: *Frontiers in Immunology* 10 (2019), p. 1283. DOI: 10.3389/fimmu.2019.01283.

[181] K. E. Johnson et al. "Cancer Cell Population Growth Kinetics at Low Densities Deviate from the Exponential Growth Model and Suggest an Allee Effect". In: *PLOS Biology* 17.8 (Aug. 5, 2019), e3000399. DOI: 10.1371/journal.pbio.3000399.

[182] K. R. Kumar, M. J. Cowley, and R. L. Davis. "Next-Generation Sequencing and Emerging Technologies". In: *Seminars in Thrombosis and Hemostasis* 45.7 (Oct. 2019), pp. 661–673. DOI: 10.1055/s-0039-1688446.

[183] Q. Li et al. "A Bayesian Hidden Potts Mixture Model for Analyzing Lung Cancer Pathology Images". In: *Biostatistics (Oxford, England)* 20.4 (Oct. 1, 2019), pp. 565–581. DOI: 10.1093/biostatistics/kxy019. pmid: 29788035.

[184] G. Lightbody et al. "Review of Applications of High-Throughput Sequencing in Personalized Medicine: Barriers and Facilitators of Future Progress in Research and Clinical Application". In: *Briefings in Bioinformatics* 20.5 (Sept. 27, 2019), pp. 1795–1811. DOI: 10.1093/bib/bby051.

[185] F. Liu et al. "Systematic Comparative Analysis of Single-Nucleotide Variant Detection Methods from Single-Cell RNA Sequencing Data". In: *Genome Biology* 20.1 (Nov. 19, 2019), p. 242. DOI: 10.1186/s13059-019-1863-4.

[186] M. D. Luecken and F. J. Theis. "Current Best Practices in Single-Cell RNA-seq Analysis: A Tutorial". In: *Molecular Systems Biology* 15.6 (2019), e8746. DOI: 10.15252/msb.20188746. eprint: https://www.embopress.org/doi/pdf/10.15252/msb.20188746.

[187] M. Mahmoud et al. "Structural Variant Calling: The Long and the Short of It". In: *Genome Biology* 20.1 (Nov. 20, 2019), p. 246. DOI: 10.1186/s13059-019-1828-7.

[188] J. Metzcar et al. "A Review of Cell-Based Computational Modeling in Cancer Biology". In: *JCO Clinical Cancer Informatics* 3 (Dec. 1, 2019), pp. 1–13. DOI: 10.1200/CCI.18.00069.

[189] M. A. Myers, G. Satas, and B. J. Raphael. "CALDER: Inferring Phylogenetic Trees from Longitudinal Tumor Samples". In: *Cell Systems* 8.6 (June 26, 2019), 514–522.e5. DOI: 10.1016/j.cels.2019.05.010.

[190] K. Nakayama and N. Kataoka. "Regulation of Gene Expression under Hypoxic Conditions". In: *International Journal of Molecular Sciences* 20.13 (July 3, 2019), E3278. DOI: 10.3390/ijms20133278. pmid: 31277312.

[191] A. S. Nam et al. "Somatic Mutations and Cell Identity Linked by Genotyping of Transcriptomes". In: *Nature* 571.7765 (7765 July 2019), pp. 355–360. DOI: 10.1038/s41586-019-1367-0.

[192] D. Ramazzotti et al. "Learning Mutational Graphs of Individual Tumour Evolution from Single-Cell and Multi-Region Sequencing Data". In: *BMC Bioinformatics* 20.1 (Apr. 25, 2019), p. 210. DOI: 10.1186/s12859-019-2795-4.

[193] O. Röhrle et al. "Multiscale Modeling of the Neuromuscular System: Coupling Neurophysiology and Skeletal Muscle Mechanics". In: *WIREs Systems Biology and Medicine* 11.6 (2019), e1457. DOI: 10.1002/wsbm.1457.

[194] J. A. Rohrs, P. Wang, and S. D. Finley. "Understanding the Dynamics of T-Cell Activation in Health and Disease Through the Lens of Computational Modeling". In: *JCO Clinical Cancer Informatics* 3 (Dec. 1, 2019), pp. 1–8. DOI: 10.1200/CCI.18.00057.

[195] P. M. Schnepp et al. “SNV Identification from Single-Cell RNA Sequencing Data”. In: *Human Molecular Genetics* 28.21 (Nov. 1, 2019), pp. 3569–3583. DOI: 10.1093/hmg/ddz207.

[196] M. Setty et al. “Characterization of Cell Fate Probabilities in Single-Cell Data with Palantir”. In: *Nature Biotechnology* 37.4 (4 Apr. 2019), pp. 451–460. DOI: 10.1038/s41587-019-0068-4.

[197] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons. “Machine Learning in Medicine: A Practical Introduction”. In: *BMC Medical Research Methodology* 19.1 (Mar. 19, 2019), p. 64. DOI: 10.1186/s12874-019-0681-4.

[198] R. Stark, M. Grzelak, and J. Hadfield. “RNA Sequencing: The Teenage Years”. In: *Nature Reviews Genetics* 20.11 (11 Nov. 2019), pp. 631–656. DOI: 10.1038/s41576-019-0150-2.

[199] A. A. Tabl et al. “A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer”. In: *Frontiers in Genetics* 10 (2019), p. 256. DOI: 10.3389/fgene.2019.00256.

[200] H. Zafar et al. “SiCloneFit: Bayesian Inference of Population Structure, Genotype, and Phylogeny of Tumor Clones from Single-Cell Genome Sequencing Data”. In: *Genome Research* 29.11 (Jan. 11, 2019), pp. 1847–1859. DOI: 10.1101/gr.243121.118. pmid: 31628257.

[201] X. Zhang, C. Xu, and N. Yosef. “Simulating Multiple Faceted Variability in Single Cell RNA Sequencing”. In: *Nature Communications* 10.1 (1 June 13, 2019), p. 2611. DOI: 10.1038/s41467-019-10500-w.

[202] E. Y. Zhao, M. Jones, and S. J. M. Jones. “Whole-Genome Sequencing in Cancer”. In: *Cold Spring Harbor Perspectives in Medicine* 9.3 (Mar. 1, 2019), a034579. DOI: 10.1101/cshperspect.a034579. pmid: 29844223.

[203] A. Alblawi et al. “Scaffold-Free: A Developing Technique in Field of Tissue Engineering”. In: *Computer Methods and Programs in Biomedicine* 185 (Mar. 1, 2020), p. 105148. DOI: 10.1016/j.cmpb.2019.105148.

[204] L. B. Alexandrov et al. “The Repertoire of Mutational Signatures in Human Cancer”. In: *Nature* 578.7793 (7793 Feb. 2020), pp. 94–101. DOI: 10.1038/s41586-020-1943-3.

[205] K. G. Andersen et al. “The Proximal Origin of SARS-CoV-2”. In: *Nature Medicine* 26.4 (4 Apr. 2020), pp. 450–452. DOI: 10.1038/s41591-020-0820-9.

[206] A. Bastola et al. “The First 2019 Novel Coronavirus Case in Nepal”. In: *The Lancet Infectious Diseases* 20.3 (Mar. 1, 2020), pp. 279–280. DOI: 10.1016/S1473-3099(20)30067-0.

[207] V. Bergen et al. "Generalizing RNA Velocity to Transient Cell States through Dynamical Modeling". In: *Nature Biotechnology* 38.12 (12 Dec. 2020), pp. 1408–1414. DOI: 10.1038/s41587-020-0591-3.

[208] M. F. Boni et al. "Evolutionary Origins of the SARS-CoV-2 Sarbecovirus Lineage Responsible for the COVID-19 Pandemic". In: *Nature Microbiology* 5.11 (11 Nov. 2020), pp. 1408–1417. DOI: 10.1038/s41564-020-0771-4.

[209] G. Caravagna et al. "Subclonal Reconstruction of Tumors by Using Machine Learning and Population Genetics". In: *Nature Genetics* 52.9 (9 Sept. 2020), pp. 898–907. DOI: 10.1038/s41588-020-0675-5.

[210] L. González-Silva, L. Quevedo, and I. Varela. "Tumor Functional Heterogeneity Unraveled by scRNA-seq Technologies". In: *Trends in Cancer* 6.1 (Jan. 1, 2020), pp. 13–19. DOI: 10.1016/j.trecan.2019.11.010. pmid: 31952776.

[211] M. C. Hansen, T. Haferlach, and C. G. Nyvold. "A Decade with Whole Exome Sequencing in Haematology". In: *British Journal of Haematology* 188.3 (2020), pp. 367–382. DOI: 10.1111/bjh.16249.

[212] M. Hong et al. "RNA Sequencing: New Technologies and Applications in Cancer Research". In: *Journal of Hematology & Oncology* 13.1 (Dec. 4, 2020), p. 166. DOI: 10.1186/s13045-020-01005-x.

[213] Z. Hu et al. "Multi-Cancer Analysis of Clonality and the Timing of Systemic Spread in Paired Primary Tumors and Metastases". In: *Nature Genetics* 52.7 (7 July 2020), pp. 701–708. DOI: 10.1038/s41588-020-0628-z.

[214] A. Jary et al. "Evolution of Viral Quasispecies during SARS-CoV-2 Infection". In: *Clinical Microbiology and Infection* 26.11 (Nov. 1, 2020), 1560.e1–1560.e4. DOI: 10.1016/j.cmi.2020.07.032. pmid: 32717416.

[215] Y. Kashima et al. "Single-Cell Sequencing Techniques from Individual to Multiomics Analyses". In: *Experimental & Molecular Medicine* 52.9 (9 Sept. 2020), pp. 1419–1427. DOI: 10.1038/s12276-020-00499-2.

[216] R. Kumar et al. "Single Cell Metabolomics: A Future Tool to Unmask Cellular Heterogeneity and Virus-Host Interaction in Context of Emerging Viral Diseases". In: *Frontiers in Microbiology* 11 (2020), p. 1152. DOI: 10.3389/fmicb.2020.01152.

[217] D. Lähnemann et al. "Eleven Grand Challenges in Single-Cell Data Science". In: *Genome Biology* 21.1 (Feb. 7, 2020), p. 31. DOI: 10.1186/s13059-020-1926-6.

[218] J. K. Marzinek, R. G. Huber, and P. J. Bond. "Multiscale Modelling and Simulation of Viruses". In: *Current Opinion in Structural Biology*. Theory and Simulation . Macromolecular Assemblies 61 (Apr. 1, 2020), pp. 146–152. DOI: 10.1016/j.sbi.2019.12.019.

[219] N. Milind et al. "Transcriptomic Stratification of Late-Onset Alzheimer's Cases Reveals Novel Genetic Modifiers of Disease Pathology". In: *PLOS Genetics* 16.6 (June 3, 2020), e1008775. DOI: 10.1371/journal.pgen.1008775.

[220] D. P. R&d et al. *COVID-19 ARTIC v3 Illumina Library Construction and Sequencing Protocol*. protocols.io. Nov. 4, 2020.

[221] J. Al-Sabah, C. Baccin, and S. Haas. "Single-Cell and Spatial Transcriptomics Approaches of the Bone Marrow Microenvironment". In: *Current Opinion in Oncology* 32.2 (Mar. 2020), pp. 146–153. DOI: 10.1097/CCO.0000000000000602.

[222] M. N. Shahbazi. "Mechanisms of Human Embryo Development: From Cell Fate to Tissue Shape and Back". In: *Development* 147.14 (July 17, 2020), dev190629. DOI: 10.1242/dev.190629.

[223] B. Shrestha. "Single-Cell Metabolomics by Mass Spectrometry". In: *Single Cell Metabolism: Methods and Protocols*. Ed. by B. Shrestha. Methods in Molecular Biology. New York, NY: Springer, 2020, pp. 1–8. ISBN: 978-1-4939-9831-9. DOI: 10.1007/978-1-4939-9831-9_1.

[224] L. Wang et al. "A Gene Expression-Based Immune Signature for Lung Adenocarcinoma Prognosis". In: *Cancer Immunology, Immunotherapy* 69.9 (Sept. 1, 2020), pp. 1881–1890. DOI: 10.1007/s00262-020-02595-8.

[225] Z. Zhou et al. "DENDRO: Genetic Heterogeneity Profiling and Subclone Detection by Single-Cell RNA Sequencing". In: *Genome Biology* 21.1 (Jan. 14, 2020), p. 10. DOI: 10.1186/s13059-019-1922-x.

[226] G. Aguadé-Gorgorió, S. Kauffman, and R. Solé. "Transition Therapy: Tackling the Ecology of Tumor Phenotypic Plasticity". In: *Bulletin of Mathematical Biology* 84.1 (Dec. 27, 2021), p. 24. DOI: 10.1007/s11538-021-00970-9.

[227] F. Angaroni, M. Antoniotti, and A. Graudenzi. *OG-SPACE: Optimized Stochastic Simulation of Spatial Models of Cancer Evolution*. Oct. 13, 2021. arXiv: 2110.06588 [physics, q-bio].

[228] F. Angaroni et al. "PMCE: Efficient Inference of Expressive Models of Cancer Evolution with High Prognostic Power". In: *Bioinformatics* (Oct. 14, 2021), btab717. DOI: 10.1093/bioinformatics/btab717.

[229] C. Cao et al. "The Architecture of the SARS-CoV-2 RNA Genome inside Virion". In: *Nature Communications* 12.1 (1 June 24, 2021), p. 3917. DOI: 10.1038/s41467-021-22785-x.

[230] M. Chiara et al. "Next Generation Sequencing of SARS-CoV-2 Genomes: Challenges, Applications and Opportunities". In: *Briefings in Bioinformatics* 22.2 (Mar. 1, 2021), pp. 616–630. DOI: 10.1093/bib/bbaa297.

[231] S. Ciccolella et al. "Inferring Cancer Progression from Single-Cell Sequencing While Allowing Mutation Losses". In: *Bioinformatics* 37.3 (Feb. 1, 2021), pp. 326–333. DOI: 10.1093/bioinformatics/btaa722.

[232] S. C. Dentro et al. "Characterizing Genetic Intra-Tumor Heterogeneity across 2,658 Human Cancer Genomes". In: *Cell* 184.8 (Apr. 15, 2021), 2239–2254.e39. DOI: 10.1016/j.cell.2021.03.009.

[233] A. W. DeVilbiss et al. "Metabolomic Profiling of Rare Cell Populations Isolated by Flow Cytometry from Tissues". In: *eLife* 10 (Jan. 20, 2021). Ed. by R. M. White, M. G. Vander Heiden, and T. Papagiannakopoulos, e61980. DOI: 10.7554/eLife.61980.

[234] J. Diaz-Colunga and R. Diaz-Uriarte. "Conditional Prediction of Consecutive Tumor Evolution Using Cancer Progression Models: What Genotype Comes Next?" In: *PLOS Computational Biology* 17.12 (Dec. 21, 2021), e1009055. DOI: 10.1371/journal.pcbi.1009055.

[235] M. N. Gondal and S. U. Chaudhary. "Navigating Multi-Scale Cancer Systems Biology Towards Model-Driven Clinical Oncology and Its Applications in Personalized Therapeutics". In: *Frontiers in Oncology* 11 (2021), p. 4767. DOI: 10.3389/fonc.2021.712505.

[236] Y. Hao et al. "Integrated Analysis of Multimodal Single-Cell Data". In: *Cell* 184.13 (June 24, 2021), 3573–3587.e29. DOI: 10.1016/j.cell.2021.04.048. pmid: 34062119.

[237] L. Huo et al. "Single-Cell Multi-Omics Sequencing: Application Trends, COVID-19, Data Analysis Issues and Prospects". In: *Briefings in Bioinformatics* 22.6 (Nov. 1, 2021), bbab229. DOI: 10.1093/bib/bbab229.

[238] S. Knyazev et al. "Epidemiological Data Analysis of Viral Quasispecies in the Next-Generation Sequencing Era". In: *Briefings in Bioinformatics* 22.1 (Jan. 1, 2021), pp. 96–108. DOI: 10.1093/bib/bbaa101.

[239] M. Kuksin et al. "Applications of Single-Cell and Bulk RNA Sequencing in Onco-Immunology". In: *European Journal of Cancer* 149 (May 1, 2021), pp. 193–210. DOI: 10.1016/j.ejca.2021.03.005.

[240] A. Lal et al. "De Novo Mutational Signature Discovery in Tumor Genomes Using SparseSignatures". In: *PLOS Computational Biology* 17.6 (June 28, 2021), e1009119. DOI: 10.1371/journal.pcbi.1009119.

[241] I. Martínez-Reyes and N. S. Chandel. "Cancer Metabolism: Looking Forward". In: *Nature Reviews Cancer* 21.10 (10 Oct. 2021), pp. 669–680. DOI: 10.1038/s41568-021-00378-6.

[242] A. R. Massarat et al. “Discovering Single Nucleotide Variants and Indels from Bulk and Single-Cell ATAC-seq”. In: *Nucleic Acids Research* 49.14 (Aug. 20, 2021), pp. 7986–7994. DOI: 10.1093/nar/gkab621.

[243] M. L. Matthews and A. Marshall-Colón. “Multiscale Plant Modeling: From Genome to Phenome and Beyond”. In: *Emerging Topics in Life Sciences* 5.2 (Feb. 5, 2021). Ed. by J. M. Jez and C. N. Topp, pp. 231–237. DOI: 10.1042/ETLS20200276.

[244] A. Modi et al. “The Illumina Sequencing Protocol and the NovaSeq 6000 System”. In: *Bacterial Pangenomics: Methods and Protocols*. Ed. by A. Mengoni, G. Bacci, and M. Fondi. Methods in Molecular Biology. New York, NY: Springer US, 2021, pp. 15–42. ISBN: 978-1-07-161099-2. DOI: 10.1007/978-1-0716-1099-2_2.

[245] F. Mölder et al. “Sustainable Data Analysis with Snakemake”. In: 10:33 (Apr. 19, 2021). DOI: 10.12688/f1000research.29032.2.

[246] P. Nieto et al. “A Single-Cell Tumor Immune Atlas for Precision Oncology”. In: *Genome Research* (Sept. 21, 2021). DOI: 10.1101/gr.273300.120. pmid: 34548323.

[247] N. M. Novikov et al. “Mutational Drivers of Cancer Cell Migration and Invasion”. In: *British Journal of Cancer* 124.1 (1 Jan. 2021), pp. 102–114. DOI: 10.1038/s41416-020-01149-0.

[248] Á. O’Toole et al. “Tracking the International Spread of SARS-CoV-2 Lineages B.1.1.7 and B.1.351/501Y-V2 with Grinch”. In: 6:121 (Sept. 17, 2021). DOI: 10.12688/wellcomeopenres.16661.2.

[249] B. B. Oude Munnink et al. “The next Phase of SARS-CoV-2 Surveillance: Real-Time Molecular Epidemiology”. In: *Nature Medicine* 27.9 (9 Sept. 2021), pp. 1518–1524. DOI: 10.1038/s41591-021-01472-w.

[250] L. Patruno et al. “A Review of Computational Strategies for Denoising and Imputation of Single-Cell Transcriptomic Data”. In: *Briefings in Bioinformatics* 22.4 (July 1, 2021), bbaa222. DOI: 10.1093/bib/bbaa222.

[251] S. Salehi et al. “Clonal Fitness Inferred from Time-Series Modelling of Single-Cell Cancer Genomes”. In: *Nature* 595.7868 (7868 July 2021), pp. 585–590. DOI: 10.1038/s41586-021-03648-3.

[252] J. O. Shaibu et al. “Full Length Genomic Sanger Sequencing and Phylogenetic Analysis of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) in Nigeria”. In: *PLOS ONE* 16.1 (Jan. 11, 2021), e0243271. DOI: 10.1371/journal.pone.0243271.

[253] P. Simmonds and M. A. Ansari. "Extensive C->U Transition Biases in the Genomes of a Wide Range of Mammalian RNA Viruses; Potential Associations with Transcriptional Mutations, Damage- or Host-Mediated Editing of Viral RNA". In: *PLOS Pathogens* 17.6 (June 1, 2021), e1009596. DOI: 10.1371/journal.ppat.1009596.

[254] G. Sun et al. "Single-Cell RNA Sequencing in Cancer: Applications, Advances, and Emerging Challenges". In: *Molecular Therapy - Oncolytics* 21 (June 25, 2021), pp. 183–206. DOI: 10.1016/j.omto.2021.04.001. pmid: 34027052.

[255] M. Tedesco et al. "Chromatin Velocity Reveals Epigenetic Dynamics by Single-Cell Profiling of Heterochromatin and Euchromatin". In: *Nature Biotechnology* (Oct. 11, 2021), pp. 1–10. DOI: 10.1038/s41587-021-01031-1.

[256] N. Wang et al. "Spatial Transcriptomics and Proteomics Technologies for Deconvoluting the Tumor Microenvironment". In: *Biotechnology Journal* 16.9 (2021), p. 2100041. DOI: 10.1002/biot.202100041.

[257] M. Zhang et al. "Spatial Molecular Profiling: Platforms, Applications and Analysis Tools". In: *Briefings in Bioinformatics* 22.3 (May 1, 2021), bbaa145. DOI: 10.1093/bib/bbaa145.

[258] S. Zhang et al. "Longitudinal Single-Cell Profiling Reveals Molecular Heterogeneity and Tumor-Immune Evolution in Refractory Mantle Cell Lymphoma". In: *Nature Communications* 12.1 (1 May 17, 2021), p. 2877. DOI: 10.1038/s41467-021-22872-z.

[259] X. Zhu et al. "Cancer Evolution: A Means by Which Tumors Evade Treatment". In: *Biomedicine & Pharmacotherapy* 133 (Jan. 1, 2021), p. 111016. DOI: 10.1016/j.biopha.2020.111016.

[260] N. Andrews et al. "Covid-19 Vaccine Effectiveness against the Omicron (B.1.1.529) Variant". In: *New England Journal of Medicine* 0.0 (Mar. 2, 2022), null. DOI: 10.1056/NEJMoa2119451.

[261] W. H. Organization. *Breast Cancer*.

[262] P. L. Tzou et al. "Comparison of an In Vitro Diagnostic Next-Generation Sequencing Assay with Sanger Sequencing for HIV-1 Genotypic Resistance Testing". In: *Journal of Clinical Microbiology* 56.6 (), e00105–18. DOI: 10.1128/JCM.00105-18.