




Variant calling from scRNA-seq data allows the assessment of cellular identity in patient-derived cell lines

Daniele Ramazzotti¹, Fabrizio Angaroni², Davide Maspero^{2,3}, Gianluca Ascolani², Isabella Castiglioni⁴, Rocco Piazza ^{1,5}, Marco Antoniotti ^{2,5} & Alex Graudenzi ^{2,3,5✉}

ARISING FROM Sharma et al. *Nature Communications* <https://doi.org/10.1038/s41467-018-07261-3> (2018)

Single-cell sequencing experiments enable the investigation of cell-to-cell heterogeneity at unprecedented resolution¹, and this is especially relevant in the study of cancer evolution². In ref. ³, the authors employed longitudinal single-cell transcriptomic data from patient-derived primary and metastatic Oral Squamous Cell Carcinomas (OSCC) cell lines (from a previous panel⁴), to investigate possible divergent modes of chemoresistance in tumor subpopulations. We integrated the analyses by performing variant calling from single-cell RNA sequencing (scRNA-seq) data via GATK Best Practices⁵, and discovered a high number of Single-Nucleotide Variants (SNVs) representative of the identity of a specific patient in the cell line derived from a second patient, and vice versa. These findings suggest the existence of a sample swap, thus jeopardizing some key translational conclusions of the article, and prove the efficacy of a joint analysis of the genotypic and transcriptomic identity of single cells.

Even though scRNA-seq data are typically employed to characterize single-cell gene expression profiles⁶, recent studies proved that data generated with full-length protocols (e.g., Smart-Seq/Smart-Seq2⁷) can be effectively used for variant calling⁸. Despite known pitfalls, such as the impossibility of calling genomic variants from non-transcribed regions and the high rates of noise and dropouts⁹, this provides a highly-available and cost-effective alternative to DNA sequencing¹⁰. The mutational profiles so obtained can be used to determine the identity of single cells, and this is useful to characterize the clonal evolution of tumors¹¹ and assess the impact of therapies, when longitudinal experiments are available¹². Furthermore, this allows a natural mapping between the genotype and the gene expression profile of single cells¹³. This aspect has significant translational relevance, given the shortage of accurate and affordable technologies for concurrent DNA and RNA sequencing of the same cells, despite the introduction of new protocols^{14,15}.

We integrated the analyses presented in³ and 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 suffixes: -P (primary line), -M (metastatic line), -CR (after cisplatin treatment), -CRDH (after drug-holiday). Since, for the HN137P cell line, single- and paired-end library layouts are provided, and HN137MCRDH is not present, we have a total of 12 datasets (GEO accession code GSE117872; refer to³ for details on the experimental setup). In detail, we selected single cells labeled as “good data” and performed variant calling with the procedure employed in¹² and described in the Supplementary Information (SI).

4,924,559 unique variants were detected on a total of 1,116 single cells included in all datasets. Quality control filters were applied to ensure high confidence to the calls and reduce the number of false alleles and miscalls. In particular, we removed: (i) indels and other structural variants—to limit the impact of sequencing and alignment artifacts, (ii) variants mapped on mitochondrial genes, (iii) variants on positions with coverage < 5 reads 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 reads supporting the alternative allele) in both HN120P and HN137P (*single-end*)—to define a list of variants clearly characterizing the identity of the two primary cell lines. We finally selected the variants observed in at least 1 cell (≥ 3 reads supporting the alternative allele, ≥ 5 coverage) of HN120P and in exactly 0 cells of HN137P (*single-end*), and the variants observed in at least 1 cell (≥ 3 reads supporting the alternative allele, ≥ 5 coverage) of a HN137P (*single-end*) and in exactly 0 cells of HN120P.

As a result, we identified 67 SNVs representative of HN120P cell identity. Such variants are observed at high frequency in HN120P

¹Dept. of Medicine and Surgery, Univ. of Milan-Bicocca, Monza, Italy. ²Dept. of Informatics, Systems and Communication, Univ. of Milan-Bicocca, Milan, Italy. ³Inst. of Molecular Bioimaging and Physiology, National Research Council (IBFM-CNR), Segrate, Milan, Italy. ⁴Dept. of Physics “Giuseppe Occhialini”, Univ. of Milan-Bicocca, Milan, Italy. ⁵Bicocca Bioinformatics, Biostatistics and Bioimaging Centre - B4, Milan, Italy. ✉email: alex.graudenzi@unimib.it

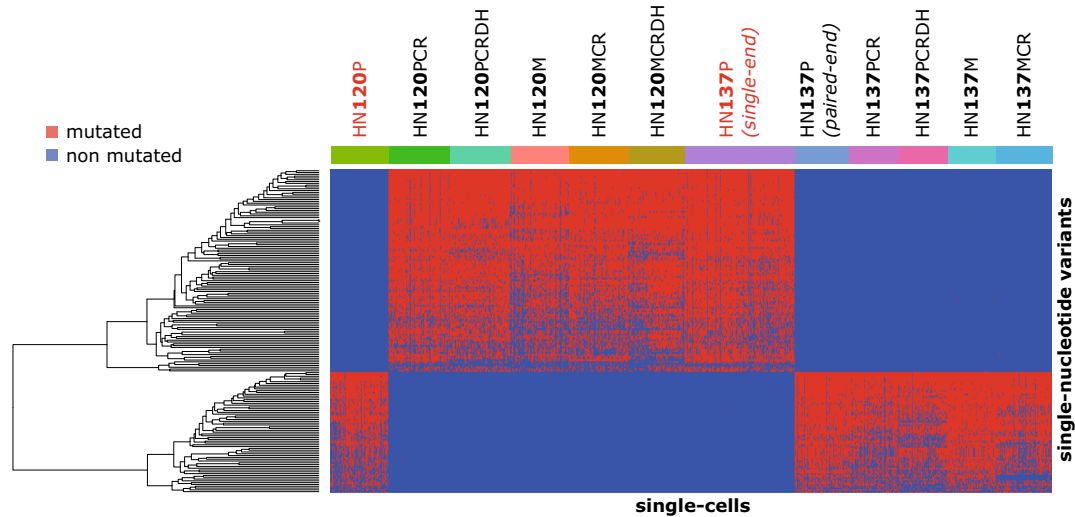
and in HN137P (*paired-end*), HN137PCR, HN137PCRDH, HN137M, HN137MCR, whereas are not observed (<1% of the cells) in HN120PCR, HN120PCRDH, HN120M, HN120MCR, HN120MCRDH and HN137P (*single-end*). In Fig. 1A, we display the mutational profiles of all single cells in all datasets (coverage information is provided in Supplementary Data 1).

Analogously, we identified 112 SNVs that are strongly informative for HN137P (*single-end*) identity (see Fig. 1A). Such variants are observed at 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,

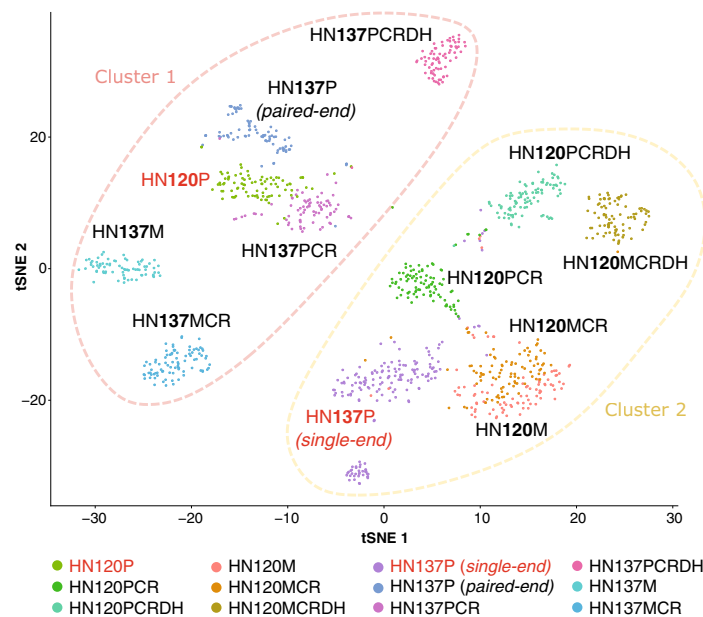
HN137MCR, and HN120P. The attributes of the SNVs are reported in Supplementary Data 2.

From the 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, while being 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 over 179 variants have a rsID). These

A Single-cell genotype analysis



B Single-cell transcriptomic analysis



C VIM expression analysis

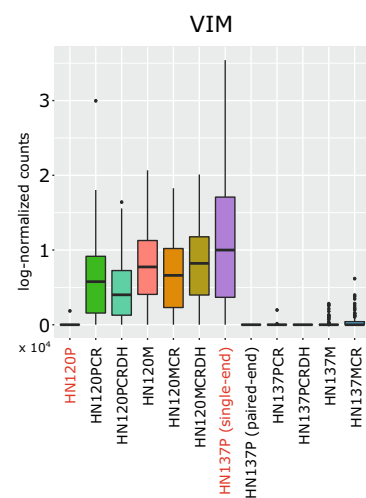


Fig. 1 Analysis of single-cell mutational and gene expression profiles of patient-derived OSCC cell lines from scRNA-seq data. A The heatmap including the mutational profiles of all single cells of the HN120 and HN137 datasets is displayed (-P: primary line, -M: metastatic line, -CR: after cisplatin treatment, -CRDH: after drug-holiday). Red entries mark cells displaying a given SNV. For the ID of single cells and SNVs please refer to Supplementary Data 1 and 2. **B** The t-SNE plot generated from the gene expression profiles of all single cells for all datasets is shown (see the SI for additional details). **C** The distribution of the expression level of *VIM* on all single cells is shown with boxplots for all datasets.

surprising results can be hardly explained by cancer-related selection phenomena, random effects, or sampling limitations. Instead, this suggests the likely presence of a methodological issue involving a label swap of samples HN120P and HN137P (*single-end*).

This hypothesis is further supported by the single-cell transcriptomic analysis performed via Seurat¹⁶ (see the SI). In Fig. 1B, one can find the t-SNE plot computed on the 1000 most variable genes. Consistently with the genotype analysis, the transcriptomic analysis 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 may have led to erroneous conclusions in refs. 3,17,18. In³, 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* (Fig. 1C), one can notice that this gene is up-regulated 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, in³ 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 gets extinct. 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. 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 analyses presented above.

Overall, our results prove that scRNA-seq data can be effectively exploited to perform an integrated analysis of the genotypic and transcriptomic identity of single cells, providing a powerful tool to decipher complex phenomena such as cancer evolution and drug resistance.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

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.

Received: 18 February 2020; Accepted: 6 April 2022;

Published online: 12 May 2022

References

- Lähnemann, D. et al. Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 1–35 (2020).
- Lawson, D. A., Kessenbrock, K., Davis, R. T., Pervolarakis, N. & Werb, Z. Tumour heterogeneity and metastasis at single-cell resolution. *Nat. Cell Biol.* **20**, 1349–1360 (2018).
- Sharma, A. 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).
- Chia, S. et al. Phenotype-driven precision oncology as a guide for clinical decisions one patient at a time. *Nat. Commun.* **8**, 1–12 (2017).
- DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491 (2011).
- Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15** (2019).
- Picelli, S. et al. Full-length RNA-seq from single cells using smart-seq2. *Nat. Protocols* **9**, 171–181 (2014).
- Liu, F. et al. Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data. *Genome Biol.* **20**, 1–15 (2019).
- Patrino, L. et al. A review of computational strategies for denoising and imputation of single-cell transcriptomic data. *Brief. Bioinform.* **22**, bbaa222 (2020).
- Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
- Ramazzotti, D., Graudenzi, A., De Sano, L., Antoniotti, M. & Caravagna, G. Learning mutational graphs of individual tumour evolution from single-cell and multi-region sequencing data. *BMC Bioinf.* **20**, 210 (2019).
- Ramazzotti, D. et al. LACE: Inference of cancer evolution models from longitudinal single-cell sequencing data. *J. Comput. Sci.* **58**, 101523 (2022).
- Zhou, Z., Xu, B., Minn, A. & Zhang, N. R. Dendro: genetic heterogeneity profiling and subclone detection by single-cell rna sequencing. *Genome Biol.* **21**, 1–15 (2020).
- Macaulay, I. C. et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519 (2015).
- Nam, A. S. et al. Somatic mutations and cell identity linked by genotyping of transcriptomes. *Nature* **571**, 355–360 (2019).
- Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
- Sharma, A. & DasGupta, R. Tracking tumor evolution one-cell-at-a-time. *Mol. Cell. Oncol.* **6**, 1590089 (2019).
- Sharma, A. Hiding in plain sight: epigenetic plasticity in drug-induced tumor evolution. *Epigenet. Insights* **12**, 2516865719870760 (2019).

Acknowledgements

This work was partially supported by the CRUK/AIRC Accelerator Award #22790 “Single-cell Cancer Evolution in the Clinic”, by the Elixir Italian Chapter and the SysBioNet project—a Italian Ministry of University and Research (MIUR) initiative for the Italian Roadmap of European Strategy Forum on Research Infrastructures—by the AIRC-IG grant 22082 to RP, by a Bicocca 2020 Starting Grant to FA and DR, and by the MIUR—Department of Excellence project PREMIA. We thank Giulio Caravagna, Chiara Damiani, Francesco Craighero and Lucrezia Patrino for helpful discussions.

Author contributions

All authors performed the analyses, interpreted the results, drafted and approved the manuscript. A.G. and D.R. supervised the study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-30230-w>.

Correspondence and requests for materials should be addressed to Alex Graudenzi.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022