



Parsimonious Clone Tree Reconciliation in Cancer

Palash Sashittal ✉ 🏠 

Department of Computer Science, University of Illinois at Urbana-Champaign,
Urbana, IL, USA

Simone Zaccaria ✉ 🏠 

Computational Cancer Genomics Research Group,
University College London Cancer Institute, London, UK
Cancer Research UK Lung Cancer Centre of Excellence,
University College London Cancer Institute, London, UK

Mohammed El-Kebir¹ ✉ 🏠 

Department of Computer Science, University of Illinois at Urbana-Champaign,
Urbana, IL, USA
Cancer Center at Illinois, University of Illinois at Urbana-Champaign,
Urbana, IL, USA

Abstract

Every tumor is composed of heterogeneous clones, each corresponding to a distinct subpopulation of cells that accumulated different types of somatic mutations, ranging from single-nucleotide variants (SNVs) to copy-number aberrations (CNAs). As the analysis of this intra-tumor heterogeneity has important clinical applications, several computational methods have been introduced to identify clones from DNA sequencing data. However, due to technological and methodological limitations, current analyses are restricted to identifying tumor clones only based on either SNVs or CNAs, preventing a comprehensive characterization of a tumor's clonal composition. To overcome these challenges, we formulate the identification of clones in terms of both SNVs and CNAs as a reconciliation problem while accounting for uncertainty in the input SNV and CNA proportions. We thus characterize the computational complexity of this problem and we introduce a mixed integer linear programming formulation to solve it exactly. On simulated data, we show that tumor clones can be identified reliably, especially when further taking into account the ancestral relationships that can be inferred from the input SNVs and CNAs. On 49 tumor samples from 10 prostate cancer patients, our reconciliation approach provides a higher resolution view of tumor evolution than previous studies.

2012 ACM Subject Classification Applied computing → Computational genomics

Keywords and phrases Intra-tumor heterogeneity, phylogenetics, mixed integer linear programming

Digital Object Identifier 10.4230/LIPIcs.WABI.2021.9

Supplementary Material *Software (Source Code)*: <https://github.com/elkebir-group/paction>
archived at `swh:1:dir:4a932ffa97a0c3be4118281e67ac2d86459de00f`

Funding *Simone Zaccaria*: Rosetrees Trust and CRUK Lung Cancer Centre of Excellence grant reference M917.

Mohammed El-Kebir: National Science Foundation award numbers CCF 1850502 and CCF 2046488.

Acknowledgements This work was a project in the course CS598MEB (Computational Cancer Genomics, Spring 2021) at UIUC. We thank the students in this course for their valuable feedback. We also thank Ron Zeira for providing the code to compute distances between copy number profiles.

¹ Corresponding author

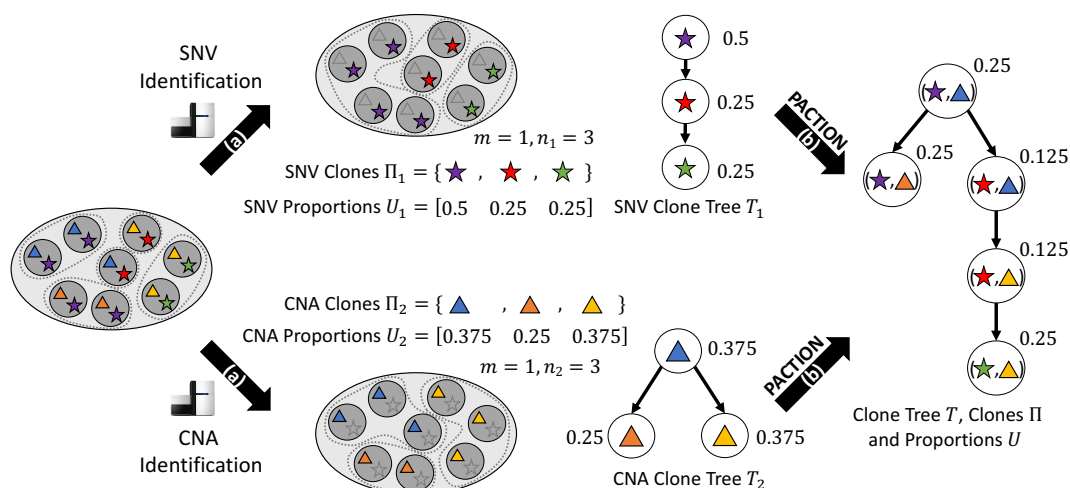


1 Introduction

Cancer results from an evolutionary process where somatic mutations accumulate in the genomes of different cells. This process yields highly heterogeneous tumors composed of different *clones*, each corresponding to a distinct subpopulation of cells with the same complement of somatic mutations [27]. The resulting intra-tumor heterogeneity has been clearly linked to critically important cancer phenotypes, including cancer prognosis and the potential of developing resistance to cancer therapy [3, 24]. Therefore, important downstream applications rely on accurate reconstructions of a tumor’s clonal architecture, which in turn requires the identification of the different clones, their proportions and their evolutionary history. However, the presence of different types of somatic mutations in the same clones renders these tasks particularly challenging. In particular, the following two types of somatic mutations are frequent in cancer [4, 39, 40]: (1) single nucleotide variants (SNVs), which are substitutions of individual DNA nucleotides, and (2) copy number alterations (CNAs), which are amplifications and deletions of large genomic regions.

Most cancer sequencing studies use bulk DNA sequencing technology, where one does not directly measure the co-occurrence of different mutations in the same clone because the generated DNA sequencing reads originate from unknown mixtures of millions of different cells in a bulk tumor sample. To identify distinct clones from such data, one thus needs to deconvolve the mixed sequencing data into the different clonal components [37]. Several computational methods have been introduced to perform this task. However, the majority of existing methods only focus on either SNVs [6, 29, 31, 35, 36] or CNAs [11, 25, 26, 28, 42–44], but rarely on both. Methods that attempt to identify clones in terms of both SNVs and CNAs do not scale to the numbers of current cancer sequencing datasets (e.g., number of samples, mutations, clones, etc.) and often require heuristics to reduce the size of input instances [5, 9, 19]. As a result, current cancer evolutionary analyses [16, 18] do not apply such proposed methods but rather perform a *post hoc* analysis, manually assigning CNAs to a tree inferred from SNVs. Furthermore, we note that similar issues arise with some single-cell DNA sequencing technologies, since the different features of these technologies only allow the reliable measurement of either SNVs or CNAs [14]. For example, targeted MDA single-cell sequencing technologies are more suited for the identification of SNVs whereas whole-exome/genome DOP-PCR single-cell technologies are more suited for the identification of CNAs, and both these technologies have been used in parallel on the same tumor sample [22].

In this study, we investigate whether tumor clonal compositions can be comprehensively reconstructed by an alternative simpler and automated approach. Leveraging the SNV and CNA clone proportions that can be independently and reliably inferred by existing methods, we introduce the PARSIMONIOUS CLONE RECONCILIATION (PCR) and PARSIMONIOUS CLONE TREE RECONCILIATION (PCTR) problems to infer clones in terms of both SNVs and CNAs, their proportions and, additionally for the PCTR problem, their evolutionary relationships (Figure 1). We prove that the proposed problems are NP-hard and we introduce PACTION (PARsimonious Clone Tree reconciliatION), an algorithm that solves these problems using two mixed integer linear programming formulations. Using simulations, we find that our approach reliably handles errors in input SNV and CNA proportions and scales to practical instance sizes. On 49 samples from prostate cancer patients [16], we find that our approach more comprehensively reconstructs tumor clonal architectures compared to the manual approach adopted in the previous analysis of the same data.



■ **Figure 1 Overview.** A tumor is composed of multiple subpopulations of cells, or clones, with distinct somatic mutations, which can be measured using DNA sequencing. (a) Due to limitations in inference algorithms and/or sequencing technologies, we are limited to characterizing tumor clones in terms of either single-nucleotide variants (SNVs, stars) or copy-number aberrations (CNAs, triangles). That is, we infer clones Π_1 , proportions U_1 and a clone tree T_1 for the SNVs. Similarly, we infer clones Π_2 , proportions U_2 and a clone tree T_2 for the CNAs. (b) PACTION solves the PARSIMONIOUS CLONE TREE RECONCILIATION problem of inferring clones $\Pi \subseteq \Pi_1 \times \Pi_2$, a clone tree T and proportions U that characterize the clones of the tumor in terms of both SNVs and CNAs.

2 Problem Statements

We introduce two reconciliation problem formulations to reconstruct tumor clonal composition from inferred SNV and CNA clone proportions². The first problem aims at inferring tumor clones and related proportions with both SNVs and CNAs given the clone proportions of SNVs and CNAs independently (Section 2.1). The second problem additionally considers phylogenetic trees describing the evolution of tumor clones with either different SNVs or CNAs (Section 2.2).

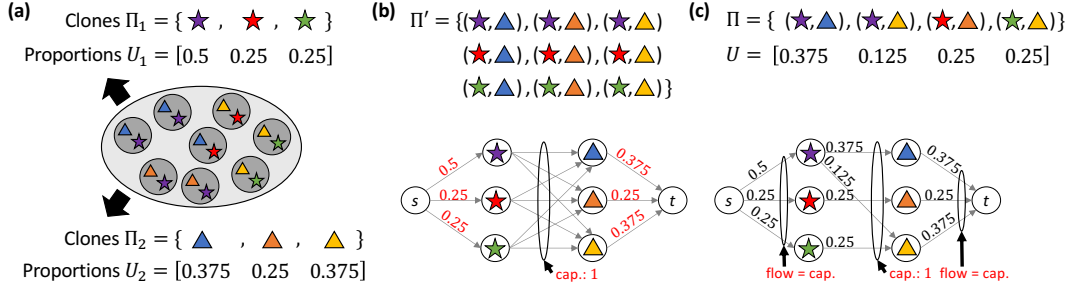
2.1 Parsimonious Clone Reconciliation

Suppose a tumor is composed of a set Π of $n = |\Pi|$ clones, which are characterised by unique complements of two different features (e.g., SNVs and CNAs). These clones occur in m samples at varying proportions, defined as follows.

► **Definition 1.** An $m \times n$ matrix $U = [u_{p,\ell}]$ is a proportion matrix for n clones Π provided (i) $u_{p,\ell} \geq 0$ for all samples $p \in [m]$ and clones $\ell \in [n]$, and (ii) $\sum_{\ell=1}^n u_{p,\ell} = 1$ for all samples $p \in [m]$.

Due to limitations in inference algorithms and/or sequencing technologies, we only infer clones and their proportions for one feature in isolation. These two features lead to two distinct partitions of all tumor cells: a set $\Pi_1 = [n_1]$ of clones induced by the first feature (e.g.,

² While reconciliation is used in species phylogenetics, particularly in the context of gene-tree species-tree reconciliation, here we will use this term to indicate the process of obtaining a comprehensive evolutionary tree of tumor clones given input trees that each focus on a distinct genomic feature.



■ **Figure 2 The Parsimonious Clone Reconciliation (PCR) problem.** (a) Given clones Π_1 and Π_2 and corresponding proportions U_1 and U_2 , we seek clones $\Pi \subseteq \Pi_1 \times \Pi_2$ and corresponding proportions U consistent with U_1 and U_2 . (b) There always exists a consistent proportion matrix U' for the trivial solution $\Pi' = \Pi_1 \times \Pi_2$, which can be identified by solving a maximum flow problem. (c) We seek the solution Π with minimum number $|\Pi|$ of clones. Here, $|\Pi| = 4$, which is smaller than ground truth (see panel (a)). The corresponding matrix U follows from solving the illustrated maximum flow problem. However, incorporating tree constraints, as in the PCTR problem, will lead to ground truth (Figure 1).

SNVs) and a set $\Pi_2 = [n_2]$ of clones induced by the second feature (e.g., CNAs). We refer to the original clones as Π -clones and the clones induced by the first and the second features as Π_1 -clones and Π_2 -clones, respectively. The proportions of the Π_1 -clones and Π_2 -clones are given by the $m \times n_1$ proportion matrix $U_1 = [u_{p,i}^{(1)}]$ and the $m \times n_2$ proportions matrix $U_2 = [u_{p,j}^{(2)}]$, respectively. How are the proportions U_1 for Π_1 -clones and the proportions U_2 for Π_2 -clones related to the proportions U of the Π -clones?

To answer this question, recall that Π is a partition of all tumor cells induced by the combination of both the two features, whereas Π_1 and Π_2 are partitions induced by each feature in isolation (Figure 2a). As such, we have that the partition Π is a refinement of partitions Π_1 and Π_2 . Thus, each Π -clone ℓ corresponds to a unique Π_1 -clone i and a unique Π_2 -clone j . In other words, we may view the set Π as a binary relation of sets Π_1 and Π_2 of clones composed of pairs $\ell = (i, j)$ of clones, i.e., $\Pi \subseteq \Pi_1 \times \Pi_2$. This relation is captured by the projection functions $\pi_1 : \Pi \rightarrow \Pi_1$ and $\pi_2 : \Pi \rightarrow \Pi_2$ such that $\pi_1((i, j)) = i$ and $\pi_2((i, j)) = j$ for all $(i, j) \in \Pi$. We relate the proportion matrix U for clones Π to the proportion matrix U_1 for clones Π_1 and the proportion matrix U_2 for clones Π_2 as follows.

► **Definition 2.** *Given projection functions $\pi_1 : \Pi \rightarrow \Pi_1$ and $\pi_2 : \Pi \rightarrow \Pi_2$ induced by the set $\Pi \subseteq \Pi_1 \times \Pi_2$ of clones, the proportion matrix $U = [u_{p,\ell}]$ for clones Π is consistent with a proportion matrix $U_1 = [u_{p,i}^{(1)}]$ for clones $\Pi_1 = [n_1]$ and proportion matrix $U_2 = [u_{p,j}^{(2)}]$ for clones $\Pi_2 = [n_2]$ provided (i) $u_{p,i}^{(1)} = \sum_{\ell:\pi_1(\ell)=i} u_{p,\ell}$ for all samples $p \in [m]$ and clones $i \in [n_1]$, and (ii) $u_{p,j}^{(2)} = \sum_{\ell:\pi_2(\ell)=j} u_{p,\ell}$ for all samples $p \in [m]$ and clones $j \in [n_2]$.*

The above definition formalizes the intuition that clones Π of the tumor are a refinement of the input clones Π_1 and Π_2 , and therefore their proportions U must be consistent with the input proportions U_1 and U_2 . Our goal is to recover the set $\Pi \subseteq \Pi_1 \times \Pi_2$ of clones and their proportions U from the proportion matrices U_1 and U_2 for clones Π_1 and Π_2 , respectively. While there always exist trivial solutions given by the full set $\Pi' = \Pi_1 \times \Pi_2$ of $n = n_1 \cdot n_2$ clones (Figure 2b), we seek a solution Π with the smallest number n of clones under the principle of parsimony (Figure 2c).

► **Problem 3 (Parsimonious Clone Reconciliation (PCR)).** *Given proportions U_1 for clones $\Pi_1 = [n_1]$ and proportions U_2 for clones $\Pi_2 = [n_2]$, find (i) the smallest set $\Pi \subseteq \Pi_1 \times \Pi_2$ of clones and (ii) proportions U for Π such that U is consistent with U_1 and U_2 .*

2.2 Parsimonious Clone Tree Reconciliation

In practice, proportions U_1 and U_2 are not measured exactly but are affected by potential measurement errors. As such, accurate recovery of the original clones Π and their proportions U requires correcting U_1 and U_2 . To accomplish this, we require additional information and constraints. In this work, we propose to use the evolutionary relationships among the clones Π_1 and Π_2 that can be inferred by existing methods in the form of clone trees [6–8, 23, 29, 33]. Specifically, a rooted tree T is a *clone tree* for clones Π provided the vertex set $V(T)$ equals Π . Moreover, the root vertex $r(T)$ of a clone tree T corresponds to the normal clone while each edge $(u, v) \in E(T)$ represents a mutation event that altered one of the features of clone u and led to the formation of the clone v .

Similarly to the PCR problem, we are given two clone trees, one for each feature in isolation. In the specific example of two features (e.g., SNVs and CNAs), let clone tree T_1 describe the evolution of clones Π_1 (e.g., SNVs) and clone tree T_2 describe the evolution of clones Π_2 (e.g., CNAs). These trees are inferred using standard algorithms in the field [6, 11, 25, 26, 28, 29, 31, 35, 36, 42–44]. Since all clones share a common evolutionary history the original clone tree T is a *refinement* [31, 41] of the clone trees T_1 and T_2 , which is defined as follows.

► **Definition 4.** *Clone tree T for clones Π is a refinement of clone trees T_1 for clones Π_1 and clone tree T_2 for clones Π_2 provided*

- (i) *for each edge $(i, i') \in E(T_1)$ there exists exactly one $j \in \Pi_2$ such that $((i, j), (i', j)) \in E(T)$,*
- (ii) *for each edge $(j, j') \in E(T_2)$ there exists exactly one $i \in \Pi_1$ such that $((i, j), (i, j')) \in E(T)$,*
- (iii) *for each $((i, j), (i', j')) \in E(T)$, it holds that $(i, i') \in E(T_1)$ and $j = j'$, or $(j, j') \in E(T_2)$ and $i = i'$.*

Intuitively, the above definition states that when collapsing vertices of T corresponding to identical Π_1 -clones one obtains T_1 , and, similarly, T_2 is obtained by collapsing vertices of T corresponding to identical Π_2 -clones.

Under a principle of parsimony and given clone trees T_1, T_2 with related proportions U_1, U_2 , our goal is to find a set $\Pi \subseteq \Pi_1 \times \Pi_2$ of clones, a clone proportion matrix U , and a T_1, T_2 -refined clone tree T that require the smallest correction in U_1 and U_2 . This motivates the following problem statement.

► **Problem 5 (Parsimonious Clone Tree Reconciliation (PCTR)).** *Given proportions U_1 and tree T_1 for clones $\Pi_1 = [n_1]$ and proportions U_2 and tree T_2 for clones $\Pi_2 = [n_2]$, find (i) the set Π of clones, (ii) clone tree T and (iii) proportions U for Π such that the clone tree T is a refinement of T_1 and T_2 and minimizes the total error $J(U, U_1, U_2)$ such that*

$$J(U, U_1, U_2) = \sum_{p=1}^m \sum_{i=1}^{n_1} |u_{p,i}^{(1)} - \sum_{\ell: \pi_1(\ell)=i} u_{p,\ell}| + \sum_{p=1}^m \sum_{j=1}^{n_2} |u_{p,j}^{(2)} - \sum_{\ell: \pi_2(\ell)=j} u_{p,\ell}|.$$

Note that $J(U, U_1, U_2) = 0$ if and only if U is consistent with U_1 and U_2 . The clone trees T , T_1 and T_2 do not appear in the objective function $J(U, U_1, U_2)$ and only provides constraints to the optimization problem. Due to these constraints, unlike the previous PCR problem, PCTR does not always admit a trivial solution with $J(U, U_1, U_2) = 0$ (as we further discuss in Section 3.2).

3 Combinatorial Characterization and Computational Complexity

We investigate the combinatorial structure and computational complexity of the two proposed PCR and PCTR problems in the following two sections, respectively.

3.1 Parsimonious Clone Reconciliation

We characterize the combinatorial structure of feasible and optimal solutions (Π, U) for the PCR problem. We first observe that the PCR problem always has a trivial solution. Specifically, given a set Π_1 of $n_1 = |\Pi_1|$ clones and a set Π_2 of $n_2 = |\Pi_2|$ clones and corresponding proportions $U_1 \in [0, 1]^{m \times n_1}$ and $U_2 \in [0, 1]^{m \times n_2}$, a trivial feasible solution is composed of $n = n_1 n_2$ clones $\Pi = \Pi_1 \times \Pi_2$, which may have many possible corresponding proportions U (Figure 2b). For example, proportions $U = [u_{p,(i,j)}]$ can be computed greedily by considering the n clones in any arbitrary order, and assigning each clone $(i, j) \in \Pi$ a proportion of $u_{p,(i,j)} = \min(u_{p,i}^{(1)}, u_{p,j}^{(2)})$ followed by subsequently updating $u_{p,i}^{(1)} := u_{p,i}^{(1)} - u_{p,(i,j)}$ and $u_{p,j}^{(2)} := u_{p,j}^{(2)} - u_{p,(i,j)}$ for each sample $p \in [m]$. Thus, $n = n_1 n_2$ is an upper bound on the number of clones needed. Can we similarly identify a lower bound on n ?

To answer this question, let the *support* $S(U)$ of an $m \times n$ proportion matrix U be defined as the number of non-zero entries in the vector $U\mathbf{1}_m$ where $\mathbf{1}_m$ is a $m \times 1$ vector with all entries equal to one. That is, the support $S(U)$ of a proportion matrix U of clones Π signifies the number of clones with non-zero proportion in at least one of the samples $p \in [m]$. Any such clone must be part of at least one clone $\ell \in \Pi$ in the solution to the PCR problem to ensure consistency of the proportion matrices. This leads to the following observation.

► **Observation 6.** *Given an instance (Π_1, U_1, Π_2, U_2) of the PCR problem with solution Π we have $n \geq \max(S(U_1), S(U_2))$ where $n = |\Pi|$.*

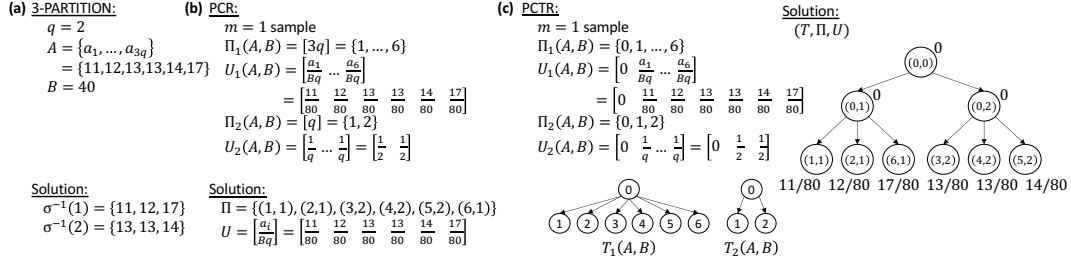
Given any set $\Pi \subseteq \Pi_1 \times \Pi_2$ of clones, deciding whether there exists a proportion matrix U that is consistent with given proportion matrix U_1 for clones Π_1 and U_2 for clones Π_2 , and constructing such a matrix is equivalent to solving a maximum flow problem, which takes polynomial time [1]. Figure 2 illustrates the construction such that there exists a consistent proportion matrix if and only if the value of the flow is 1. Note that for $m > 1$ samples, we need to solve a multi-commodity rather than a single-commodity flow problem. However, the PCR problem, where we simultaneously seek Π and U , is NP-hard and the hardness comes from having to identify the smallest set Π of clones.

► **Theorem 7.** *The PCR problem is NP-hard even for number $m = 1$ of samples.*

This follows by reduction from the 3-PARTITION problem, a known NP-complete problem [12, 13] stated as follows.

► **Problem 8 (3-PARTITION).** *Given an integer $B \in \mathbb{N}^{>0}$, a multiset $A = \{a_1, \dots, a_{3q}\}$ of $3q$ positive integers such that $a_i \in (B/4, B/2)$ for all $i \in [3q]$, and $\sum_{i=1}^{3q} a_i = Bq$, does there exist a partition of A into q disjoint subsets such that the sum of the integers in each subset equals B ?*

Note that since each a_i occurs within the open interval $(B/4, B/2)$ and the elements in each subset of the desired partition sum to B , it holds that each subset must be composed of exactly three elements from the multiset A – hence the name of the problem.



■ **Figure 3 Reduction from 3-PARTITION.** (a) Example instance of 3-PARTITION with a multiset A of 6 elements and target sum $B = 40$. (b) Corresponding PCR instance (Π_1, U_1, Π_2, U_2) and solution (Π, U) . (c) Corresponding PCTR instance $(T_1, \Pi_1, U_1, T_2, \Pi_2, U_2)$ and solution (T, Π, U) .

We represent the solution to an instance (A, B) of the 3-PARTITION problem as a function $\sigma : [3q] \rightarrow [q]$, which encodes the division of the elements of $A = \{a_1, \dots, a_{3q}\}$ into q disjoint subsets. The inverse of this function specifies the subset corresponding to each $j \in [q]$ as $\sigma^{-1}(j) = \{i \in [3q] : \sigma(i) = j\}$. Note that any solution $\sigma : [3q] \rightarrow [q]$ of the 3-PARTITION problem satisfies the following constraint.

$$\sum_{i \in \sigma^{-1}(j)} a_i = B, \quad \forall j \in [q]. \quad (1)$$

Figure 3a provides an example 3-PARTITION instance and solution.

Given a 3-PARTITION problem instance (A, B) , we construct an instance of the PCR problem with number $m = 1$ of samples as follows. The set $\Pi_1(A, B)$ of clones is given by the set $[3q]$. The corresponding proportions are given by the $1 \times 3q$ proportion matrix $U_1(A, B) = [u_{1,i}^{(1)}]$ where $u_{1,i}^{(1)} = a_i/Bq$ for all $i \in [3q]$. Clearly, $U_1(A, B) = [u_{1,i}^{(1)}]$ is a proportion matrix for $\Pi_1(A, B)$ as, by construction, we have that $\sum_{i=1}^{3q} u_{1,i}^{(1)} = 1$ and $u_{1,i}^{(1)} \geq 0$ for all $i \in [3q]$. The second set $\Pi_2(A, B)$ of clones is given by $[q]$. The corresponding proportions are given by the $1 \times q$ proportion matrix $U_2(A, B) = [u_{1,j}^{(2)}]$ where $u_{1,j}^{(2)} = 1/q$ for all $j \in [q]$. It is easy to verify that $U_2(A, B)$ is a proportion matrix for $\Pi_2(A, B)$. Clearly, this construction takes polynomial time. Figure 3b shows an example. Hardness follows from the following lemma whose proof is omitted due to space constraints.

► **Lemma 9.** *Given proportions $U_1(A, B)$ for clones $\Pi_1(A, B) = [3q]$ and proportions $U_2(A, B)$ for clones $\Pi_2(A, B) = [q]$, there exists a set Π of clones of size $n = |\Pi| \leq 3q$ with proportions U that are consistent with $U_1(A, B)$ and $U_2(A, B)$ if and only if there exists a solution to the 3-PARTITION instance (A, B) .*

3.2 Parsimonious Clone Tree Reconciliation

We now characterize the combinatorial structure of feasible and optimal solutions (Π, U, T) for the PCTR problem. Let T_1 be the first input clone tree for the input set Π_1 of $n_1 = |\Pi_1|$ clones. Similarly, let T_2 be the second input clone tree for the input set Π_2 of $n_2 = |\Pi_2|$ clones. Let T be a solution clone tree that is a refinement of both T_1 and T_2 . First, we observe that the clones that label the root vertices $r(T_1)$ and $r(T_2)$ of the two input trees together label the root vertex $r(T)$ of the output tree T , i.e., $r(T) = (r(T_1), r(T_2))$.

► **Observation 10.** *If clones Π , clone tree T and proportion matrix U form a solution to the PCTR instance $(\Pi_1, T_1, U_1, \Pi_2, T_2, U_2)$, then $(r(T_1), r(T_2)) \in \Pi$ and $r(T) = (r(T_1), r(T_2))$.*

Next, from Definition 4 it follows that in the output clone tree T it must hold that along each edge there is either a change in corresponding Π_1 -clones or Π_2 -clones but not both.

► **Observation 11.** *For each $(i, j) \in V(T) \setminus \{r(T)\}$ it holds that either $((i', j), (i, j)) \in E(T)$ or $((i, j'), (i, j)) \in E(T)$ where $(i', i) \in E(T_1)$ and $(j', j) \in E(T_2)$.*

Combining these observations, we get that the number of vertices/clones in T equals $n = n_1 + n_2 - 1$.

► **Observation 12.** *The number of clones $V(T)$ equals $n = n_1 + n_2 - 1$.*

We note that T is a multi-state perfect phylogeny with two characters, i.e. each character state labels at most one edge of T , whose two sets of states correspond to Π_1 and Π_2 . Moreover, T_1 and T_2 impose an ordering of two sets of states to which T must adhere – i.e., the two characters are cladistic [10]. The problem of deciding whether there exists an error-free solution of PCTR with $J(U, U_1, U_2) = 0$ is equivalent to a special case of the CLADISTIC MULTI-STATE PERFECT PHYLOGENY DECONVOLUTION problem [9]. Details and precise definitions of these concepts are omitted due to space constraints. Although the tree constraints alter the solution space of PCTR problem compared to the PCR problem (see Figure 1 and Figure 2c), PCTR remains NP-hard, as we will show in the following.

► **Theorem 13.** *The PCTR problem is NP-hard even for number $m = 1$ of samples.*

For a given instance (A, B) of the 3-PARTITION problem, we construct an instance of the PCTR problem as follows. The first set $\Pi_1(A, B)$ of clones equals $\{0\} \cup [3q]$ with corresponding $1 \times (3q + 1)$ proportion matrix $U_1(A, B) = [u_{1,i}^{(1)}]$ where $u_{1,i}^{(1)} = a_i/(Bq)$ for all $i \in [3q]$, and $u_{1,0}^{(1)} = 0$. The second set $\Pi_2(A, B)$ of clones equals $\{0\} \cup [q]$ with corresponding $1 \times (q + 1)$ proportion matrix $U_2(A, B) = [u_{1,j}^{(2)}]$ where $u_{1,j}^{(2)} = 1/q$ for all $j \in [q]$, and $u_{1,0}^{(2)} = 0$. The clone tree $T_1(A, B)$ is a star phylogeny rooted at Π_1 -clone $i = 0$ with outgoing edges to each of the remaining Π_1 -clones. Similarly, clone tree $T_2(A, B)$ is also a *star* phylogeny rooted at Π_2 -clone $j = 0$ with outgoing edges to each of the remaining Π_2 -clones. It is easy to verify that $U_1(A, B)$ and $U_2(A, B)$ are proportion matrices for $\Pi_1(A, B)$ and $\Pi_2(A, B)$, respectively. Clearly, this construction takes polynomial time. Figure 3c shows an example. The hardness follows from the following lemma whose proof is omitted due to space constraints.

► **Lemma 14.** *Given proportions $U_1(A, B)$ and clone tree T_1 for clones $\Pi_1(A, B) = \{0\} \cup [3q]$ and proportions $U_2(A, B)$ and clone tree T_2 for clones $\Pi_2(A, B) = \{0\} \cup [q]$, there exists a set Π of clones of size $n = |\Pi| = 4q + 1$, clone tree T and proportion matrix U such that T is a refinement of T_1 and T_2 and $J(U, U_1, U_2) = 0$ if and only if there exists a solution of the 3-PARTITION instance (A, B) .*

4 Methods

We introduce two mixed integer linear programming (MILP) formulations to solve the PCR (Section 4.1) and the PCTR problems (Section 4.2). We implement these two formulations within the algorithm PACTION (PARsimonious Clone Tree reconciliatION), which uses the MILP-solver Gurobi version 9.1. PACTION is available at <https://github.com/elkebir-group/paction>.

4.1 Parsimonious Clone Reconciliation

To solve the PCR problem, we introduce an MILP formulation composed of $\mathcal{O}(n_1 n_2 m)$ variables (including $\mathcal{O}(n_1 n_2)$ binary variables) and $\mathcal{O}(n_1 n_2 m)$ constraints. We introduce binary variables $x_{i,j} \in \{0, 1\}$ for each Π_1 -clone $i \in [n_1]$ and Π_2 -clone $j \in [n_2]$ that indicate if clone (i, j) belongs to Π . As such, the corresponding proportion of clone (i, j) in sample $p \in [m]$ is denoted by the continuous variable $u_{p,i,j} \in [0, 1]$. In the following we define the constraints on these variables by first describing the constraints for consistency and next those for encoding the objective function.

Consistency constraints. This first set of constraints ensure that proportion matrix U is consistent with proportion matrices U_1 and U_2 . We begin by forcing $u_{p,i,j}$ to 0 if (i, j) is not a clone in the solution Π .

$$u_{p,i,j} \leq x_{i,j} \quad \forall p \in [m], i \in [n_1], j \in [n_2].$$

These above constraints allow us to model consistency of the solution U with input proportions $U_1 = [u_{p,i}^{(1)}]$ and $U_2 = [u_{p,j}^{(2)}]$ as follows.

$$\begin{aligned} \sum_{j=1}^{n_2} u_{p,i,j} &= u_{p,i}^{(1)} & \forall p \in [m], i \in [n_1], \\ \sum_{i=1}^{n_1} u_{p,i,j} &= u_{p,j}^{(2)} & \forall p \in [m], j \in [n_2]. \end{aligned}$$

Note that these two sets of constraints imply that $\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} u_{p,i,j} = 1$ for all $p \in [m]$.

Objective function. We minimize the total number of clones in the set Π by minimizing the following objective function.

$$\min \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} x_{i,j}.$$

4.2 Parsimonious Clone Tree Reconciliation

To solve the PCTR problem, we introduce an MILP formulation composed of $\mathcal{O}(n_1 n_2 m)$ variables (including $\mathcal{O}(n_1 n_2)$ binary variables) and $\mathcal{O}(n_1 n_2 m)$ constraints. Similarly to the PCR MILP, we introduce binary variables $x_{i,j} \in \{0, 1\}$ for $i \in [n_1]$ and $j \in [n_2]$ that indicate if clone (i, j) belongs to Π . As such, the corresponding proportion of clone (i, j) in sample $p \in [m]$ is denoted by the continuous variable $u_{p,i,j} \in [0, 1]$. We introduce constraints to model the error $J(U, U_1, U_2)$ used in the objective function, as well constraints to enforce that U is a proportion matrix, and finally constraints to enforce that T is a refinement of T_1 and T_2 .

Correction constraints. Unlike the PCR problem, the proportion matrix U need not be consistent with proportion matrices U_1 and U_2 . We introduce continuous variables $c_{p,i}^{(1)} \in [0, 1]$ for $p \in [m], i \in [n_1]$ and $c_{p,j}^{(2)} \in [0, 1]$ for $p \in [m], j \in [n_2]$ to model the entry-wise absolute differences, i.e., $c_{p,i}^{(1)} = |\sum_{j=1}^{n_2} u_{p,i,j} - u_{p,i}^{(1)}|$ and $c_{p,j}^{(2)} = |\sum_{i=1}^{n_1} u_{p,i,j} - u_{p,j}^{(2)}|$. We do so with the following constraints.

9:10 Parsimonious Clone Tree Reconciliation

$$\begin{aligned}
c_{p,i}^{(1)} &\geq \sum_{j=1}^{n_2} u_{p,i,j} - u_{p,i}^{(1)} && \forall p \in [m], i \in [n_1], \\
c_{p,i}^{(1)} &\geq u_{p,i}^{(1)} - \sum_{j=1}^{n_2} u_{p,i,j} && \forall p \in [m], i \in [n_1], \\
c_{p,j}^{(2)} &\geq \sum_{i=1}^{n_1} u_{p,i,j} - u_{p,j}^{(2)} && \forall p \in [m], j \in [n_2], \\
c_{p,j}^{(2)} &\geq u_{p,j}^{(2)} - \sum_{i=1}^{n_1} u_{p,i,j} && \forall p \in [m], j \in [n_2].
\end{aligned}$$

Proportion matrix constraints. To model that our output matrix U is a proportion matrix, we begin by ensuring that $u_{p,i,j} = 0$ with $x_{i,j} = 0$, i.e., the proportion of clone (i, j) is zero when it is not part of the solution Π with the following constraints.

$$u_{p,i,j} \leq x_{i,j} \quad \forall p \in [m], i \in [n_1], j \in [n_2].$$

Next, we ensure that matrix U is a valid proportion matrix by enforcing that the proportions of the clones in each sample sum to 1.

$$\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} u_{p,i,j} = 1 \quad \forall p \in [m].$$

Refinement constraints. We introduce constraints that ensure that the clone tree T is a refinement of the clone trees T_1 and T_2 . Following condition (iii) in Definition 4, we require that for each clone $(i, j) \neq (r(T_1), r(T_2))$ there only two possible parents, i.e., either (i', j) or (i, j') where $(i', i) \in E(T_1)$ and $(j', j) \in E(T_2)$. We model the first case with continuous variables $z_{(i,i'),j}^{(1)} \in [0, 1]$ and the second case with continuous variables $z_{i,(j,j')}^{(2)}$. More specifically, we model the products $z_{(i,i'),j}^{(1)} = x_{i,j}x_{i',j}$ and $z_{i,(j,j')}^{(2)} = x_{i,j}x_{i,j'}$ with the following constraints.

$$\begin{aligned}
z_{(i,i'),j}^{(1)} &\leq x_{i,j} && \forall (i, i') \in E(T_1), j \in [n_2], \\
z_{(i,i'),j}^{(1)} &\leq x_{i',j} && \forall (i, i') \in E(T_1), j \in [n_2], \\
z_{(i,i'),j}^{(1)} &\geq x_{i,j} + x_{i',j} - 1 && \forall (i, i') \in E(T_1), j \in [n_2], \\
z_{i,(j,j')}^{(2)} &\leq x_{i,j} && \forall i \in [n_1], (j, j') \in E(T_2), \\
z_{i,(j,j')}^{(2)} &\leq x_{i,j'} && \forall i \in [n_1], (j, j') \in E(T_2), \\
z_{i,(j,j')}^{(2)} &\geq x_{i,j} + x_{i,j'} - 1 && \forall i \in [n_1], (j, j') \in E(T_2).
\end{aligned}$$

We now enforce conditions (i) and (ii) in Definition 4 as follows.

$$\begin{aligned}
\sum_{j=1}^{n_2} z_{(i,i'),j}^{(1)} &= 1 && \forall (i, i') \in E(T_1), \\
\sum_{i=1}^{n_1} z_{i,(j,j')}^{(2)} &= 1 && \forall (j, j') \in E(T_2).
\end{aligned}$$

Objective function. Our goal is to minimize the difference between projections of proportion matrix U with U_1 and U_2 . To that end, we minimize the following objective function

$$\min \sum_{p=1}^m \sum_{i=1}^{n_1} c_{p,i}^{(1)} + \sum_{p=1}^m \sum_{j=1}^{n_2} c_{p,j}^{(2)}.$$

We provide the full MILP for reference in Appendix A.

5 Results

5.1 Simulations

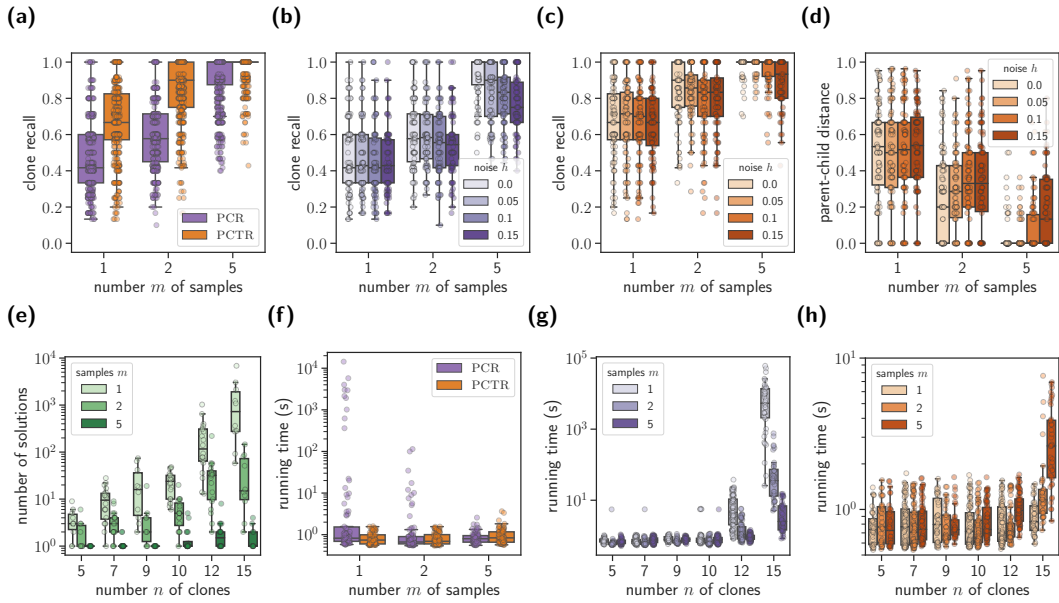
We perform simulations to investigate the performance of PACTION when solving the PCR and PCTR problems under different simulation regimes.

Setup. Given numbers n_1, n_2 of clones, number m of samples and noise parameter $h \in [0, 1]$, we use a three-step procedure to simulate a set Π of $n = n_1 + n_2$ clones whose SNV and CNA evolution is described by a clone tree T and with clone proportions U on m samples. From T and U , we obtain input trees T_1 and T_2 as well as input proportion matrices U_1 and U_2 subject to additional noise h . We detail the three steps in the following.

First, we use an approach based on growing random networks [21] to simulate T : starting from the root vertex (representing the normal clone $(1, 1)$) T 's topology is built by iteratively adding descendant vertices, choosing each parent uniformly at random. Specifically, we label each edge with a single event from either the first set $\{2, \dots, n_1\}$ or second set $\{2, \dots, n_2\}$ of features. Thus, the overall clones Π are obtained by labeling all vertices with a depth-first traversal. Second, we obtain the clone trees T_1 and T_2 by collapsing vertices of T corresponding to identical Π_1 -clones and collapsing vertices of T corresponding to identical Π_2 -clones, respectively. Third, the proportions U of the Π -clones in each sample are simulated by using a Dirichlet distribution with all concentration parameters equal to 1, similarly to previous methods [6, 23]. Proportions U_1 and U_2 are thus obtained following the consistency condition (Definition 2). Furthermore, we introduce noise in these two proportion matrices by mixing in a second draw from the same Dirichlet distribution using the parameter $h \in [0, 1]$ – a value of $h = 0$ indicates the absence of noise. Details are in Appendix B.

We ran PACTION in both PCR and PCTR mode on 360 simulated instances that we obtained by generating 10 instances for each combination of varying parameters. Matching numbers observed in recent cancer genomics studies [16, 18, 44], we varied the numbers $n_1 \in \{3, 5, 8\}$ and $n_2 \in \{3, 5, 8\}$ of clones, the number $m \in \{1, 2, 5\}$ of samples and noise level $h \in \{0, 0.05, 0.1, 0.15\}$. Note that both proportions U_1, U_2 and the simulated trees T_1, T_2 are taken in input in PCTR mode, while only proportions U_1, U_2 are considered in PCR mode.

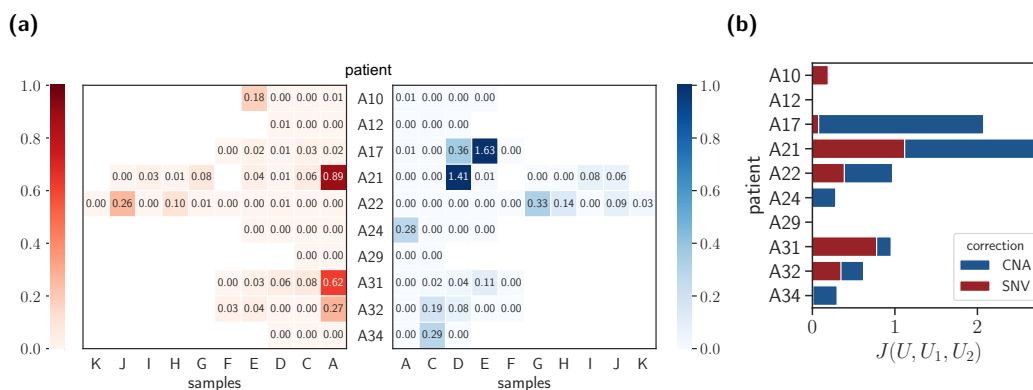
Results. We measure the performance of PACTION based on recall, which is the fraction of ground truth clones that are predicted by our method, i.e., the *clone recall* equals $|\Pi \cap \Pi^*|/|\Pi^*|$ where Π is the set of clones inferred by PACTION and Π^* are the ground truth clones. As expected, PACTION in PCTR mode leverages additional information from the clone trees T_1 and T_2 and thus resulted in higher recall compared to PCR mode (Figure 4a). Interestingly, recall increased with increasing number m of samples, as each additional samples provides additional constraints regarding consistency of the output clone proportions. Breaking down the clone recall by noise level h , we found that performance decreased with increasing noise levels in both PCR mode (Figure 4b) as well as PCTR mode (Figure 4c). However, we



■ **Figure 4 Simulations show that PACTION quickly and accurately reconstructs comprehensive clonal architectures.** (a) Clone recall of PACTION in the PCR and PCTR mode for simulation instances with increasing number m of samples. Clone recall of PACTION in the (b) PCR mode and (c) PCTR mode for different noise levels h and number m of samples. (d) Parent-child distance between the clone tree in the ground truth and the solution of PACTION in the PCTR mode for simulation instances with increasing number m of samples. (e) Number of solutions to the error-free version of the PCTR problem (with additional constraint of $J(U, U_1, U_2) = 0$) by SPRUCE [9] for increasing number n of clones. (f) Running time of PACTION in the PCR and PCTR modes for simulation instances with increasing number m of samples. Running time of PACTION in the (g) PCR mode and the (h) PCTR mode for simulation instances with increasing number n of clones and number m of samples.

found that the PCTR solver better handles increasing noise levels h , with a medial clone recall of 1 for noise level $h = 0$ as well as $h = 0.05$ when number m of samples is 5 (Figure 4c and Figure S1).

Next, we investigated how well PACTION in PCTR mode infers ground truth clone trees T^* . To that end, we computed the parent-child distance [15] between the predicted clone tree T and the clone tree T^* in the ground truth. Specifically, the *parent-child distance* equals the ratio between the size $|E(T) \Delta E(T^*)|$ of the symmetric difference of the edge sets by the size $|E(T) \cup E(T^*)|$ of the union of edge sets. We observed that the clone tree distance is inversely correlated with the clone recall and when the clone recall is 1, the predicted clone tree matches the ground truth perfectly (Figure 4d). Indeed, we observed that performance increases with increasing number m of samples, e.g., for $m = 5$ samples the median parent-child distance is 0 for noise levels $h \in \{0, 0.05, 0.1\}$ indicating that in the majority of these instances PACTION perfectly inferred ground truth trees. The reason why performance drops for decreasing number of samples is because the number of solutions increases with decreasing number of samples (Figure 4e). We used the correspondence between the PCTR problem (subject to the constraint that $J(U, U_1, U_2) = 0$, i.e., the proportions are error-free) and the perfect phylogeny mixture problem solved by SPRUCE [9] to enumerate all solutions for $h = 0$ instances. For instances with a large number of optimal solutions, the PCTR problem and consequently the MILP lacks additional constraints to disambiguate between solutions, thus sometimes reporting solutions that do not match the ground truth.



■ **Figure 5 Overview of PACTION results on samples from 10 metastatic prostate cancer patients [16].** (a) The corrections made by PACTION to the SNV and CNA clone proportions in the samples from each of the 10 patients. (b) The total correction made to clone proportions $J(U, U_1, U_2)$ in samples from each patient.

Finally, we investigated the running times of PACTION in PCR and PCTR modes. Overall, the running times in PCR mode (median of 0.79 s and mean of 385.52 s) were larger than PCTR mode (median of 0.77 s and mean of 0.95 s), likely due to the tree constraints providing more guidance for the MILP solver (Table S1). Interestingly, while running time decreased with increasing number m of samples in PCR mode, the opposite is true in PCTR mode. The reason is that in PCTR mode the MILP is often solved in the first iteration prior to branching, where the running time of solving the linear programming relaxation will depend on the size of the formulation, which in turn depends on m . However, in PCR mode, the solver requires branching, and here additional constraints due to more samples will provide stronger bounds that will lead to more pruning and reduction in overall running time.

In summary, our simulations demonstrate that PACTION is able to quickly and accurately reconstruct ground truth clonal architectures under varying noise levels h , especially when the number m is large and when run in PCTR mode.

5.2 Metastatic prostate cancer

In this study, we analyze whole-genome sequencing data from 49 tumor samples from 10 metastatic prostate cancer patients [16]. In a previous analysis of this data, Gundem et al. [16] identified SNV clones and reconstructed the SNV clone tree for each of the 10 patients. To further investigate the role of CNAs on tumor evolution, the authors annotated the SNV clone trees with CNA events in a *post hoc* analysis by manually comparing and matching frequencies of SNVs and CNAs. However, this approach does not allow us to identify tumor clones that are only distinguished by different CNAs and have the same SNVs. Therefore, there is no information about CNA-only driven tumor clones nor information about the ordering of the CNA events and the SNV events on the same edge of the tree. Such information is crucial to understand cancer progression [38] and is the subject of numerous studies [17, 20, 34]. Therefore, we investigated whether we can use PACTION to provide a more comprehensive analysis of these tumor clonal compositions by jointly considering SNVs and CNAs.

We applied PACTION to previously inferred SNV and CNA clone proportions. First, we used the SNV clone proportions as well as the SNV clone tree T_1 inferred for each patient by Gundem et al. [16]. Note that each edge of the SNV tree represents a cluster of SNV

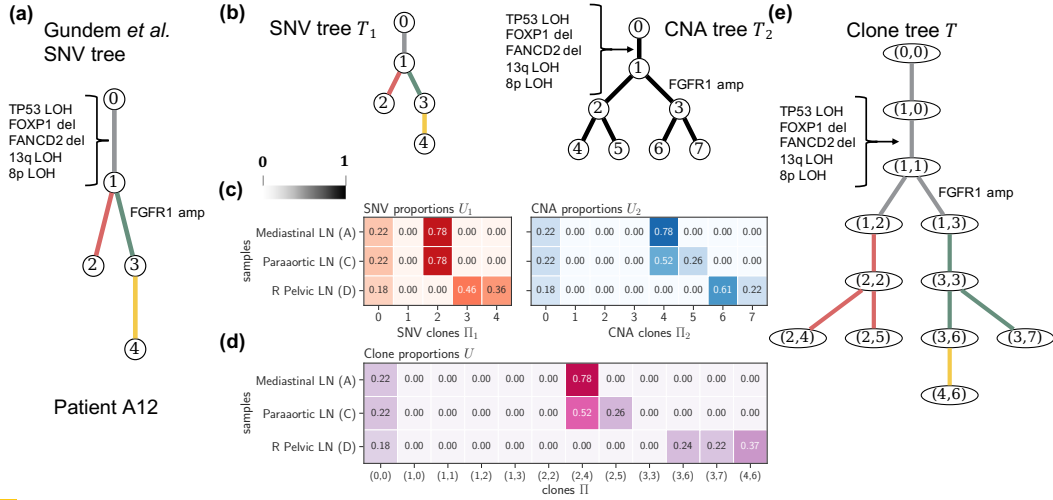


Figure 6 PACTION results for patient A12. (a) The SNV clone tree reported by Gudem et al. [16] where the authors manually annotated edges with CNA events. (b) SNV clone tree T_1 and CNA clone tree T_2 describing the evolution of the SNV clones Π_1 and CNA clones Π_2 in the tumor samples of patient A12, respectively. (c) Proportions U_1 of SNV clones Π_1 and proportions U_2 of CNA clones Π_2 in the four samples of patient A12. (d) Proportions U of tumor clones Π in the four samples of patient A12 inferred by PACTION. (e) Reconciled clone tree T inferred by PACTION. amp: amplification, del: deletion, LOH: loss of heterozygosity.

mutations. As such, we computed the SNV clone proportions U_1 using the published cancer cell fractions of SNVs (details in Appendix C). Second, we used the CNA clones obtained from a previous copy-number analysis [44] of the same patients. Since this previous analysis does not provide CNA clone trees, we enumerated all possible binary trees [2] with the CNA clones as the leaves and independently ran PACTION in PCTR mode with each tree as input. We then selected the CNA clone tree with the smallest correction $J(U, U_1, U_2)$, which for each patient was unique. Overall, we ultimately obtained SNV trees with $n_1 \in \{5, \dots, 16\}$ clones and CNA trees with $n_2 \in \{4, \dots, 8\}$ clones across $m \in \{2, \dots, 10\}$ samples (Table S2).

In all patients but A29, we found that one cannot reconcile independently-inferred SNV and CNA clone trees without additional corrections to the clone proportions. Importantly, this observation highlights that the clone proportions inferred by existing methods are generally characterized by errors (Figure 5a). As previously demonstrated in our simulation study, PACTION, however, reliably handles the presence of noise, enabling the inference of the complete clonal composition and tumor evolution with limited corrections for all patients. Specifically, the corrections applied by PACTION were limited to only a few samples per patient, potentially indicating sample-specific errors in previous analysis or samples with higher levels of noise. Importantly, we also observed that corrections were uniformly needed for both SNV and CNA clone proportions (Figure 5). This important observation highlights that both features are generally characterized by errors and, therefore, one cannot simply leave one feature fixed and use it to reconcile the other feature, as done previously [16].

Notably, we found that the reconciled clone trees inferred by PACTION reveal additional branching events that were previously missed. As an example, in patient A12, Gudem et al. [16] inferred an SNV clone tree with five clones and annotated this tree with five clonal CNA events, including loss-of-heterozygosity (LOH) of gene TP53 and chromosomes 8p and 13q, as well as deletions of genes FOXP1 and FANCD2 (gray edge in Figure 6a). The

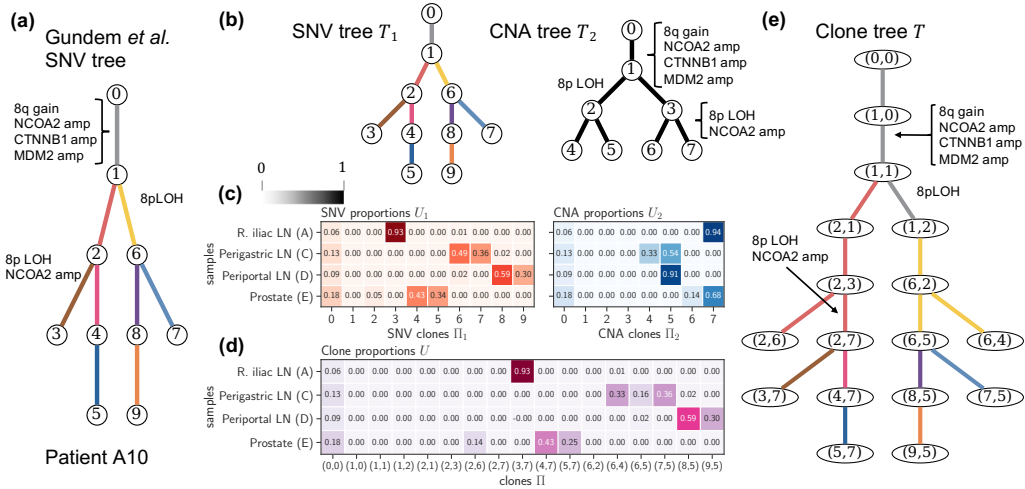


Figure 7 PACTION results for patient A10. (a) The SNV clone tree reported by Gundem et al. [16] where the authors manually annotated edges with CNA events. (b) SNV clone tree T_1 and CNA clone tree T_2 describing the evolution of the SNV clones Π_1 and CNA clones Π_2 in the tumor samples of patient A12, respectively. (c) Proportions U_1 of SNV clones Π_1 and proportions U_2 of CNA clones Π_2 in the four samples of patient A10. (d) Proportions U of tumor clones Π in the four samples of patient A10 inferred by PACTION. (e) Reconciled clone tree T inferred by PACTION. amp: amplification, LOH: loss of heterozygosity.

tree also contains a single subclonal CNA event, amplification of gene *FGFR1* (green edge in Figure 6a). When using PACTION to analyze the previously-inferred SNV and CNA clone proportions, we reconstructed a reconciled clone tree with higher resolution. In fact, PACTION reconstructed a more refined clone tree with 12 clones while only applying modest corrections to the input clone proportions (Figure 5a). Similarly to the published tree, PACTION’s inferred clone tree contains a trunk with the same four clonal CNA events. However, PACTION’s tree contains additional branching events that are absent in the published SNV tree. Specifically, we observed that two SNV clones in the published tree (i.e., 2 and 3) were split into multiple clones in PACTION’s refined tree (i.e., (2, 2), (2, 4), and (2, 5) for SNV clone 2, and (3, 3), (3, 6), and (3, 7) for SNV clone 3). Importantly, a subset of these refined clones are present at large proportions in the sequenced samples (Figure 6d), thus showing that PACTION enables a more fine-grained analysis of current sequencing data.

Finally, we found that the more refined clone trees inferred by PACTION also reveal novel insights about the relative temporal ordering of SNVs and CNAs. This phenomenon is particularly interesting in patient A10 (Figure 7a), for which PACTION inferred a clone tree with 17 clones and relatively high corrections to the previous SNV clone proportions (Figure 7b-d). PACTION’s tree recapitulates the same four clonal CNAs identified in the previous tree, including gain of chromosome 8q and amplifications of genes *NCOA2*, *CTNNB1* and *MDM2* (gray edge in Figure 7a). Importantly, PACTION’s tree also recapitulates subclonal CNA events as in the previous tree but further revealed that these CNA events precede the SNV events placed on the same edges in the published SNV clone tree (Figure 7e). More specifically, PACTION revealed that LOH of chromosome 8p and amplification of gene *NCOA2* occur on the edge from clone (2, 3) to (2, 7) which precedes the SNV cluster represented by the edge from clone (2, 7) to (3, 7). Similarly, PATION revealed that LOH of chromosome 8p occurs on the edge from clone (1, 1) to (1, 2) which precedes the SNV cluster represented by the edge from clone (1, 2) to (6, 2).

In summary, we demonstrated on metastatic prostate cancer patients that PACTION is able to resolve the temporal ordering of mutations and reveal branching events that are either unclear or hidden when the SNV tree or the CNA tree are considered in isolation.

6 Discussion

In this paper, we introduced PACTION, a new algorithm that infers comprehensive tumor clonal compositions by reconciling the clones proportions of both SNVs and CNAs that are inferred by existing methods. Our algorithm can additionally leverage SNV and CNA clone trees reconstructed by existing methods to obtain a refined tumor clone tree and correct potential errors in the input proportions. We formulated two problems, the PCR problem to infer the clones and their proportions, and the PCTR problem to additionally infer tumor clone trees with both SNVs and CNAs. We showed that both problems are NP-hard and can be solved exactly by PACTION using two mixed integer linear programming formulations. We demonstrated the performance of PACTION on simulations, showing that our method accurately reconciles clone trees, reliably handles errors in clone proportions, and scales to practical input sizes. Finally, we applied our method to whole-genome sequencing data from 10 metastatic prostate cancer patients [16], obtaining a higher resolution view of tumor evolution than previously reported.

In addition to the contributions of this study, we foresee four major avenues for future research. First, building upon the established relationship of the error-free PCTR and the cladistic multi-state perfect phylogeny deconvolution problems, we can adapt the existing method SPRUCE [9] to enumerate all possible solution of the PCTR problem in the presence of errors in the input proportions. Second, PACTION can be extended to account for uncertainty in the input clone trees and quantify its effect on the solution space. One way of incorporating the uncertainty in the input clone trees, is to consider a set of possible clone trees for each feature instead of a single input tree, choosing the best tree that leads to the most parsimonious solution. Moreover, we plan to adapt the PCR and PCTR to incorporate probabilistic models that account for uncertainty in the estimated clone proportions. Third, the PCR and PCTR problems can be generalized to reconcile more than two features. For instance, in addition to SNVs and CNAs, tumor cells may be partitioned into clones based on RNA expression or DNA methylation profiles. Finally, a likelihood-based objective function could be used to incorporate a joint evolutionary model for SNVs and CNAs [32].

References

- 1 Ravindra K Ahuja, Thomas L Magnanti, James B Orlin, and K Weihe. Network flows: theory, algorithms and applications. *ZOR-methods and models of operations research*, 41(3):252–254, 1995.
- 2 Johnathan Barnett, Hannah Correia, Peter Johnson, Michael Laughlin, and Kathryn Wilson. Darwin meets graph theory on a strange planet: Counting full n-ary trees with labeled leaves. *Alabama Journal of Mathematics*, 2010.
- 3 Rebecca A Burrell, Nicholas McGranahan, Jiri Bartek, and Charles Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345, 2013.
- 4 Giovanni Ciriello, Martin L Miller, Bülent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz, and Chris Sander. Emerging landscape of oncogenic signatures across human cancers. *Nature genetics*, 45(10):1127–1133, 2013.
- 5 Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. Phylowgs: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome biology*, 16(1):1–20, 2015.

- 6 Mohammed El-Kebir, Layla Oesper, Hannah Acheson-Field, and Benjamin J Raphael. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70, 2015.
- 7 Mohammed El-Kebir, Benjamin J Raphael, Ron Shamir, Roded Sharan, Simone Zaccaria, Meirav Zehavi, and Ron Zeira. Copy-number evolution problems: complexity and algorithms. In *International Workshop on Algorithms in Bioinformatics*, pages 137–149. Springer, 2016.
- 8 Mohammed El-Kebir, Benjamin J Raphael, Ron Shamir, Roded Sharan, Simone Zaccaria, Meirav Zehavi, and Ron Zeira. Complexity and algorithms for copy-number evolution problems. *Algorithms for Molecular Biology*, 12(1):1–11, 2017.
- 9 Mohammed El-Kebir, Gryte Satas, Layla Oesper, and Benjamin J. Raphael. Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures. *Cell Systems*, 3(1):43–53, 2016. doi:10.1016/j.cels.2016.07.004.
- 10 D Fernández-Baca. The perfect phylogeny problem. In D Z Zu and X Cheng, editors, *Steiner Trees in Industries*. Kluwer Academic Publishers, 2000.
- 11 Andrej Fischer, Ignacio Vázquez-García, Christopher JR Illingworth, and Ville Mustonen. High-definition reconstruction of clonal composition in cancer. *Cell reports*, 7(5):1740–1752, 2014.
- 12 Michael R Garey and David S. Johnson. Complexity results for multiprocessor scheduling under resource constraints. *SIAM Journal on Computing*, 4(4):397–411, 1975.
- 13 Michael R. Garey and David S. Johnson. *Computers and intractability. a guide to the theory of np-completeness*, 1983.
- 14 Charles Gawad, Winston Koh, and Stephen R Quake. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 17(3):175, 2016.
- 15 Kiya Govek, Camden Sikes, and Layla Oesper. A consensus approach to infer tumor evolutionary histories. In *Proceedings of the 2018 Acm international conference on bioinformatics, computational biology, and health informatics*, pages 63–72, 2018.
- 16 Gunes Gundem, Peter Van Loo, Barbara Kremeyer, Ludmil B Alexandrov, Jose MC Tubio, Elli Papaemmanuil, Daniel S Brewer, Heini ML Kallio, Gunilla Högnäs, Matti Annala, et al. The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547):353–357, 2015.
- 17 Jun Guo, Hanliang Guo, and Zhanyi Wang. Inferring the temporal order of cancer gene mutations in individual tumor samples. *PLoS One*, 9(2):e89244, 2014.
- 18 Mariam Jamal-Hanjani, Gareth A Wilson, Nicholas McGranahan, Nicolai J Birkbak, Thomas BK Watkins, Selvaraju Veeriah, Seema Shafi, Diana H Johnson, Richard Mitter, Rachel Rosenthal, et al. Tracking the evolution of non-small-cell lung cancer. *New England Journal of Medicine*, 376(22):2109–2121, 2017.
- 19 Yuchao Jiang, Yu Qiu, Andy J Minn, and Nancy R Zhang. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proceedings of the National Academy of Sciences*, 113(37):E5528–E5537, 2016.
- 20 Sahand Khakabimamaghani, Dujian Ding, Oliver Snow, and Martin Ester. Uncovering the subtype-specific temporal order of cancer pathway dysregulation. *PLoS computational biology*, 15(11):e1007451, 2019.
- 21 Paul L Krapivsky and Sidney Redner. Organization of growing random networks. *Physical Review E*, 63(6):066123, 2001.
- 22 Marco L Leung, Alexander Davis, Ruli Gao, Anna Casasent, Yong Wang, Emi Sei, Eduardo Vilar, Dipen Maru, Scott Kopetz, and Nicholas E Navin. Single-cell dna sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome research*, 27(8):1287–1299, 2017.
- 23 Salem Malikic, Andrew W McPherson, Nilgun Donmez, and Cenk S Sahinalp. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31(9):1349–1356, 2015.
- 24 Nicholas McGranahan and Charles Swanton. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer cell*, 27(1):15–26, 2015.

- 25 Andrew W McPherson, Andrew Roth, Gavin Ha, Cedric Chauve, Adi Steif, Camila PE de Souza, Peter Eirew, Alexandre Bouchard-Côté, Sam Aparicio, S Cenk Sahinalp, et al. Remixt: clone-specific genomic structure estimation in cancer. *Genome biology*, 18(1):1–14, 2017.
- 26 Faiyaz Notta, Michelle Chan-Seng-Yue, Mathieu Lemire, Yilong Li, Gavin W Wilson, Ashton A Connor, Robert E Denroche, Sheng-Ben Liang, Andrew MK Brown, Jaeseung C Kim, et al. A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature*, 538(7625):378–382, 2016.
- 27 Peter C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.
- 28 Layla Oesper, Ahmad Mahmoody, and Benjamin J Raphael. Theta: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome biology*, 14(7):1–21, 2013.
- 29 Victoria Popic, Raheleh Salari, Iman Hajirasouliha, Dorna Kashef-Haghighi, Robert B West, and Serafim Batzoglou. Fast and scalable inference of multi-sample cancer lineages. *Genome biology*, 16(1):1–17, 2015.
- 30 Dikshant Pradhan and Mohammed El-Kebir. On the non-uniqueness of solutions to the perfect phylogeny mixture problem. In *RECOMB International conference on Comparative Genomics*, pages 277–293. Springer, 2018.
- 31 Gryte Satas and Benjamin J Raphael. Tumor phylogeny inference using tree-constrained importance sampling. *Bioinformatics*, 33(14):i152–i160, 2017.
- 32 Gryte Satas, Simone Zaccaria, Geoffrey Mon, and Benjamin J Raphael. Scarlet: Single-cell tumor phylogeny inference with copy-number constrained mutation losses. *Cell Systems*, 10(4):323–332, 2020.
- 33 Roland F Schwarz, Anne Trinh, Botond Sipos, James D Brenton, Nick Goldman, and Florian Markowitz. Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Comput Biol*, 10(4):e1003535, 2014.
- 34 Kathleen Sprouffske, John W Pepper, and Carlo C Maley. Accurate reconstruction of the temporal order of mutations in neoplastic progression. *Cancer prevention research*, 4(7):1135–1144, 2011.
- 35 Francesco Strino, Fabio Parisi, Mariann Micsinai, and Yuval Kluger. Trap: a tree approach for fingerprinting subclonal tumor composition. *Nucleic acids research*, 41(17):e165–e165, 2013.
- 36 Linda K Sundermann, Jeff Wintersinger, Gunnar Rätsch, Jens Stoye, and Quaid Morris. Reconstructing tumor evolutionary histories and clone trees in polynomial-time with submarine. *PLoS computational biology*, 17(1):e1008400, 2021.
- 37 Maxime Tarabichi, Adriana Salcedo, Amit G Deshwar, Máire Ni Leathlobhair, Jeff Wintersinger, David C Wedge, Peter Van Loo, Quaid D Morris, and Paul C Boutros. A practical guide to cancer subclonal reconstruction from dna sequencing. *Nature methods*, 18(2):144–155, 2021.
- 38 Hamid Teimouri and Anatoly B Kolomeisky. Temporal order of mutations influences cancer initiation dynamics. *bioRxiv*, 2021.
- 39 ICGC The, TCGA Pan-Cancer Analysis of Whole, Genomes Consortium, et al. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82, 2020.
- 40 Thomas BK Watkins, Emilia L Lim, Marina Petkovic, Sergi Elizalde, Nicolai J Birkbak, Gareth A Wilson, David A Moore, Eva Grönroos, Andrew Rowan, Sally M Dewhurst, et al. Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature*, 587(7832):126–132, 2020.
- 41 Taoyang Wu, Vincent Moulton, and Mike Steel. Refining phylogenetic trees given additional data: An algorithm based on parsimony. *IEEE/ACM transactions on computational biology and bioinformatics*, 6(1):118–125, 2008.
- 42 Simone Zaccaria, Mohammed El-Kebir, Gunnar W Klau, and Benjamin J Raphael. The copy-number tree mixture deconvolution problem and applications to multi-sample bulk sequencing tumor data. In *International Conference on Research in Computational Molecular Biology*, pages 318–335. Springer, 2017.

- 43 Simone Zaccaria, Mohammed El-Kebir, Gunnar W Klau, and Benjamin J Raphael. Phylogenetic copy-number factorization of multiple tumor samples. *Journal of Computational Biology*, 25(7):689–708, 2018.
- 44 Simone Zaccaria and Benjamin J Raphael. Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. *Nature communications*, 11(1):1–13, 2020.

A MILP formulation for the PCTR problem

$$\min \sum_{p=1}^m \sum_{i=1}^{n_1} c_{p,i}^{(1)} + \sum_{p=1}^m \sum_{j=1}^{n_2} c_{p,j}^{(2)} \quad (2)$$

$$\text{s.t. } c_{p,i}^{(1)} \geq \sum_{j=1}^{n_2} u_{p,i,j} - u_{p,i}^{(1)} \quad \forall p \in [m], i \in [n_1], \quad (3)$$

$$c_{p,i}^{(1)} \geq u_{p,i}^{(1)} - \sum_{j=1}^{n_2} u_{p,i,j} \quad \forall p \in [m], i \in [n_1], \quad (4)$$

$$c_{p,j}^{(2)} \geq \sum_{i=1}^{n_1} u_{p,i,j} - u_{p,j}^{(2)} \quad \forall p \in [m], j \in [n_2], \quad (5)$$

$$c_{p,j}^{(2)} \geq u_{p,j}^{(2)} - \sum_{i=1}^{n_1} u_{p,i,j} \quad \forall p \in [m], j \in [n_2], \quad (6)$$

$$u_{p,i,j} \leq x_{i,j} \quad \forall p \in [m], i \in [n_1], j \in [n_2], \quad (7)$$

$$\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} u_{p,i,j} = 1 \quad \forall p \in [m], \quad (8)$$

$$z_{(i,i'),j}^{(1)} \leq x_{i,j} \quad \forall (i, i') \in E(T_1), j \in [n_2], \quad (9)$$

$$z_{(i,i'),j}^{(1)} \leq x_{i',j} \quad \forall (i, i') \in E(T_1), j \in [n_2], \quad (10)$$

$$z_{(i,i'),j}^{(1)} \geq x_{i,j} + x_{i',j} - 1 \quad \forall (i, i') \in E(T_1), j \in [n_2]. \quad (11)$$

$$z_{i,(j,j')}^{(2)} \leq x_{i,j} \quad \forall i \in [n_1], (j, j') \in E(T_2), \quad (12)$$

$$z_{i,(j,j')}^{(2)} \leq x_{i,j'} \quad \forall i \in [n_1], (j, j') \in E(T_2), \quad (13)$$

$$z_{i,(j,j')}^{(2)} \geq x_{i,j} + x_{i,j'} - 1 \quad \forall i \in [n_1], (j, j') \in E(T_2), \quad (14)$$

$$\sum_{j=1}^{n_2} z_{(i,i'),j}^{(1)} = 1 \quad \forall (i, i') \in E(T_1), \quad (15)$$

$$\sum_{i=1}^{n_1} z_{i,(j,j')}^{(2)} = 1 \quad \forall (j, j') \in E(T_2), \quad (16)$$

$$x_{i,j} \in \{0, 1\}, \quad \forall i \in [n_1], j \in [n_2], \quad (17)$$

$$u_{p,i}^{(1)} \in [0, 1], \quad \forall p \in [m], i \in [n_1], \quad (18)$$

$$u_{p,j}^{(2)} \in [0, 1], \quad \forall p \in [m], j \in [n_2], \quad (19)$$

$$c_{p,i}^{(1)} \in [0, 1], \quad \forall p \in [m], i \in [n_1], \quad (20)$$

$$c_{p,j}^{(2)} \in [0, 1], \quad \forall p \in [m], j \in [n_2], \quad (21)$$

$$z_{(i,i'),j}^{(1)} \geq 0 \quad \forall (i, i') \in E(T_1), j \in [n_2], \quad (22)$$

$$z_{i,(j,j')}^{(2)} \geq 0 \quad \forall i \in [n_1], (j, j') \in E(T_2). \quad (23)$$

B Simulation Details

We perturb the proportion matrices U_1 and U_2 by introducing noise following a user-defined level $h \in [0, 1]$. For each sample $p \in [m]$, let $\mathbf{u}_p^{(1)} = [u_{p,i}^{(1)}]$ for $i \in [n_1]$ and $\mathbf{u}_p^{(2)} = [u_{p,j}^{(2)}]$ for $j \in [n_2]$. The perturbed proportions $\bar{\mathbf{u}}_p^{(1)}$ and $\bar{\mathbf{u}}_p^{(2)}$ are drawn from the following distributions

$$\begin{aligned}\bar{\mathbf{u}}_p^{(1)} &\sim (1-h)\mathbf{u}_p^{(1)} + h\text{Dir}(\mathbf{1}_{n_1}), \quad \forall p \in [m], \\ \bar{\mathbf{u}}_p^{(2)} &\sim (1-h)\mathbf{u}_p^{(2)} + h\text{Dir}(\mathbf{1}_{n_2}), \quad \forall p \in [m].\end{aligned}$$

The resulting proportion matrices are $\bar{U}_1 = [\bar{u}_{p,i}^{(1)}]$ for $p \in [m], i \in [n_1]$ and $\bar{U}_2 = [\bar{u}_{p,j}^{(2)}]$ for $p \in [m], j \in [n_2]$. Note that when noise level $h = 0$, we have $\bar{U}_1 = U_1$ and $\bar{U}_2 = U_2$. Also, for any $h \in [0, 1]$, the matrices \bar{U}_1 and \bar{U}_2 satisfy the conditions laid out in the definition of proportion matrices (Definition 2).

C Computation of SNV Clone Proportions

Each edge of the SNV clone tree T_1 reported by Gundem et al. [16] represents a set of mutations, also known as mutation clusters. As such, for a SNV clone tree T_1 with n_1 vertices, there are $n_1 - 1$ mutation clusters. The authors have provided the cancer cell fraction (CCF) for each of the mutation clusters in each sample of the ten patients. They used pigeonhole principle (PPH) to construct the SNV clone tree manually. For a given patient, let $F \in [0, 1]^{m \times (n_1 - 1)}$ be the CCF matrix such that $F = [f_{p,k}]$ and $f_{p,k}$ is the CCF of mutation cluster $k \in [n_1 - 1]$ in sample $p \in [m]$. The SNV clone tree T_1 , excluding the root vertex which represent the normal cell, is used to construct a perfect phylogeny matrix B [30]. We use the perfect phylogeny matrix B and the CCF matrix F to get the proportion U' of SNV clones, excluding the normal clone, in each sample of the ten patients by solving the following linear program

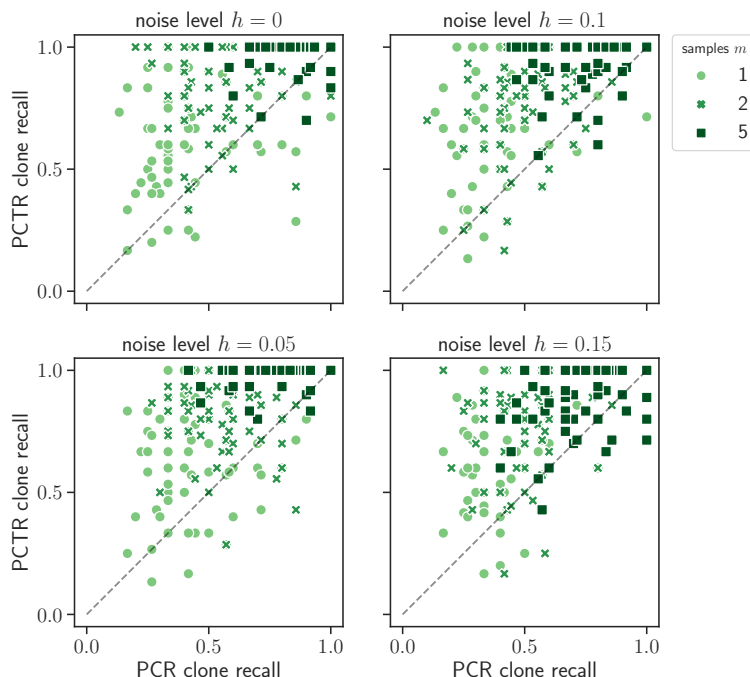
$$\begin{aligned}\min & |F - BU'|_1, \\ \text{s.t.} & 0 \leq u_{p,i} \leq 1, \quad \forall p \in [m], i \in [n_1 - 1], \\ & \sum_{i=1}^{n_1-1} u_{p,i} = 1, \quad \forall p \in [m],\end{aligned}$$

where $|\cdot|_1$ is the entry-wise L_1 norm. Finally, we correct the proportion matrix U' for the purity of the tumor samples (also known as tumor cellularity), which is the proportion of cancer cells in the tumor. We use the proportion of normal cells in each sample, inferred by HATCHet [44], to compute the purity of the tumor samples. Let $\gamma \in [0, 1]^{m \times 1}$ be a vector such that $\gamma_{p,1}$ is the purity of sample $p \in [m]$ inferred using HATCHet. The proportion matrix $U \in [0, 1]^{m \times n_1}$ of the SNV clones is given by

$$U = [\text{Diag}(\gamma)U' \quad \mathbf{1}_m - \gamma]$$

where $\mathbf{1}_m$ is a $m \times 1$ vector with all entries equal to 1 and $\text{Diag}(\gamma)$ is a $m \times m$ diagonal matrix with the diagonal elements given by the entries of the vector γ . It is easy to see that the proportion matrix U satisfies the conditions for being a proportion matrix (see Definition 1).

D Supplementary Figures and Tables



■ **Figure S1** Clone recall for the two modes of PACTION on the simulated instances. We show the clone recall of PACTION with the PCR and the PCTR mode on the simulated instances for varying noise levels h and number m of samples. For majority of simulated instances, PACTION in the PCTR mode has a higher recall compared to the PCR mode.

■ **Table S1** Median running time of PACTION in PCT and PCTR modes for simulation instances with varying number of samples m .

number of samples m	PCR runtime (s)	PCTR runtime (s)
1	0.84820	0.74365
2	0.6949	0.7379
5	0.81985	0.84460

■ **Table S2** Statistics of the metastatic prostate cancer data [16]. Number m of samples, number n_1 of SNV clones and number n_2 of CNA clones for the 10 patients from Gundem et al. [16]. The CNA clones were identified using HATCHet [44].

patient	number m of samples	number n_1 of SNV clones	number n_2 of CNA clones
A10	4	10	8
A12	3	5	8
A17	5	11	6
A21	8	15	6
A22	10	16	4
A24	4	10	4
A29	2	6	4
A31	5	11	6
A32	5	13	6
A34	3	14	6