On the Non-uniqueness of Solutions to the Perfect Phylogeny Mixture Problem

Dikshant Pradhan[1] and Mohammed El-Kebir[2]

[1]Department of Bioengineering, University of Illinois at Urbana-Champaign, IL 61801. [2]Department of Computer Science, University of Illinois at Urbana-Champaign, IL 61801.

Abstract. Tumors exhibit extensive intra-tumor heterogeneity, the presence of groups of cellular populations with distinct sets of somatic mutations. This heterogeneity is the result of an evolutionary process, described by a phylogenetic tree. The problem of reconstructing a phylogenetic tree T given bulk sequencing data from a tumor is more complicated than the classic phylogeny inference problem. Rather than observing the leaves of T directly, we are given mutation frequencies that are the result of mixtures of the leaves of T. The majority of current tumor phylogeny inference methods employ the perfect phylogeny evolutionary model. In this work, we show that the underlying Perfect Phylogeny Mixture combinatorial problem typically has multiple solutions. We provide a polynomial-time computable upper bound on the number of solutions. We use simulations to identify factors that contribute to and counteract non-uniqueness of solutions. In addition, we study the sampling performance of current methods, identifying significant biases.

1 Introduction

Cancer is characterized by somatic mutations that accumulate in a population of cells, leading to the formation of genetically distinct *clones* within the same tumor [19]. This *intra-tumor heterogeneity* is the main cause of relapse and resistance to treatment [24]. The evolutionary process that led to the formation of a tumor can be described by a *phylogenetic tree* whose leaves correspond to tumor cells at the present time and whose edges are labeled by somatic mutations. To elucidate the mechanisms behind tumorigenesis [22, 24] and identify treatment strategies [6, 28], we require algorithms that accurately infer a phylogenetic tree from DNA sequencing data of a tumor.

Most cancer sequencing studies, including those from The Cancer Genome Atlas [12] and the International Cancer Genome Consortium [8], use bulk DNA sequencing technology, where samples are a mixture of millions of cells. While in classic phylogenetics, one is asked to infer a phylogenetic tree given its leaves, with bulk sequencing data we are asked to infer a phylogenetic tree given mixtures of its leaves in the form of mutation frequencies. More specifically, one first identifies a set of loci containing somatic mutations present in the tumor by sequencing and comparing the aligned reads of a matched normal sample and

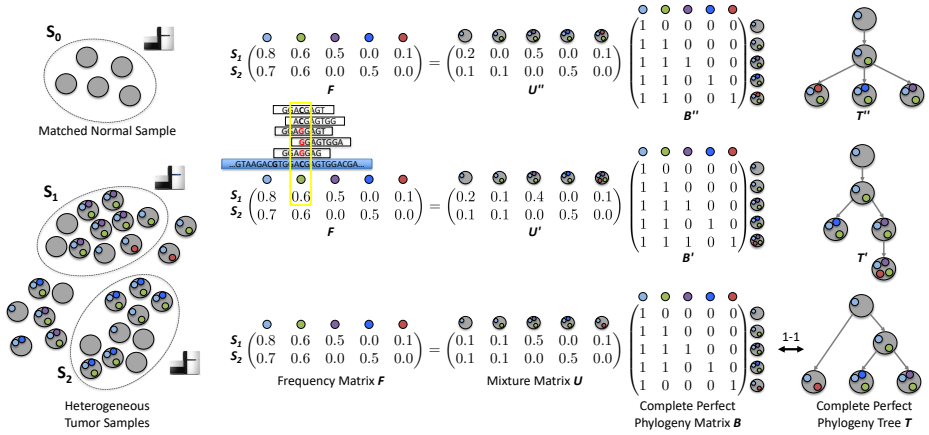

Fig. 1: **Overview of the Perfect Phylogeny Mixture (PPM) problem.** By comparing the aligned reads obtained from bulk DNA sequencing data of a matched normal sample and m tumor samples, we identify n somatic mutations and their frequencies $F = [f_{p,c}]$. In the PPM problem, we are asked to factorize F into a mixture matrix U and a complete perfect phylogeny matrix B, explaining the composition of the m tumor samples and the evolutionary history of the n mutations present in the tumor, respectively. Typically, an input frequency matrix admits multiple distinct solutions. Here, matrix F has three solutions: (U, B), (U', B') and (U'', B''), where only (U, B) is the correct solution.

one or more tumor samples. Based on the number reads of each mutation locus in a sample, we obtain *mutation frequencies* indicating the fraction of cells in the tumor sample that contain each mutation. From these frequencies, the task is to infer the phylogenetic tree under an appropriate evolutionary model that generated the data.

The most commonly used evolutionary model in cancer phylogenetics is the *two-state perfect phylogeny* model, where mutations adhere to the infinite sites assumption [2, 3, 10, 11, 16, 17, 20, 23, 29]. That is, for each mutation locus the actual mutation occurred exactly once in the evolutionary history of the tumor and was subsequently never lost. The underlying combinatorial problem of the majority of current methods is the Perfect Phylogeny Mixture (PPM) problem. Given an $m \times n$ frequency matrix F, we are asked to explain the composition of the m tumor samples and the evolutionary history of the n mutations. More specifically, we wish to factorize F into a mixture matrix U and a perfect phylogeny matrix B. Not only is this problem NP-complete [3], but multiple perfect phylogeny trees may be inferred from the same input matrix F (Fig. 1). Tumor phylogenies have been used to identify mutations that drive cancer progression [9,18], to assess the interplay between the immune system and the clonal architecture of a tumor [15, 30] and to identify common evolutionary patterns in tumorigenesis and metastasis [25, 26]. To avoid any bias in such downstream

analyses, all possible solutions must be considered. While non-uniqueness of solutions to PPM has been recognized in the field [4,17], a rigorous analysis of its extent and consequences on sampling by current methods has been missing.

In this paper, we study the non-uniqueness of solutions to the PPM problem. We give a upper bound on the number of solutions that can be computed in polynomial time. Using simulations, we identify the factors that contribute to non-uniqueness. In addition, we empirically study how, in addition to bulk sequencing, incorporating single-cell and long-read sequencing technologies affects non-uniqueness. Upon finding that current Markov chain Monte Carlo methods fail to sample uniformly from the solution space, we describe a simple rejection sampling algorithm that is able to sample uniformly for modest numbers n of mutations.

2 Preliminaries

In this section, we review the Perfect Phylogeny Mixture problem, as introduced in [3] (where it was the called the Variant Allele Frequency Factorization Problem or VAFFP). As input, we are given a frequency matrix $F = [f_{p,c}]$ composed of allele frequencies of n single-nucleotide variants (SNVs) measured in m bulk DNA sequencing samples. In the following, we refer to SNVs as mutations.

Definition 1. *An $m \times n$ matrix $F = [f_{p,c}]$ is a* frequency matrix *provided $f_{p,c} \in [0, 1]$ for all samples $p \in [m]$ and mutations $c \in [n]$.*

Each frequency $f_{p,c}$ indicates the proportion of cells in sample p that have mutation c. The evolutionary history of all n mutations is described by a phylogenetic tree. We assume the absence of homoplasy and define a complete perfect phylogeny tree T as follows.

Definition 2. *A rooted tree T on n vertices is a* complete perfect phylogeny tree *provided each edge of T is labeled with exactly one mutation from $[n]$ and no mutation appears more than once in T.*

We call the unique mutation $r \in [n]$ that does not label any edge of a complete perfect phylogeny tree T the *founder mutation*. Equivalently, we may represent a complete perfect phylogeny tree by an $n \times n$ binary matrix B subject to the following constraints.

Definition 3. *An $n \times n$ binary matrix $B = [b_{c,d}]$ is an* n-complete perfect phylogeny matrix *provided:*

1. *There exists exactly one $r \in [n]$ such that $\sum_{c=1}^{n} b_{r,c} = 1$.*
2. *For each $d \in [n] \setminus \{r\}$ there exists exactly one $c \in [n]$ such that $\sum_{e=1}^{n} b_{d,e} - \sum_{e=1}^{n} b_{c,e} = 1$, and $b_{d,e} \geq b_{c,e}$ for all $e \in [n]$.*
3. *$b_{c,c} = 1$ for all $c \in [n]$.*

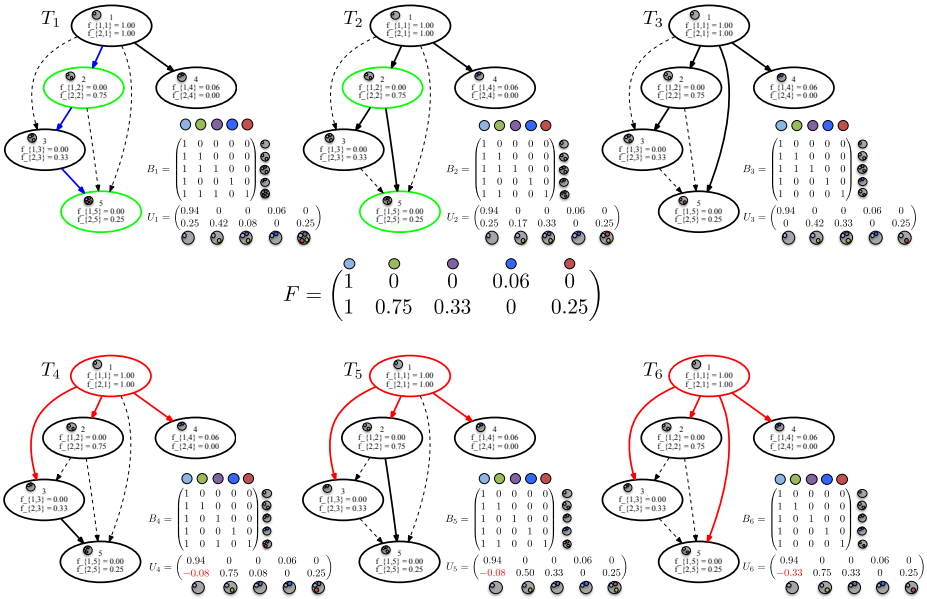

Fig. 2: **Example PPM instance F has three solutions.** Frequency matrix F corresponds to a simulated $n = 5$ instance (#9) and has $m = 2$ samples. The ancestry graph G_F has six spanning arborescences. Among these, only trees T_1, T_2 and T_3 satisfy the sum condition (SC), whereas trees T_4, T_5 and T_6 violate (SC) leading to negative entries in U_4, U_5 and U_6. Tree T_1 is the simulated tree of this instance. Trees T_2 and T_3 differ from T_1 by only one edge, and thus each have an edge recall of $3/4 = 0.75$.

While the rows of a perfect phylogeny matrix B correspond to the leaves of a perfect phylogeny tree T (as per Definition 1), a *complete* perfect phylogeny matrix B includes all vertices of T. The final ingredient is an $m \times n$ mixture matrix U defined as follows.

Definition 4. *An $m \times n$ matrix $U = [u_{p,c}]$ is a* mixture matrix *provided $u_{p,c} \in [0, 1]$ for all samples $p \in [m]$ and mutations $c \in [n]$, and $\sum_{c=1}^{n} u_{p,c} \leq 1$ for all samples $p \in [m]$.*

The forward problem of obtaining a frequency matrix F from a complete perfect phylogeny matrix B and mixture matrix U is trivial. That is, $F = UB$. We are interested in the inverse problem, which is defined as follows.

*Problem 1 (*Perfect Phylogeny Mixture *(*PPM*)).* Given a frequency matrix F, find a complete perfect phylogeny matrix B and mixture matrix U such that $F = UB$.

El-Kebir et al. [3] showed that a solution to PPM corresponds to a constrained spanning arborescence of a directed graph G_F obtained from F, as

illustrated in Fig. 8. This directed graph G_F is called the *ancestry graph* and is defined as follows.

Definition 5. *The* ancestry graph G_F *obtained from frequency matrix* $F = [f_{p,c}]$ *has* n *vertices* $V(G_F) = \{1, \ldots, n\}$ *and there is a directed edge* $(c, d) \in E(G_F)$ *if and only if* $f_{p,c} \geq f_{p,d}$ *for all samples* $p \in [m]$.

As shown in [3], the square matrix B is invertible and thus matrix U is determined by F and B. We denote the set of children of the vertex corresponding to a mutation $c \in [n] \setminus \{r\}$ by $\delta(c)$, and we define $\delta(r) = \{r(T)\}$.

Proposition 1 (Ref. [3]). *Given frequency matrix* $F = [f_{p,c}]$ *and complete perfect phylogeny matrix* $B = [b_{c,d}]$, *matrix* $U = [u_{p,c}]$ *where* $u_{p,c} = f_{p,c} - \sum_{d \in \delta(c)} f_{p,d}$ *is the the unique matrix* U *such that* $F = UB$.

For matrix U to be a mixture matrix, it is necessary and sufficient to enforce non-negativity as follows.

Theorem 1 (Ref. [3]). *Let* $F = [f_{p,c}]$ *be a frequency matrix and* G_F *be the corresponding ancestry graph. Then, complete perfect phylogeny matrix* B *and associated matrix* U *are a solution to* PPM *instance* F *if and only if* B *encodes a spanning arborescence* T *of* G_F *satisfying*

$$f_{p,c} \geq \sum_{d \in \delta_{out}(c)} f_{p,d} \qquad \forall p \in [m], c \in [n]. \tag{SC}$$

The above equation is known as the sum condition (SC), which requires that any mutation with multiple children have a greater frequency than the sum of the frequencies of its children in all samples. In this equation, $\delta_{\text{out}}(c)$ denotes the set of children of vertex c in rooted tree T. A *spanning arborescence* T of a directed graph G_F is defined as a subset of edges that induce a rooted tree that spans all vertices of G_F.

While finding a spanning arborescence in a directed graph can be done in linear time (e.g., using a depth-first or breadth-first search), the problem of finding a spanning arborescence in G_F adhering to (SC) is NP-hard [3,4]. Moreover, the same input frequency matrix F may admit more than one solution (Fig. 2).

3 Methods

3.1 Characterization of the solution space

Let F be a frequency matrix and let G_F be the corresponding ancestry graph. By Theorem 1, we have that solutions to the PPM instance F are spanning arborescences T in the ancestry graph G_F that satisfy (SC). In this section, we describe additional properties that further characterize the solution space. We start with the ancestry graph G_F.

Fact 1. *If there exists a path from vertex* c *to vertex* d *then* $(c, d) \in E(G_F)$.

A pair of mutations that are not connected by a path in G_F correspond to two mutations that must occur on distinct branches in any solution. Such pairs of incomparable mutations are characterized as follows.

Fact 2. *Ancestry graph G_F does not contain the edge (c, d) nor the edge (d, c) if and only if there exist two samples $p, q \in [m]$ such that $f_{p,c} > f_{p,d}$ and $f_{q,c} < f_{q,d}$.*

We define the branching coefficient as follows.

Definition 6. *The* branching coefficient *$\gamma(G_F)$ is the fraction of unordered pairs (c, d) of distinct mutations such that $(c, d) \notin E(G_F)$ and $(d, c) \notin E(G_F)$.*

In the single-sample case, where frequency matrix F has $m = 1$ sample, we have that $\gamma(G_F) = 0$. This is because either $f_{1,c} \geq f_{1,d}$ or $f_{1,d} \geq f_{1,c}$ for any ordered pair (c, d) of distinct mutations. Since an arborescence is a rooted tree, we have the following fact.

Fact 3. *For G_F to contain a spanning arborescence there must exist a vertex in G_F from which all other vertices are reachable.*

Note that G_F may contain multiple source vertices from which all other vertices are reachable. Such source vertices correspond to repeated columns in F whose entries are greater than or equal to every other entry in the same row. In most cases the ancestry graph G_F does not contain any directed cycles because of the following property.

Fact 4. *Ancestry graph G_F is a directed acyclic graph (DAG) if and only if F has no repeated columns.*

In the case where G_F is a DAG and contains at least one spanning arborescence, we know that all spanning arborescence T of G_F share the same root vertex. This root vertex r is the unique vertex of G_F with in-degree 0.

Fact 5. *If G_F is a DAG and contains a spanning arborescence then there exists exactly one vertex r in G_F from which all other vertices are reachable.*

Fig. 2 shows the solutions to a PPM instance F with $m = 2$ tumor samples and $n = 5$ mutations. Since F has no repeated columns, the corresponding ancestry graph G_F is a DAG. Vertex $r = 1$ is the unique vertex of G_F without any incoming edges. There are three solutions to F, i.e. T_1, T_2 and T_3 are spanning arborescences of G_F, each rooted at vertex $r = 1$ and each satisfying (SC). How do we know that F has three solutions in total? This leads to the following problem.

*Problem 2 (#-*Perfect Phylogeny Mixture *(#*PPM*)).* Given a frequency matrix F, count the number of pairs (U, B) such that B is a complete perfect phylogeny matrix, U is a mixture matrix and $F = UB$.

Since deciding whether a frequency matrix F can be factorized into a complete perfect phylogeny matrix B and a mixture matrix U is NP-complete [3,4], the corresponding counting problem is NP-hard.[1] Since solutions to F correspond to a subset of spanning arboscences of G_F that satisfy (SC), we have the following fact.

Fact 6. *The number of solutions to a* PPM *instance F is at most the number of spanning arborescences in the ancestry graph G_F.*

Kirchhoff's elegant matrix tree theorem [13] uses linear algebra to count the number of spanning trees in a simple graph. Tutte extended this theorem to count spanning arborescences in a directed graph $G = (V, E)$ [27]. Briefly, the idea is to construct the $n \times n$ Laplacian matrix $L = [\ell_{i,j}]$ of G, where

$$\ell_{i,j} = \begin{cases} \deg_{\text{in}}(j), & \text{if } i = j, \\ -1, & \text{if } i \neq j \text{ and } (i,j) \in E \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Then, the number of spanning arborescences N_i rooted at vertex i is $\det(\hat{L}_i)$, where $\hat{L}_i$ is the matrix obtained from L by removing the i-th row and column. Thus, the total number of spanning arborescences in G is $\sum_{i=1}^{n} \det(\hat{L}_i)$.

By Fact 4, we have that G_F is a DAG if F has no repeated columns. In addition, by Fact 5, we know that G_F must have a unique vertex r with no incoming edges. We have the following technical lemma.

Lemma 1. *Let G_F be a DAG and let $r(G_F)$ be its unique source vertex. Let π be a topological ordering of the vertices of G_F. Let $L' = [\ell'_{i,j}]$ be the matrix obtained from $L = [\ell_{i,j}]$ by permuting its rows and columns according to π, i.e. $\ell'_{i,j} = \ell_{\pi(i),\pi(j)}$. Then, L' is an upper triangular matrix and $\pi(1) = r(G_F)$.*

Proof. Assume for a contradiction that L' is not upper triangular. Thus, there must exist vertices $i, j \in [n]$ such that $j > i$ and $\ell'_{j,i} \neq 0$. By definition of L and L', we have that $\ell'_{j,i} = -1$. Thus $(\pi(j), \pi(i)) \in E(G_F)$, which yields a contradiction with π being a topological ordering of G_F. Hence, L' is upper triangular. From Fact 5 it follows that $\pi(1) = r(G_F)$. □

Since the determinant of an upper triangular matrix is the product of its diagonal entries, it follows from the previous lemma that $\det(\hat{L}'_1) = \prod_{i=1}^{n-1} \hat{\ell}'_{i,i}$. Combining this fact with Tutte's directed matrix-tree theorem, yields the following result.

Theorem 2. *Let F be a frequency matrix without any repeated columns and let r be the unique mutation such that $f_{p,r} \geq f_{p,c}$ for all mutations c and samples p. Then the number of solutions to F is at most the product of the in-degrees of all vertices $c \neq r$ in G_F.*

[1] We expect the counting problem #PPM to be #P-complete, as to date no NP-complete problem has been found whose counting version is not NP-complete [14]. To prove that #PPM is #P-complete, we need to give a parsimonious reduction from a known #P-complete problem to #PPM.

In Fig. 2, the number of spanning arborescences in G_F is $\deg_{\text{in}}(2) \cdot \deg_{\text{in}}(3) \cdot \deg_{\text{in}}(4) \cdot \deg_{\text{in}}(5) = 1 \cdot 2 \cdot 1 \cdot 3 = 6$. To compute the number of spanning arborescences of G_F that satisfy (SC), we can simply enumerate all spanning arborescences using, for instance, the Gabow-Myers algorithm [7] and only output those that satisfy (SC). El-Kebir et al. [4] extended this algorithm such that it maintains (SC) as an invariant while growing arborescences. Applying both algorithms on the instance in Fig. 2 reveals that trees T_1, T_2 and T_3 comprise all solutions to F. We note that the enumeration algorithm in [4] has not been shown to be an output-sensitive algorithm.

3.2 Additional constraints on the solution space

Long-read sequencing. Most cancer sequencing studies are performed using next-generation sequencing technology, producing short reads containing between 100 and 1000 basepairs. Due to the small size of short reads, it is highly unlikely to observe two mutations that occur on the same read (or read pair). With (synthetic) long read sequencing technology, including 10X Genomics, Pacbio and Oxford Nanopore, one is able to obtain reads with millions of basepairs. Thus, it becomes possible to observe long reads that contain more than one mutation.

As described in [1], the key insight is that a pair (c, d) of mutations that occur on the same read orginate from a single DNA molecule of a single cell, and thus c and d must occur on the same path in the phylogenetic tree. Such mutation pairs provide very strong constraints to the PPM problem. For example in Fig. 2, in addition to frequency matrix F, we may be given that mutations 2 and 5 have been observed on a single read. Thus, in T_1 and T_2 the pair is highlighted in green because it is correctly placed on the same path from the root on the inferred trees. However, the two mutations occur on distinct branches on T_3, which is therefore ruled out as a possible solution.

Single-cell sequencing. With single-cell sequencing, we are able to identify the mutations that are present in a single tumor cell. If in addition to bulk DNA sequencing samples, we are given single cell DNA sequencing data from the same tumor, we can constrain the solution space to PPM considerably. In particular, each single cell imposes that its comprising mutations must correspond to a connected path in the phylogenetic tree. These constraints have been described recently in [16].

For an example of these constraints, consider frequency matrix F described in Fig. 2. In addition to frequency matrix F, we may observe a single cell with mutations $\{1, 2, 3, 5\}$. T_1 is the only potential solution as this is the only tree which places all four mutations on a single path, highlighted in blue. Trees T_2 and T_3 would be ruled out because the mutation set $\{1, 2, 3, 5\}$ does not induce a connected path in these two trees.

We note that the constraints described above for single-cell sequencing and long-read sequencing assume error-free data. In practice, one must incorporate

an error model and adjust the constraints accordingly. However, the underlying principles will remain the same.

3.3 Uniform sampling of solutions

For practical PPM problem instances, the number n of mutations ranges from 10 to 1000. In particular, for solid tumors in adults we typically observe thousands of point mutations in the genome. As such, exhaustive enumeration of solutions is infeasible in practice. To account for non-uniqueness of solutions and to identify common features shared among different solutions, it would be desirable to have an algorithm that samples uniformly from the solution space. However, as the underlying decision problem is NP-complete, the problem of sampling uniformly from the solution space for arbitrary frequency matrices F is NP-hard. Thus, one must resort to heuristic approaches.

One class of such approaches employs Markov chain Monte Carlo (MCMC) for sampling from the solution space [2, 10, 11]. Here, we describe an alternative method based on rejection sampling. This method is guaranteed to sample uniformly from the solution space. Briefly, the idea is to generate a spanning arborescence T from G_F uniformly at random and then test whether T satisfies (SC). In the case where T satisfies (SC), we report T as a solution and otherwise reject T.

For the general case where G_F may have a directed cycle, we use the cycle-popping algorithm of Propp and Wilson [21]. This algorithm generates a uniform spanning arborescence in time $O(\tau(\tilde{G}_F))$ where $\tau(\tilde{G}_F)$ is the expected hitting time of $\tilde{G}_F$. More precisely, $\tilde{G}_F$ is the multi-graph obtained from G_F by including self-loops such that the out-degrees of all its vertices are identical.

For the case where G_F is a DAG with a unique source vertex r, there is a much simpler sampling algorithm. We simply assign each vertex $c \neq r$ to a parent $\pi(c) \in \delta_{\text{in}}(c)$ uniformly at random. It is easy to verify that the resulting function π encodes a spanning arborescence of G_F. Thus, the running time of this procedure is $O(E(G_F))$. In both cases, the probability of success equals the fraction of spanning arborescences of G_F that satisfy (SC) among all spanning arborescences of G_F.

An implementation of the rejection sampling for the case where G_F is a DAG is available on `https://github.com/elkebir-group/OncoLib`.

4 Results

Fig. 1 and Fig. 2 show anecdotal examples of non-uniqueness of solutions to the Perfect Phylogeny Mixture problem. The following questions arise: Is non-uniqueness a widespread phenomenon in PPM instances? Which factors contribute to non-uniqueness and how does information from long-read sequencing and single-cell sequencing reduce non-uniqueness? Finally, are current MCMC methods able to sample uniformly from the space of solutions?

To answer these questions, we used simulated data generated by a previously published tumor simulator [5]. For each number $n \in \{3, 5, 7, 9, 11, 13\}$ of mutations, we generated 10 complete perfect phylogeny trees T^*. The simulator assigned each vertex $v \in V(T^*)$ a frequency $f(v) \geq 0$ such that $\sum_{v \in V(T^*)} f(v) = 1$. For each simulated complete perfect phylogeny tree T^*, we generated $m \in \{1, 2, 5, 10\}$ bulk samples by partitioning the vertex set $V(T^*)$ into m disjoint parts followed by normalizing the frequencies in each sample. This yielded a frequency matrix F for each combination of n and m. In total, we generated $10 \cdot 6 \cdot 4 = 240$ instances (Tables 1–6). The raw data and scripts to generate the results are available on `https://github.com/elkebir-group/PPM-NonUniq`.

4.1 What contributes to non-uniqueness?

The two main factors that influence non-uniqueness are the number n of mutations and the number m of samples taken from the tumor. The former contributes to non-uniqueness while the latter reduces it. As we increased the number n of mutations from 3 to 13, we observed that the number of solutions increased exponentially (Fig. 3A). On the other hand, the number m of samples had an opposing effect: with increasing m the number of solutions decreased.

To understand why we observed these two counteracting effects, we computed the number of spanning arborescences in each ancestry graph G_F. Fig. 3B shows that the number of spanning arborescences exhibited an exponential increase with increasing number n of mutations, whereas increased number m of samples decreased the number of spanning arborescences. The latter can be explained by studying the effect of the number m of samples on the branching coefficient $\gamma(G_F)$. Fig. 3C shows that the branching coefficient increased with increasing m, with branching coefficient $\gamma(G_F) = 0$ for all $m = 1$ instances F. This finding illustrates that additional samples reveal branching of mutations. That is, in the case where $m = 1$ one does not observe branching in G_F, whereas as $m \to \infty$ each sample will be composed of a single cell with binary frequencies and the ancestry graph G_F will be a rooted tree.

Adding mutations increases the complexity of the problem, as reflected by the number of solutions. To quantify how distinct each solution T is to the simulated tree T^*, we computed the edge recall of T defined as $|E(T) \cap E(T^*)|/|E(T^*)|$ (note that $|E(T^*)| = n - 1$ by definition). A recall value of 1 indicates that the inferred tree T is identical to the true tree T^*. Fig. 3D shows that the median recall decreased with increasing number n of mutations. However, as additional samples provide more information, the recall increased with increasing number m of samples.

4.2 How to reduce non-uniqueness?

As discussed in Section 3.2, the non-uniqueness of solutions can be reduced through various sequencing techniques such as single-cell sequencing and long-read sequencing. We considered the effect of both technologies on the $n = 9$ instances (Table 4).

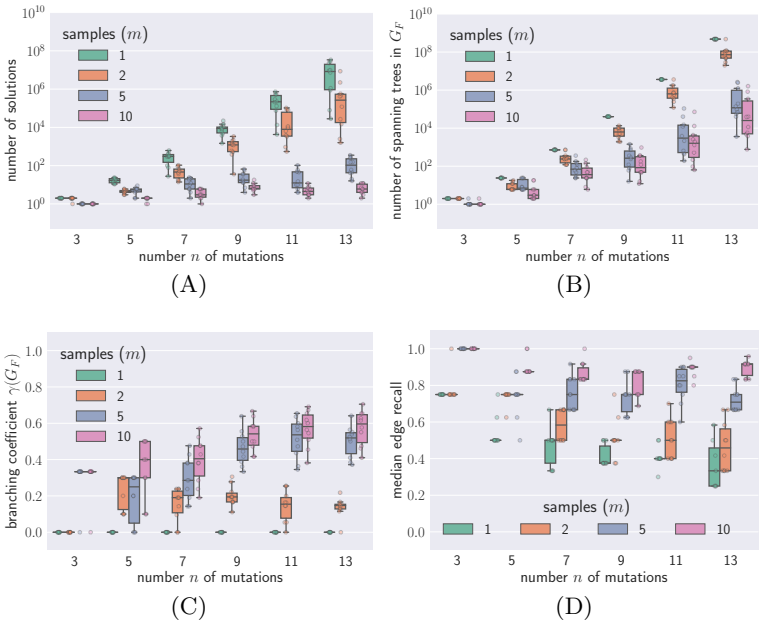

Fig. 3: **Factors that contribute to non-uniqueness.** (A) The number of solutions increased with increasing number n of mutations, but decreased with increasing number m of bulk samples. (B) Every solution of an PPM instance F is a spanning arborescence in the ancestry graph G_F. The number of spanning arborescences in G_F also increased with increasing n and decreased with increasing m. (C) The decrease in the number of solutions and spanning arborescences with increasing m is explained by the branching coefficient of $\gamma(G_F)$, which is the fraction of distinct pairs of mutations that occur on distinct branches in G_F. The fraction of such pairs increased with increasing m. (D) The median edge recall of the inferred trees T increased with increasing m.

By taking longer reads of the genome, long-read sequencing can identify mutations which coexist in a clone if they appear near one another on the genome. If two mutations are observed together on a long read, then one mutation is ancestral to the other. That is, on the true phylogenetic tree T^* there must exist a path from the root to a leaf containing both mutations. We varied the number of mutation pairs observed together from 0 to 5 and observed that increasing this number reduced the size of the solution space (Fig. 4A). In addition, incorporating more simulated long-read information resulted in increased recall of the inferred trees (Fig. 4B).

Single-cell sequencing illuminates all of the mutations present in a single clone in a tumor. This reveals a path from the root of the true phylogenetic tree T^* down to a leaf. Fig. 5A shows the effect that single-cell sequencing has on the size of the solution space. We found that, as we increased the number of

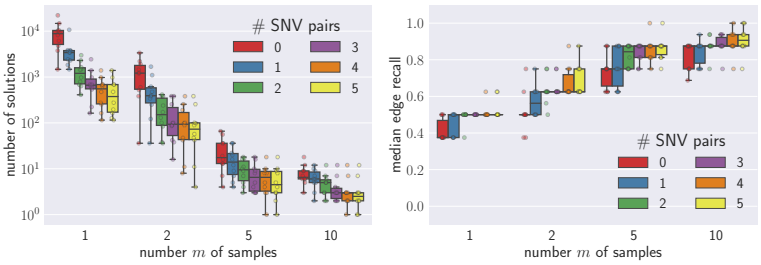

Fig. 4: **Long-read sequencing reduces the size of the solution space.** (A) The number of solutions decreased with increasing pairs of mutations that occurred on the same read. (B) The median edge recall increased with increasing pairs of mutations that co-occur on a read.

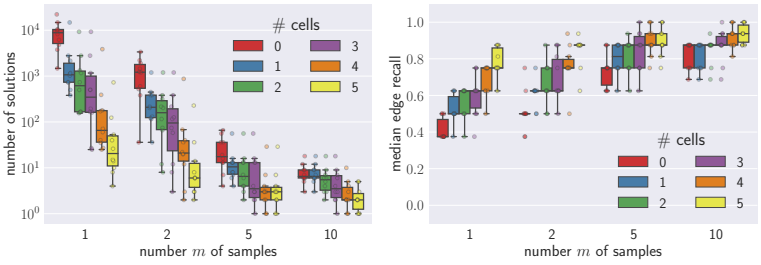

Fig. 5: **Joint bulk and single-cell sequencing reduces the size of the solution space.** (A) The number of solutions decreased with increasing number of single cells. (B) The median edge recall increased with increasing number of single cells.

known paths (sequenced single cells) in the tree from 0 to 5, the solution space decreased exponentially. Additionally, the inferred trees were more accurate with more sequenced cells, as shown in Fig. 5B by the increase in median edge recall. These effects are more pronounced when fewer samples are available.

In summary, while both single-cell and long-read sequencing reduce the extent of non-uniqueness in the solution space, single-cell sequencing achieves a larger reduction than long-read sequencing.

4.3 How does non-uniqueness affect current methods?

To study the effect of non-uniqueness, we considered two current methods, PhyloWGS [2] and Canopy [10], both of which use Markov chain Monte Carlo to sample solutions from the posterior distribution. Rather than operating from

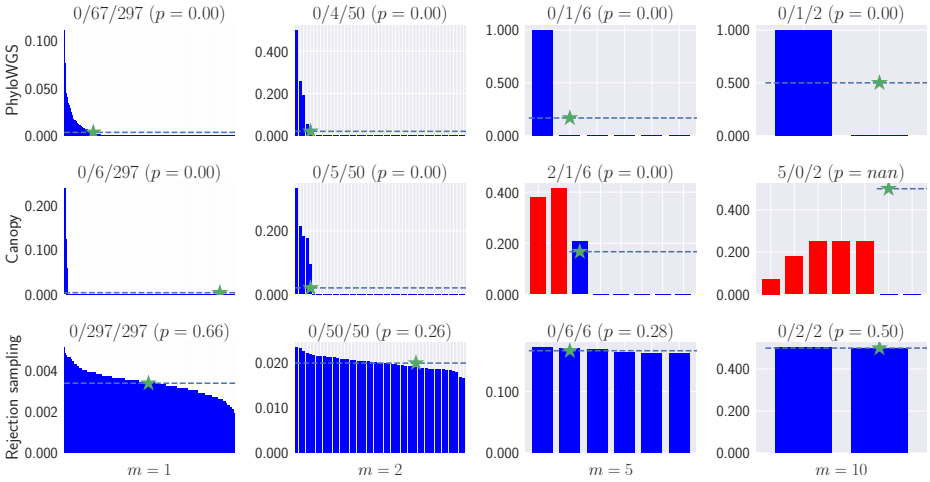

Fig. 6: **PhyloWGS and Canopy do not sample uniformly from the solution space.** We consider an $n = 7$ instance (#81) with varying number $m \in \{1, 2, 5, 10\}$ of bulk samples (columns), from which we sample solutions using different methods (rows). Each plot shows the relative frequency (y-axis) of identical trees (x-axis) output by each method, with the simulated tree indicated by '$\star$'. While blue bars are correct solutions (satisfying (SC)), red bars correspond to incorrect solutions (violating (SC)). Dashed line indicates the expected relative frequency in the case of uniformity. The title of each plot lists the number of incorrect solutions, the number of recovered correct solutions, the total number of correct solutions and the p-value of the chi-squared test of uniformity (null hypothesis is that the samples come from a uniform distribution).

frequencies $F = [f_{p,c}]$, these two methods take as input two integers $a_{p,c}$ and $d_{p,c}$ for each mutation c and sample p. These two integers are, respectively, the number of reads with mutation c and the total number of reads. Given $A = [a_{p,c}]$ and $D = [d_{p,c}]$, PhyloWGS and Canopy aim to infer a frequency matrix $\hat{F}$ and phylogenetic tree T with maximum data likelihood $\Pr(D, A \mid \hat{F})$ such that T satisfies (SC) for matrix $\hat{F}$. In addition, the two methods cluster mutations that are inferred to have similar frequencies across all samples. To use these methods in our error-free setting, where we are given matrix $F = [f_{p,c}]$, we set the total number of reads for each mutation c in each sample p to a large number, i.e. $d_{p,c} = 1,000,000$. The number of variant reads is simply set as $a_{p,c} = f_{p,c} \cdot d_{p,c}$. Since both PhyloWGS and Canopy model variant reads $a_{p,c}$ as draws from a binomial distribution parameterized by $d_{p,c}$ and $\hat{f}_{p,c}$, the data likelihood is maximized when $\hat{F} = F$. We also discard generated solutions where mutations are clustered. Hence, we can use these methods in the error-free case.

We ran PhyloWGS, Canopy, and our rejection sampling method (Section 3.3) on all $n = 7$ instances (Table 3). We used the default settings for PhyloWGS (2500 MCMC samples, burnin of 1000) and Canopy (burnin of 100 and 1 out

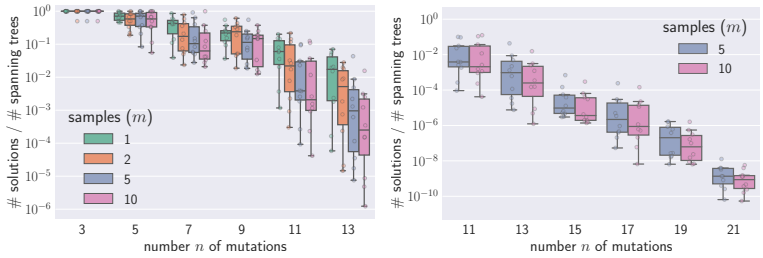

Fig. 7: **Although rejection sampling achieves uniformity, it becomes impractical with increasing number n of mutations.** (A) Plot shows the ratio of the number of solutions to spanning arborescences. Observe that the number of spanning trees increased with the number n of mutations far more rapidly than the number of solutions. (B) With further increases in n, the ratio rapidly decreased and the odds of randomly sampling a solution from the space of spanning arborescences becomes infeasible.

of 5 thinning), with 20 chains per instance for PhyloWGS and 15 chains per instance for Canopy. For each instance, we ran the rejection sampling algorithm until it generated 10,000 solutions that satisfy (SC).

Fig. 6 shows one $n = 7$ instance (#81) with varying number $m \in \{1, 2, 5, 10\}$ of samples. For this instance, all the trees output by PhyloWGS satisfied the sum condition. However, the set of solutions was not sampled uniformly, with only 67 out 297 trees generated for $m = 1$ samples. For $m = 5$, this instance had six unique solutions, with PhyloWGS only outputting trees that corresponded to a single solution among these six solutions (Fig. 9). Similarly, Canopy failed to sample solutions uniformly at random. In addition, Canopy failed to recover any of the two $m = 10$ solutions and recovered incorrect solutions for $m = 5$. The rejection sampling method recovered all solutions for each value of m. In addition, it sampled solutions uniformly at random. Fig. 10, Fig. 11 and Fig. 12 show similar patterns for the other $n = 7$ instances.

Given a frequency matrix F, the success probability of the rejection sampling approach equals the fraction between the number of solutions and the number of spanning arborescences in G_F, as shown empirically in Table 7. As such, this approach does not scale with increasing n. Indeed, Fig. 7A shows that the fraction of spanning trees which also fulfill the sum condition is initially high when the number of mutations is low. With $n = 11$ mutations, the fraction is approximately 10^{-2} and rejection sampling can be considered to be feasible. However, as the number of mutations is increased further, rejection sampling become infeasible as the fraction can drop to 10^{-10} for $n = 21$ mutations (Fig. 7B). Therefore, a better sampling approach is required.

5 Discussion

In this work, we studied the problem of non-uniqueness of solutions to the Perfect Phylogeny Mixture (PPM) problem. In this problem, we are given a frequency matrix F that determines a directed graph G_F called the ancestry graph. The task is to identify a spanning arborescence T of G_F whose internal vertices satisfy a linear inequality whose terms are entries of matrix F. We formulated the #PPM problem of counting the number of solutions to an PPM instance. We showed that the number of solutions is at most the number of spanning arborescences in G_F, a number that can be computed in polynomial time. For the case where G_F is a directed acyclic graph, we gave a simple algorithm for counting the number of spanning arborescences. This algorithm formed the basis of a rejection sampling scheme that samples solutions to a PPM instance uniformly at random.

Using simulations, we showed that the number of solutions increases with increasing number n of mutations but decreases with increasing number m of samples. In addition, we showed that the median recall of all solutions increases with increasing m but decreases with increasing n. We showed how constraints from single-cell and long-read sequencing reduce the number of solutions. Finally, we showed that current MCMC methods fail to sample uniformly from the solution space. This is problematic as it leads to biases that propagate to downstream analyses.

There are a couple of avenues for future research. First, it remains to show that #PPM is #P-complete. Second, while the rejection sampling algorithm achieves uniformity, it does not scale to practical problem instance sizes. Further research is needed to develop sampling algorithms that achieve near-uniformity and have reasonable running time for practical problem instances. Third, in terms of practical applications, the problem of sampling solutions uniformly at random in the case of noisy frequencies must be studied. Fourth, just as single-cell sequencing and long-read sequencing impose constraints on the solution space of PPM, it will be worthwhile to include additional prior knowledge to further constrain the solution space. Finally, the PPM problem and the simulations in this paper assumed error-free data. Further research is needed to study the effect of sequencing, sampling and mapping errors. It is to be expected that the problem of non-uniqueness is further exacerbated with additional sources uncertainty.

Acknowledgements. This research is part of the Blue Waters sustained-petascale computing project, which is supported by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. The authors thank the anonymous referees for insightful comments that have improved the manuscript.

References

1. Amit G Deshwar, Levi Boyles, Jeff Wintersinger, Paul C Boutros, Yee Whye Teh, and Quaid Morris. Abstract B2-59: PhyloSpan: Using multi-mutation reads to resolve subclonal architectures from heterogeneous tumor samples. *Cancer Research*, 75(22 Supplement 2):B2–59–B2–59, November 2015.
2. Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun H Jang, Lincoln Stein, and Quaid Morris. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):35, February 2015.
3. 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, June 2015.
4. 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, July 2016.
5. Mohammed El-Kebir, Gryte Satas, and Benjamin J Raphael. Inferring parsimonious migration histories for metastatic cancers. *Nature Genetics*, 50(5):718–726, May 2018.
6. R Fisher, L Pusztai, and C Swanton. Cancer heterogeneity: implications for targeted therapeutics. *British journal of cancer*, 108(3):479–485, 2013.
7. Harold N. Gabow and Eugene W. Myers. Finding all spanning trees of directed and undirected graphs. *SIAM J. Comput.*, 7(3):280–287, 1978.
8. Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner, Stefan C Dentro, Santiago Gonzalez, Thomas J Mitchell, Yulia Rubanova, Pavana Anur, Daniel Rosebrock, Kaixan Yu, Maxime Tarabichi, Amit Deshwar, Jeff Wintersinger, Kortine Kleinheinz, Ignacio Vázquez-García, Kerstin Haase, Subhajit Sengupta, Geoff Macintyre, Salem Malikic, Nilgun Donmez, Dimitri G Livitz, Marek Cmero, Jonas Demeulemeester, Steve Schumacher, Yu Fan, Xiaotong Yao, Juhee Lee, Matthias Schlesner, Paul C Boutros, David D Bowtell, Hongtu Zhu, Gad Getz, Marcin Imielinski, Rameen Beroukhim, S Cenk Sahinalp, Yuan Ji, Martin Peifer, Florian Markowetz, Ville Mustonen, Ke Yuan, Wenyi Wang, Quaid D Morris, Paul T Spellman, David C Wedge, Peter Van Loo, PCAWG Evolution, Heterogeneity Working Group, and PCAWG network. The evolutionary history of 2,658 cancers. *bioRxiv*, page 161562, July 2017.
9. Mariam Jamal-Hanjani, Gareth A Wilson, Nicholas McGranahan, Nicolai J Birkbak, Thomas B K Watkins, Selvaraju Veeriah, Seema Shafi, Diana H Johnson, Richard Mitter, Rachel Rosenthal, Max Salm, Stuart Horswell, Mickael Escudero, Nik Matthews, Andrew Rowan, Tim Chambers, David A Moore, Samra Turajlic, Hang Xu, Siow Ming Lee, Martin D Forster, Tanya Ahmad, Crispin T Hiley, Christopher Abbosh, Mary Falzon, Elaine Borg, Teresa Marafioti, David Lawrence, Martin Hayward, Shyam Kolvekar, Nikolaos Panagiotopoulos, Sam M Janes, Ricky Thakrar, Asia Ahmed, Fiona Blackhall, Yvonne Summers, Rajesh Shah, Leena Joseph, Anne M Quinn, Phil A Crosbie, Babu Naidu, Gary Middleton, Gerald Langman, Simon Trotter, Marianne Nicolson, Hardy Remmen, Keith Kerr, Mahendran Chetty, Lesley Gomersall, Dean A Fennell, Apostolos Nakas, Sridhar Rathinam, Girija Anand, Sajid Khan, Peter Russell, Veni Ezhil, Babikir Ismail, Melanie Irvin-sellers, Vineet Prakash, Jason F Lester, Malgorzata Kornaszewska, Richard Attanoos, Haydn Adams, Helen Davies, Stefan Dentro, Philippe Taniere, Brendan O'Sullivan, Helen L Lowe, John A Hartley, Natasha Iles, Harriet Bell,

Yenting Ngai, Jacqui A Shaw, Javier Herrero, Zoltan Szallasi, Roland F Schwarz, Aengus Stewart, Sergio A Quezada, John Le Quesne, Peter Van Loo, Caroline Dive, Allan Hackshaw, Charles Swanton, and TRACERx Consortium. Tracking the Evolution of Non-Small-Cell Lung Cancer. *New England Journal of Medicine*, 376(22):2109–2121, June 2017.

10. 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 of the United States of America*, 113(37):E5528–37, September 2016.
11. Wei Jiao, Shankar Vembu, Amit G Deshwar, Lincoln Stein, and Quaid Morris. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*, 15:35, 2014.
12. Cyriac Kandoth, Michael D McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F McMichael, Matthew A Wyczalkowski, Mark D M Leiserson, Christopher a Miller, John S Welch, Matthew J Walter, Michael C Wendl, Timothy J Ley, Richard K Wilson, Benjamin J Raphael, and Li Ding. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339, October 2013.
13. G. Kirchhoff. Ueber die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanischer Ströme geführt wird. *Annalen der Physik*, 148:497–508, 1847.
14. Noam Livne. A note on #P-completeness of NP-witnessing relations. *Information Processing Letters*, 109(5):259–261, February 2009.
15. Marta Łuksza, Nadeem Riaz, Vladimir Makarov, Vinod P Balachandran, Matthew D Hellmann, Alexander Solovyov, Naiyer A Rizvi, Taha Merghoub, Arnold J Levine, Timothy A Chan, Jedd D Wolchok, and Benjamin D Greenbaum. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature*, 551(7681):517, November 2017.
16. Salem Malikic, Katharina Jahn, Jack Kuipers, Cenk Sahinalp, and Niko Beerenwinkel. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *bioRxiv*, page 234914, December 2017.
17. Salem Malikic, Andrew W McPherson, Nilgun Donmez, and Cenk S Sahinalp. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31(9):1349–1356, May 2015.
18. Nicholas McGranahan, Francesco Favero, Elza C de Bruin, Nicolai Juul Birkbak, Zoltan Szallasi, and Charles Swanton. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science Translational Medicine*, 7(283):283ra54–283ra54, April 2015.
19. P C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–8, Oct 1976.
20. 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):91, May 2015.
21. James Gary Propp and David Bruce Wilson. How to Get a Perfectly Random Sample from a Generic Markov Chain and Generate a Random Spanning Tree of a Directed Graph. *Journal of Algorithms*, 27(2):170–217, May 1998.
22. Russell Schwartz and Alejandro A Schäffer. The evolution of tumour phylogenetics: principles and practice. *Nature reviews. Genetics*, 18(4):213–229, April 2017.
23. Francesco Strino, Fabio Parisi, Mariann Micsinai, and Yuval Kluger. Trap: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Res*, 41(17):e165, Sep 2013.

24. Doris P Tabassum and Kornelia Polyak. Tumorigenesis: it takes a village. *Nature Reviews Cancer*, 15(8):473–483, July 2015.
25. Samra Turajlic, Hang Xu, Kevin Litchfield, Andrew Rowan, Tim Chambers, Jose I Lopez, David Nicol, Tim O'Brien, James Larkin, Stuart Horswell, Mark Stares, Lewis Au, Mariam Jamal-Hanjani, Ben Challacombe, Ashish Chandra, Steve Hazell, Claudia Eichler-Jonsson, Aspasia Soultati, Simon Chowdhury, Sarah Rudman, Joanna Lynch, Archana Fernando, Gordon Stamp, Emma Nye, Faiz Jabbar, Lavinia Spain, Sharanpreet Lall, Rosa Guarch, Mary Falzon, Ian Proctor, Lisa Pickering, Martin Gore, Thomas B K Watkins, Sophia Ward, Aengus Stewart, Renzo DiNatale, Maria F Becerra, Ed Reznik, James J Hsieh, Todd A Richmond, George F Mayhew, Samantha M Hill, Catherine D McNally, Carol Jones, Heidi Rosenbaum, Stacey Stanislaw, Daniel L Burgess, Nelson R Alexander, and Charles Swanton. Tracking Cancer Evolution Reveals Constrained Routes to Metastases: TRACERx Renal. *Cell*, 0(0), April 2018.
26. Samra Turajlic, Hang Xu, Kevin Litchfield, Andrew Rowan, Stuart Horswell, Tim Chambers, Tim O'Brien, Jose I Lopez, Thomas B K Watkins, David Nicol, Mark Stares, Ben Challacombe, Steve Hazell, Ashish Chandra, Thomas J Mitchell, Lewis Au, Claudia Eichler-Jonsson, Faiz Jabbar, Aspasia Soultati, Simon Chowdhury, Sarah Rudman, Joanna Lynch, Archana Fernando, Gordon Stamp, Emma Nye, Aengus Stewart, Wei Xing, Jonathan C Smith, Mickael Escudero, Adam Huffman, Nik Matthews, Greg Elgar, Ben Phillimore, Marta Costa, Sharmin Begum, Sophia Ward, Max Salm, Stefan Boeing, Rosalie Fisher, Lavinia Spain, Carolina Navas, Eva Gronroos, Sebastijan Hobor, Sarkhara Sharma, Ismaeel Aurangzeb, Sharanpreet Lall, Alexander Polson, Mary Varia, Catherine Horsfield, Nicos Fotiadis, Lisa Pickering, Roland F Schwarz, Bruno Silva, Javier Herrero, Nick M Luscombe, Mariam Jamal-Hanjani, Rachel Rosenthal, Nicolai J Birkbak, Gareth A Wilson, Orsolya Pipek, Dezso Ribli, Marcin Krzystanek, Istvan Csabai, Zoltan Szallasi, Martin Gore, Nicholas McGranahan, Peter Van Loo, Peter Campbell, James Larkin, and Charles Swanton. Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal. *Cell*, April 2018.
27. W T Tutte. The dissection of equilateral triangles into equilateral triangles. *Mathematical Proceedings of the Cambridge Philosophical Society*, 44(4):463–482, October 1948.
28. Subramanian Venkatesan and Charles Swanton. Tumor Evolutionary Principles: How Intratumor Heterogeneity Influences Cancer Treatment and Outcome. *American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Meeting*, 35:e141–9, 2016.
29. Ke Yuan, Thomas Sakoparnig, Florian Markowetz, and Niko Beerenwinkel. BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome biology*, 16(1):1, 2015.
30. Allen W Zhang, Andrew McPherson, Katy Milne, David R Kroeger, Phineas T Hamilton, Alex Miranda, Tyler Funnell, Nicole Little, Camila P E de Souza, Sonya Laan, Stacey LeDoux, Dawn R Cochrane, Jamie L P Lim, Winnie Yang, Andrew Roth, Maia A Smith, Julie Ho, Kane Tse, Thomas Zeng, Inna Shlafman, Michael R Mayo, Richard Moore, Henrik Failmezger, Andreas Heindl, Yi Kan Wang, Ali Bashashati, Diljot S Grewal, Scott D Brown, Daniel Lai, Adrian N C Wan, Cydney B Nielsen, Curtis Huebner, Basile Tessier-Cloutier, Michael S Anglesio, Alexandre Bouchard-Côté, Yinyin Yuan, Wyeth W Wasserman, C Blake Gilks, Anthony N Karnezis, Samuel Aparicio, Jessica N McAlpine, David G Huntsman, Robert a Holt, Brad H Nelson, and Sohrab P Shah. Interfaces of Malignant and

Immunologic Clonal Dynamics in Ovarian Cancer. *Cell*, 173(7):1755–1769.e22, June 2018.

A Supplementary Results

We have the following the figures and tables in the supplement.

- Fig. 8 illustrates how an ancestry graph is derived from a frequency matrix.
- Fig. 9 shows the six solutions of instance #81 with $n = 7$ mutations and $m = 5$ samples.
- Fig. 10 illustrates the distribution of samples drawn by PhyloWGS for all $n = 7$ instances.
- Fig. 11 illustrates the distribution of samples drawn by Canopy for all $n = 7$ instances.
- Fig. 12 illustrates the distribution of samples drawn by rejection sampling for all $n = 7$ instances.
- Table 1 lists the parameters and results of all instances where $n = 3$.
- Table 2 lists the parameters and results of all instances where $n = 5$.
- Table 3 lists the parameters and results of all instances where $n = 7$.
- Table 4 lists the parameters and results of all instances where $n = 9$.
- Table 5 lists the parameters and results of all instances where $n = 11$.
- Table 6 lists the parameters and results of all instances where $n = 13$.
- Table 7 lists the parameters and results of rejection sampling over all $n = 7$ instances.

#	samples m	solutions	spanning arborescences	ratio	median recall
2	1	2	2	1.000000	0.750
	2	1	2	0.500000	1.000
	5	1	2	0.500000	1.000
	10	1	2	0.500000	1.000
8	1	2	2	1.000000	0.750
	2	2	2	1.000000	0.750
	5	1	1	1.000000	1.000
	10	1	1	1.000000	1.000
12	1	2	2	1.000000	0.750
	2	2	2	1.000000	0.750
	5	1	1	1.000000	1.000
	10	1	1	1.000000	1.000
15	1	2	2	1.000000	0.750
	2	2	2	1.000000	0.750
	5	1	1	1.000000	1.000
	10	1	1	1.000000	1.000
30	1	2	2	1.000000	0.750
	2	2	2	1.000000	0.750
	5	1	1	1.000000	1.000
	10	1	1	1.000000	1.000
39	1	2	2	1.000000	0.750
	2	2	2	1.000000	0.750
	5	1	1	1.000000	1.000
	10	1	1	1.000000	1.000
50	1	2	2	1.000000	0.750
	2	2	2	1.000000	0.750
	5	1	1	1.000000	1.000
	10	1	1	1.000000	1.000
104	1	2	2	1.000000	0.750
	2	2	2	1.000000	0.750
	5	1	1	1.000000	1.000
	10	1	1	1.000000	1.000
119	1	2	2	1.000000	0.750
	2	2	2	1.000000	0.750
	5	1	1	1.000000	1.000
	10	1	1	1.000000	1.000
129	1	2	2	1.000000	0.750
	2	2	2	1.000000	0.750
	5	1	1	1.000000	1.000
	10	1	1	1.000000	1.000

Table 1: **Result for $n = 3$ instances**. From left to right, we list the instance identifier, the number m of samples, the number of solutions, the number of spanning arborescences in the ancestry graph of the instance, the ratio between the solutions and spanning arborescences and the median edge recall.

#	samples m	solutions	spanning arborescences	ratio	median recall
3	1	15	24	0.625000	0.500
	2	6	12	0.500000	0.750
	5	4	6	0.666667	0.625
	10	1	2	0.500000	1.000
5	1	16	24	0.666667	0.500
	2	5	6	0.833333	0.750
	5	5	6	0.833333	0.750
	10	2	2	1.000000	0.875
9	1	21	24	0.875000	0.500
	2	3	6	0.500000	0.750
	5	9	24	0.375000	0.500
	10	2	6	0.333333	0.875
18	1	21	24	0.875000	0.500
	2	5	6	0.833333	0.750
	5	6	8	0.750000	0.750
	10	2	3	0.666667	0.875
37	1	24	24	1.000000	0.500
	2	6	6	1.000000	0.750
	5	6	6	1.000000	0.750
	10	2	2	1.000000	0.875
45	1	12	24	0.500000	0.625
	2	3	16	0.187500	0.750
	5	5	24	0.208333	0.750
	10	2	18	0.111111	0.875
62	1	18	24	0.750000	0.500
	2	5	6	0.833333	0.750
	5	6	6	1.000000	0.750
	10	2	2	1.000000	0.875
66	1	11	24	0.458333	0.500
	2	4	12	0.333333	0.750
	5	2	6	0.333333	0.875
	10	2	6	0.333333	0.875
69	1	22	24	0.916667	0.500
	2	4	6	0.666667	0.625
	5	7	8	0.875000	0.750
	10	2	3	0.666667	0.875
71	1	11	24	0.458333	0.750
	2	4	18	0.222222	0.750
	5	2	24	0.083333	0.875
	10	1	18	0.055556	1.000

Table 2: **Result for $n = 5$ instances**. From left to right, we list the instance identifier, the number m of samples, the number of solutions, the number of spanning arborescences in the ancestry graph of the instance, the ratio between the solutions and spanning arborescences and the median edge recall.

#	samples m	solutions	spanning arborescences	ratio	median recall
7	1	432	720	0.600000	0.500
	2	94	120	0.783333	0.500
	5	24	60	0.400000	0.667
	10	6	24	0.250000	0.833
10	1	28	720	0.038889	0.667
	2	17	720	0.023611	0.667
	5	4	144	0.027778	0.833
	10	3	144	0.020833	0.833
12	1	315	720	0.437500	0.333
	2	43	120	0.358333	0.500
	5	12	80	0.150000	0.750
	10	6	48	0.125000	0.833
23	1	79	720	0.109722	0.500
	2	18	360	0.050000	0.667
	5	10	180	0.055556	0.750
	10	3	90	0.033333	0.833
30	1	293	720	0.406944	0.500
	2	70	120	0.583333	0.667
	5	22	24	0.916667	0.667
	10	6	6	1.000000	0.833
43	1	618	720	0.858333	0.333
	2	54	720	0.075000	0.500
	5	21	360	0.058333	0.667
	10	6	216	0.027778	0.833
49	1	398	720	0.552778	0.333
	2	37	270	0.137037	0.500
	5	2	24	0.083333	0.917
	10	1	24	0.041667	1.000
61	1	328	720	0.455556	0.500
	2	106	240	0.441667	0.500
	5	19	30	0.633333	0.667
	10	3	8	0.375000	0.833
66	1	101	720	0.140278	0.500
	2	14	240	0.058333	0.667
	5	6	120	0.050000	0.833
	10	2	48	0.041667	0.917
81	1	297	720	0.412500	0.500
	2	50	240	0.208333	0.667
	5	6	48	0.125000	0.833
	10	2	24	0.083333	0.917

Table 3: **Result for $n = 7$ instances**. From left to right, we list the instance identifier, the number m of samples, the number of solutions, the number of spanning arborescences in the ancestry graph of the instance, the ratio between the solutions and spanning arborescences and the median edge recall.

#	samples m	solutions	spanning arborescences	ratio	median recall
0	1	1472	40320	0.036508	0.500
	2	36	1920	0.018750	0.750
	5	7	360	0.019444	0.875
	10	5	360	0.013889	0.875
5	1	10445	40320	0.259053	0.375
	2	2200	5040	0.436508	0.500
	5	4	16	0.250000	0.875
	10	3	12	0.250000	0.875
18	1	6180	40320	0.153274	0.500
	2	1450	5040	0.287698	0.500
	5	13	60	0.216667	0.750
	10	9	48	0.187500	0.750
24	1	4776	40320	0.118452	0.375
	2	522	10080	0.051786	0.500
	5	36	1440	0.025000	0.625
	10	12	960	0.012500	0.750
27	1	3755	40320	0.093130	0.500
	2	382	7560	0.050529	0.625
	5	16	864	0.018519	0.750
	10	6	360	0.016667	0.875
31	1	8183	40320	0.202951	0.375
	2	600	13440	0.044643	0.500
	5	19	288	0.065972	0.750
	10	6	180	0.033333	0.875
32	1	14760	40320	0.366071	0.375
	2	1196	3360	0.355952	0.500
	5	56	720	0.077778	0.625
	10	18	120	0.150000	0.688
48	1	9436	40320	0.234028	0.375
	2	1906	10080	0.189087	0.375
	5	36	240	0.150000	0.625
	10	9	48	0.187500	0.750
56	1	10122	40320	0.251042	0.375
	2	1234	2016	0.612103	0.500
	5	66	120	0.550000	0.750
	10	6	16	0.375000	0.875
70	1	22151	40320	0.549380	0.375
	2	3364	10080	0.333730	0.375
	5	13	80	0.162500	0.750
	10	7	48	0.145833	0.875

Table 4: **Result for $n = 9$ instances**. From left to right, we list the instance identifier, the number m of samples, the number of solutions, the number of spanning arborescences in the ancestry graph of the instance, the ratio between the solutions and spanning arborescences and the median edge recall.

#	samples m	solutions	spanning arborescences	ratio	median recall
24	1	74138	3628800	0.020430	0.400
	2	11301	1451520	0.007786	0.400
	5	4	2160	0.001852	0.900
	10	2	1296	0.001543	0.950
27	1	211022	3628800	0.058152	0.400
	2	9456	544320	0.017372	0.500
	5	46	12096	0.003803	0.700
	10	3	2592	0.001157	0.900
35	1	13338	3628800	0.003676	0.500
	2	3350	3628800	0.000923	0.600
	5	10	108000	0.000093	0.850
	10	3	72000	0.000042	0.900
69	1	224451	3628800	0.061853	0.400
	2	3898	120960	0.032226	0.600
	5	15	3840	0.003906	0.800
	10	5	1920	0.002604	0.900
83	1	129706	3628800	0.035743	0.400
	2	936	414720	0.002257	0.600
	5	104	40320	0.002579	0.600
	10	4	4320	0.000926	0.900
89	1	4249	3628800	0.001171	0.500
	2	547	1814400	0.000301	0.700
	5	4	15120	0.000265	0.900
	10	3	12960	0.000231	0.900
109	1	546559	3628800	0.150617	0.400
	2	78547	362880	0.216454	0.500
	5	48	480	0.100000	0.800
	10	7	64	0.109375	0.900
115	1	288866	3628800	0.079604	0.300
	2	6428	241920	0.026571	0.400
	5	6	192	0.031250	0.900
	10	6	512	0.011719	0.900
129	1	522216	3628800	0.143909	0.400
	2	103994	725760	0.143290	0.400
	5	60	640	0.093750	0.750
	10	12	96	0.125000	0.800
139	1	729024	3628800	0.200899	0.400
	2	84747	725760	0.116770	0.400
	5	10	432	0.023148	0.850
	10	8	216	0.037037	0.850

Table 5: **Result for $n = 11$ instances**. From left to right, we list the instance identifier, the number m of samples, the number of solutions, the number of spanning arborescences in the ancestry graph of the instance, the ratio between the solutions and spanning arborescences and the median edge recall.

#	samples m	solutions	spanning arborescences	ratio	median recall
3	1	9863339	479001600	0.020591	0.250
	2	409393	47900160	0.008547	0.417
	5	252	194400	0.001296	0.667
	10	4	11520	0.000347	0.917
12	1	1118667	479001600	0.002335	0.250
	2	14892	19958400	0.000746	0.583
	5	16	138240	0.000116	0.833
	10	6	40320	0.000149	0.917
15	1	867056	479001600	0.001810	0.417
	2	26834	114048000	0.000235	0.500
	5	42	2419200	0.000017	0.750
	10	4	829440	0.000005	0.917
19	1	7318619	479001600	0.015279	0.333
	2	120419	65318400	0.001844	0.500
	5	60	97200	0.000617	0.750
	10	6	38880	0.000154	0.917
25	1	78781	479001600	0.000164	0.500
	2	2488	119750400	0.000021	0.667
	5	44	1244160	0.000035	0.750
	10	3	345600	0.000009	0.917
40	1	9300931	479001600	0.019417	0.250
	2	436744	47900160	0.009118	0.333
	5	352	40320	0.008730	0.667
	10	12	4800	0.002500	0.833
43	1	33809749	479001600	0.070584	NA
	2	575588	29030400	0.019827	0.333
	5	152	69120	0.002199	0.667
	10	7	5760	0.001215	0.917
45	1	28053	479001600	0.000059	0.583
	2	1592	108864000	0.000015	0.667
	5	20	2646000	0.000008	0.833
	10	2	1620000	0.000001	0.958
56	1	23086684	479001600	0.048198	NA
	2	2235187	79833600	0.027998	0.333
	5	280	57600	0.004861	0.667
	10	12	3840	0.003125	0.833
84	1	27236653	479001600	0.056861	NA
	2	8623319	479001600	0.018003	0.333
	5	156	3600	0.043333	0.667
	10	12	768	0.015625	0.833

Table 6: **Result for $n = 13$ instances**. From left to right, we list the instance identifier, the number m of samples, the number of solutions, the number of spanning arborescences in the ancestry graph of the instance, the ratio between the solutions and spanning arborescences and the median edge recall.

#	samples m	solutions	spanning arborescences	solution ratio	trials	success ratio
7	1	432	720	0.600	16585	0.603
	2	94	120	0.783	12753	0.784
	5	24	60	0.400	24821	0.403
	10	6	24	0.250	40859	0.245
10	1	28	720	0.039	256090	0.039
	2	17	720	0.024	419637	0.024
	5	4	144	0.028	358360	0.028
	10	3	144	0.021	481517	0.021
12	1	315	720	0.438	23109	0.433
	2	43	120	0.358	28009	0.357
	5	12	80	0.150	67803	0.147
	10	6	48	0.125	78530	0.127
23	1	79	720	0.110	90828	0.110
	2	18	360	0.050	197369	0.051
	5	10	180	0.056	180518	0.055
	10	3	90	0.033	300223	0.033
30	1	293	720	0.407	24665	0.405
	2	70	120	0.583	17204	0.581
	5	22	24	0.917	10942	0.914
	10	6	6	1.000	10000	1.000
43	1	618	720	0.858	11606	0.862
	2	54	720	0.075	132441	0.076
	5	21	360	0.058	169685	0.059
	10	6	216	0.028	354898	0.028
49	1	398	720	0.553	18115	0.552
	2	37	270	0.137	73073	0.137
	5	2	24	0.083	120731	0.083
	10	1	24	0.042	239816	0.042
61	1	328	720	0.456	21939	0.456
	2	106	240	0.442	22626	0.442
	5	19	30	0.633	15896	0.629
	10	3	8	0.375	26864	0.372
66	1	101	720	0.140	71260	0.140
	2	14	240	0.058	171753	0.058
	5	6	120	0.050	199703	0.050
	10	2	48	0.042	239576	0.042
81	1	297	720	0.412	24528	0.408
	2	50	240	0.208	49137	0.204
	5	6	48	0.125	79423	0.126
	10	2	24	0.083	120821	0.083

Table 7: **Rejection sampling results for $n = 7$ instances.** From left to right, we list the instance identifier, the number m of samples, the number of solutions (satisfying (SC)), the number of spanning arborescences in the ancestry graph of the instance, the ratio between the solutions and spanning trees, the total number of samples (trials) used by the rejection sampling algorithm, the fraction of accepted samples (successful trials). Observe that 'success ratio' $\approx$ 'solution ratio'.

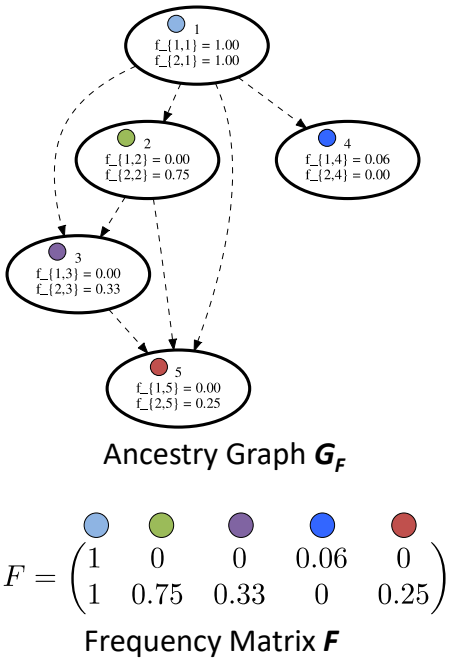

Fig. 8: **Example ancestry graph.** Frequency matrix F corresponds to a simulated $n = 5$ instance (#9) and has $m = 2$ samples. The corresponding ancestry graph G_F illustrates the potential parental relationships.

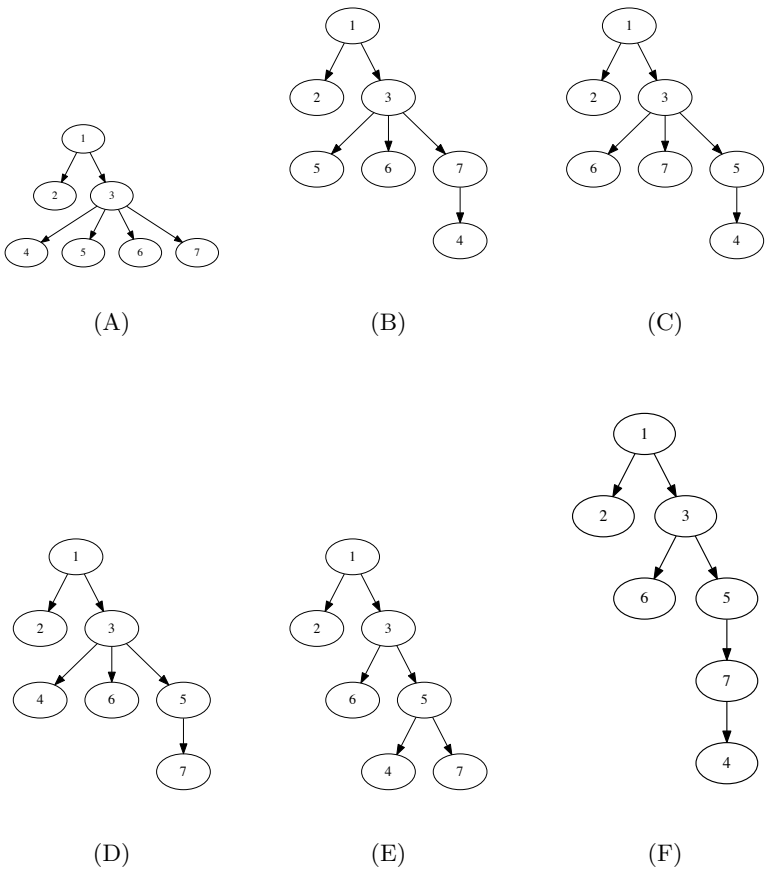

Fig. 9: **Instance #81 with $n = 7$ mutations and $m = 5$ samples has six solutions.** Solution (A) is the true solution. All 7500 samples generated by PhyloWGS correspond to (F). Canopy generated a total of 387 samples corresponding to three different trees. Two out of the three trees were incorrect (307 samples), the remaining 80 samples correspond to (A). Our rejection sampling procedure generated 10000 samples corresponding to each of the six trees in roughly equal proportions.

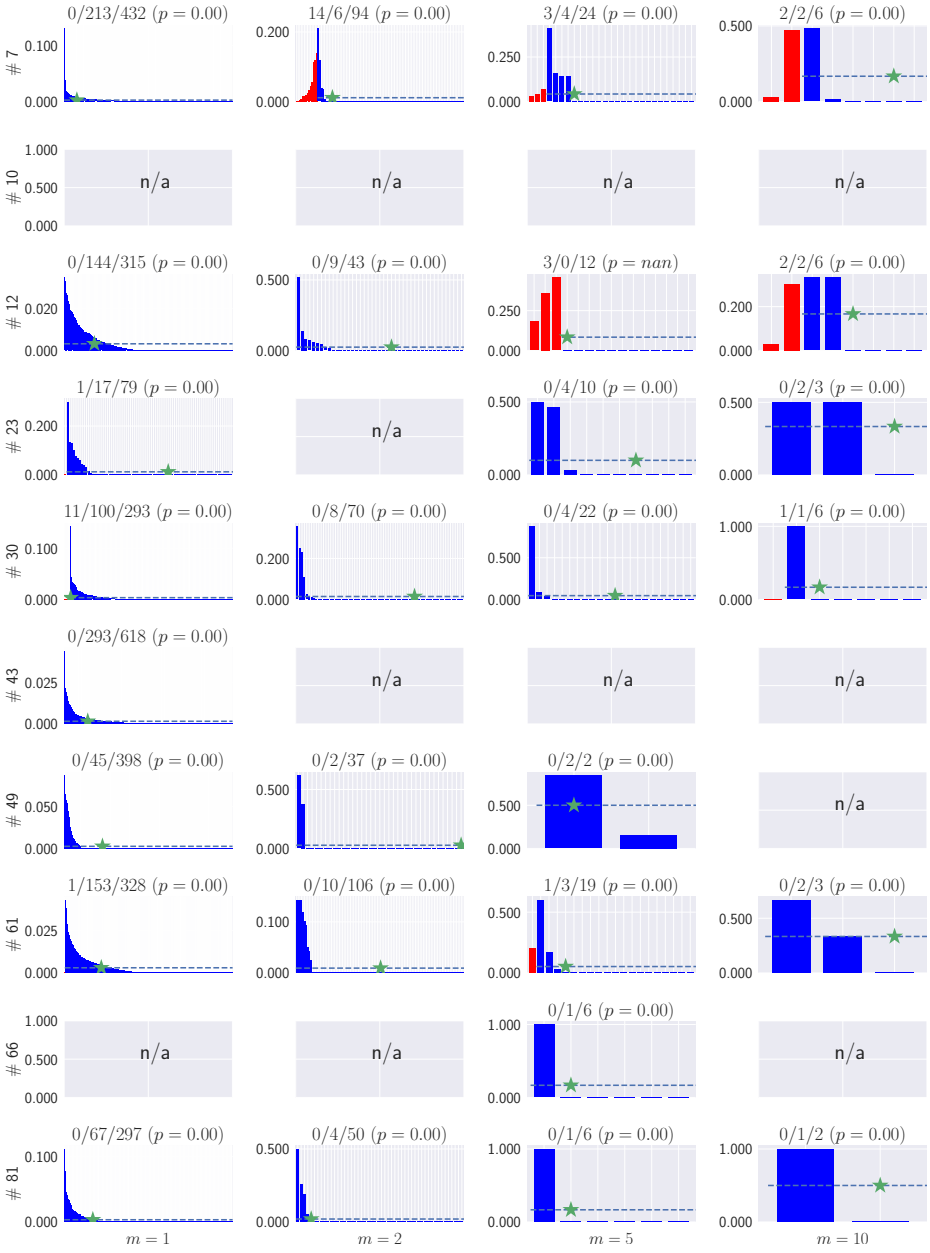

Fig. 10: **PhyloWGS results.** Each plot shows the relative frequency of correct solutions (satisfying (SC)) output by PhyloWGS (blue bars), with the simulated solution indicated by '$\star$'. Red bars correspond to incorrect solutions (violating (SC)). Dashed line indicates the expected relative frequency in the case of uniformity. The title of each plot lists the number of incorrect solutions, the number of recovered correct solutions, the total number of correct solutions and the p-value of the chi-squared test of uniformity. PhyloWGS did not generate any trees without clustered mutations for the instances marked by 'n/a'.

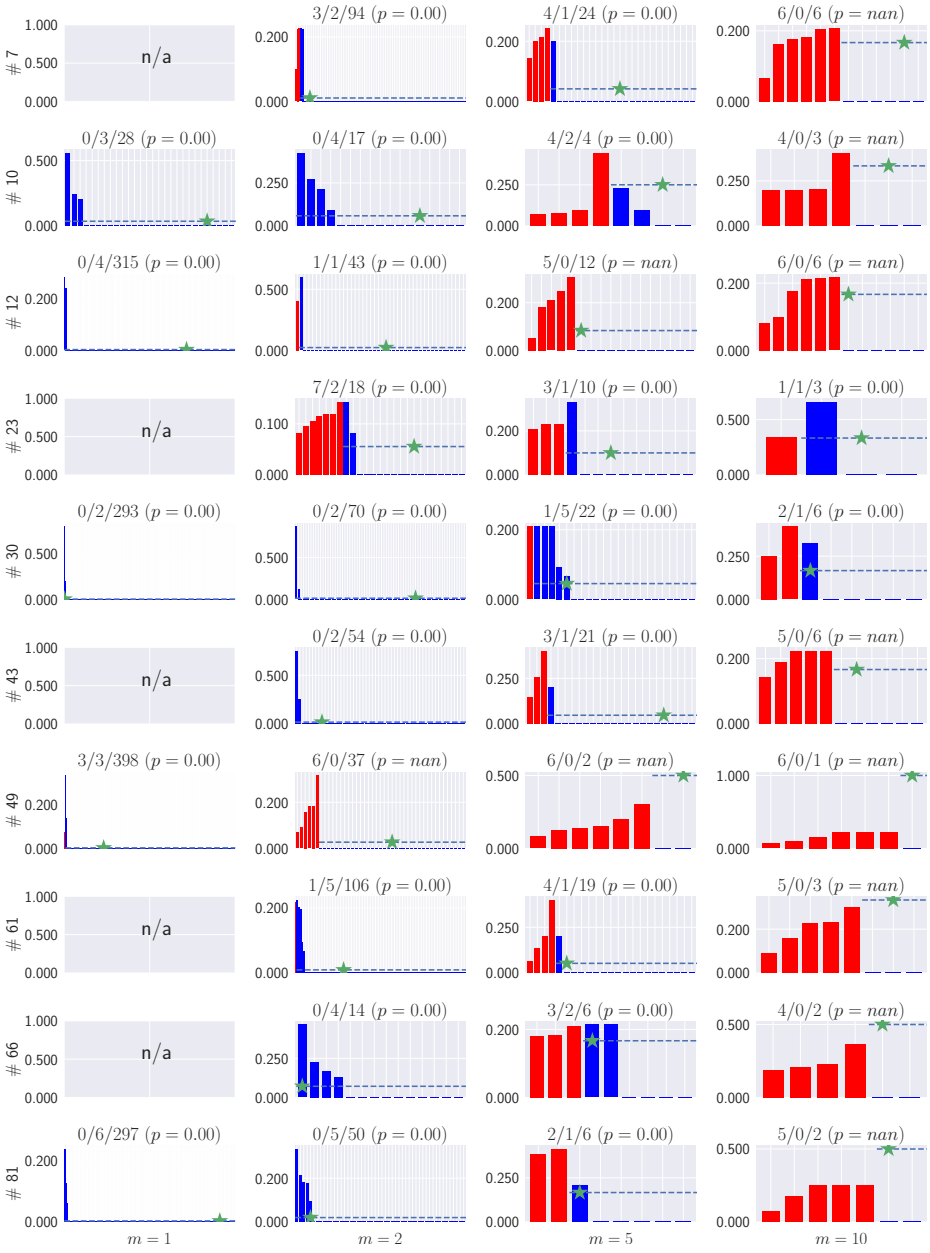

Fig. 11: **Canopy results.** Each plot shows the relative frequency of correct solutions (satisfying (SC)) output by Canopy (blue bars), with the simulated solution indicated by '$\star$'. Red bars correspond to incorrect solutions (violating (SC)). Dashed line indicates the expected relative frequency in the case of uniformity. The title of each plot lists the number of incorrect solutions, the number of recovered correct solutions, the total number of correct solutions and the p-value of the chi-squared test of uniformity. Canopy did not generate any trees without clustered mutations for the instances marked by 'n/a'.

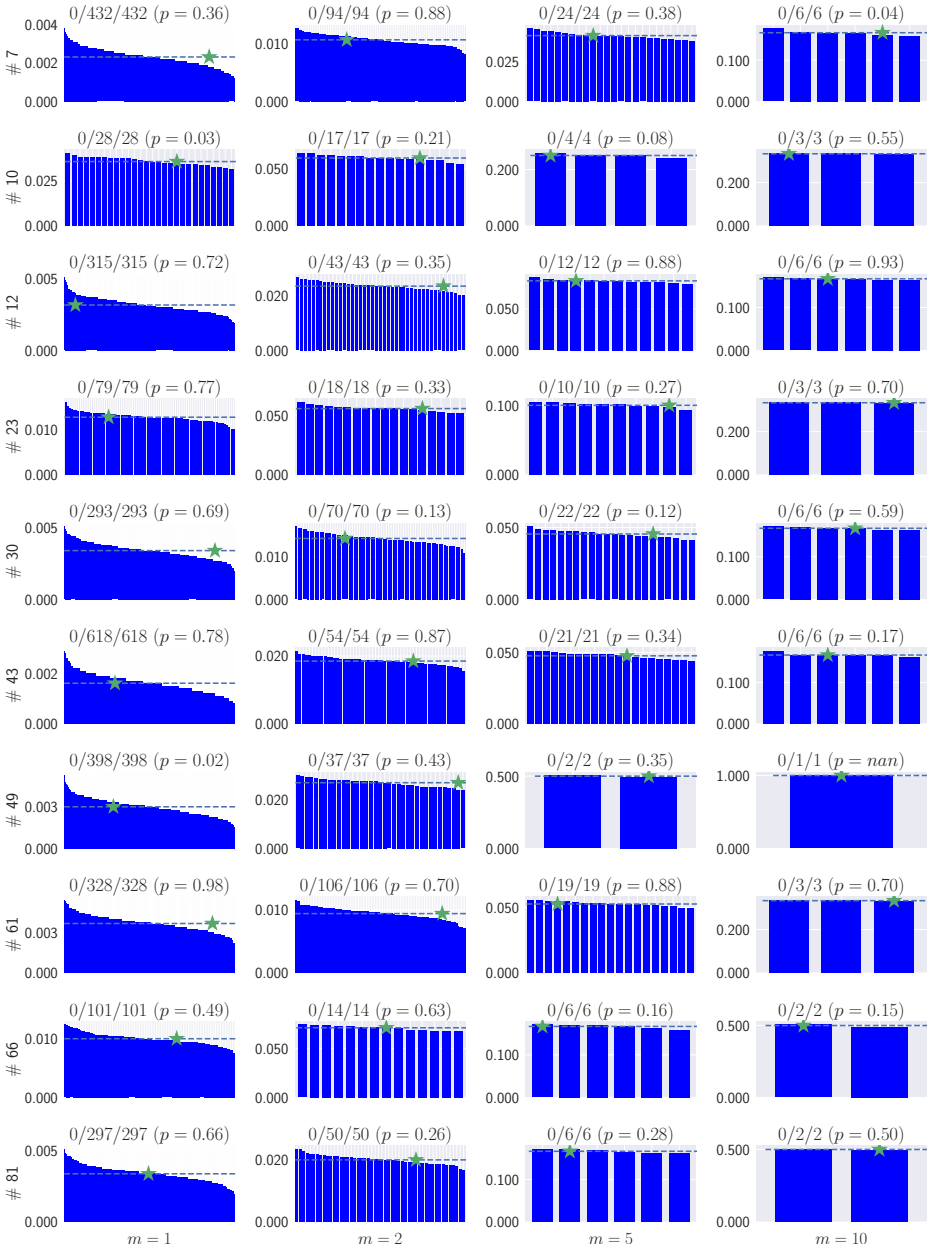

Fig. 12: **Rejection sampling results.** Each plot shows the relative frequency of correct solutions (satisfying (SC)) output by our rejection sampling procedure (blue bars), with the simulated solution indicated by '$\star$'. Red bars correspond to incorrect solutions (violating (SC)). Dashed line indicates the expected relative frequency in the case of uniformity. The title of each plot lists the number of recovered correct solutions, the total number of correct solutions and the p-value of the chi-squared test of uniformity.