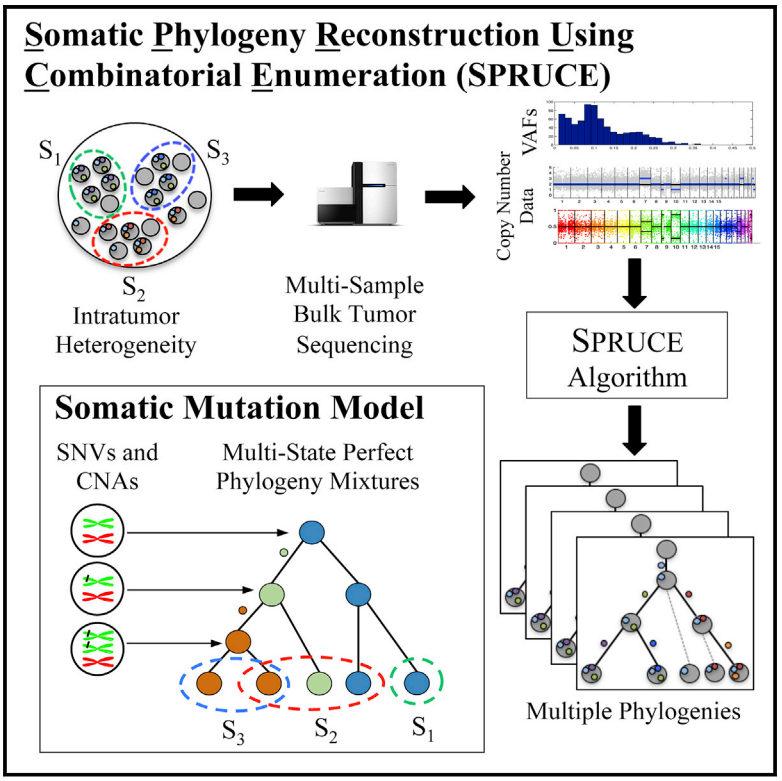
Report

Cell Systems

Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures

Graphical Abstract



Highlights

- Bulk tumor sequencing measures somatic mutations in a population of cells
- Specialized algorithms are required to infer phylogenetic trees from mixtures
- The SPRUCE algorithm constructs phylogenetic trees jointly from SNVs and CNAs
- Typically, many phylogenies are consistent with multi-sample bulk sequencing data

Authors

Mohammed El-Kebir, Gryte Satas, Layla Oesper, Benjamin J. Raphael

Correspondence

braphael@brown.edu

In Brief

Representing mutations in cancer genomes using multiple "states" enables improved analysis of phylogenetic relationships between the populations of cells in a tumor.

El-Kebir et al., 2016, Cell Systems *3*, 43–53
July 27, 2016
http://dx.doi.org/10.1016/j.cels.2016.07.004

Cell Systems
Focus on RECOMB

Report

Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures

Mohammed El-Kebir,[1,3] Gryte Satas,[1,3] Layla Oesper,[1,2] and Benjamin J. Raphael[1,*]
[1]Department of Computer Science and Center for Computational Molecular Biology, Brown University, Providence, RI 02912, USA
[2]Department of Computer Science, Carleton College, Northfield, MN 55057, USA
[3]Co-first author
*Correspondence: braphael@brown.edu
http://dx.doi.org/10.1016/j.cels.2016.07.004

SUMMARY

Phylogenetic techniques are increasingly applied to infer the somatic mutational history of a tumor from DNA sequencing data. However, standard phylogenetic tree reconstruction techniques do not account for the fact that bulk sequencing data measures mutations in a population of cells. We formulate and solve the multi-state perfect phylogeny mixture deconvolution problem of reconstructing a phylogenetic tree given mixtures of its leaves, under the multi-state perfect phylogeny, or infinite alleles model. Our somatic phylogeny reconstruction using combinatorial enumeration (SPRUCE) algorithm uses this model to construct phylogenetic trees jointly from single-nucleotide variants (SNVs) and copy-number aberrations (CNAs). We show that SPRUCE addresses complexities in simultaneous analysis of SNVs and CNAs. In particular, there are often many possible phylogenetic trees consistent with the data, but the ambiguity decreases considerably with an increasing number of samples. These findings have implications for tumor sequencing strategies, suggest caution in drawing strong conclusions based on a single tree reconstruction, and explain difficulties faced by applying existing phylogenetic techniques to tumor sequencing data.

INTRODUCTION

Cancer is an evolutionary process, characterized by the accumulation of somatic mutations in a population of cells (Nowell, 1976). As such, tumors are a heterogeneous mixture of cells with different complements of somatic mutations. Recently, researchers have begun to quantify this intra-tumor heterogeneity by sequencing DNA from one or more samples of a tumor, demonstrating the importance of intra-tumor heterogeneity in cancer treatment (Venkatesan and Swanton, 2016). A simple characterization of intra-tumor heterogeneity classifies mutations as clonal (present in all tumor cells) versus subclonal (present in a subset of tumor cells). However, one can obtain more information about the relationship between tumor clones by applying phylogenetic techniques to reconstruct the evolutionary history of the tumor (Nik-Zainal et al., 2012).

Importantly, the process of clonal evolution in a tumor occurs at the level of single cells. In phylogenetic terminology, the somatic evolutionary process is modeled by a phylogenetic tree, whose leaves correspond to extant entities called taxa and whose edges describe the ancestral relationships among the taxa. The taxa are the individual cells in a tumor. Yet, due to technical and financial constraints, the majority of cancer sequencing projects do not sequence individual cells but rather bulk tumor samples containing thousands to millions of cells. All of the datasets from The Cancer Genome Atlas (TCGA) and nearly all of the datasets from the International Cancer Genome Consortium (ICGC) measure mutations in a single bulk tumor sample. More recently, sequencing of multiple bulk samples from the same tumor has been undertaken (Bolli et al., 2014; Brastianos et al., 2015; Gerlinger et al., 2012; Gundem et al., 2015; Ling et al., 2015; Newburger et al., 2013; Sanborn et al., 2015; Schuh et al., 2012). In some of these studies, researchers have used phylogenetic trees to represent the relationships among these individual samples. Importantly, bulk sequencing data do not reveal the presence/absence of a mutation in an individual cell; rather, the fraction of sequence reads that indicate a mutation provide an estimate of the fraction of cells that contain the mutation. In phylogenetic terminology, we do not measure the individual taxa but rather mixtures of taxa. Thus, proper phylogenetic analysis of bulk cancer sequencing data demands specialized techniques that handle such mixtures.

Since the goal of cancer phylogenetic studies is to understand ancestral relationships among mutations, one uses character-based phylogenetic techniques. In a character-based phylogeny, each of the m taxa is a sequence of n characters, where each character exhibits one of several distinct states. The goal is to find a vertex-labeled tree whose leaves are labeled by the states of each taxon and whose internal vertices are labeled by an ancestral state for each character, such that the resulting tree maximizes an objective function (e.g., maximum parsimony or maximum likelihood) over all such labeled trees. Motivated by bulk cancer sequencing data, we consider a phylogenetic tree mixture problem, where the input is not the set of states for each taxon but rather mixtures of these states. The goal is to explain these given mixtures as a mixing of the leaves of an unknown phylogenetic tree in unknown proportions (Figure 1). Stated another way, given m mixtures of the leaves of a phylogenetic tree, can one reconstruct the tree and the mixing proportions?

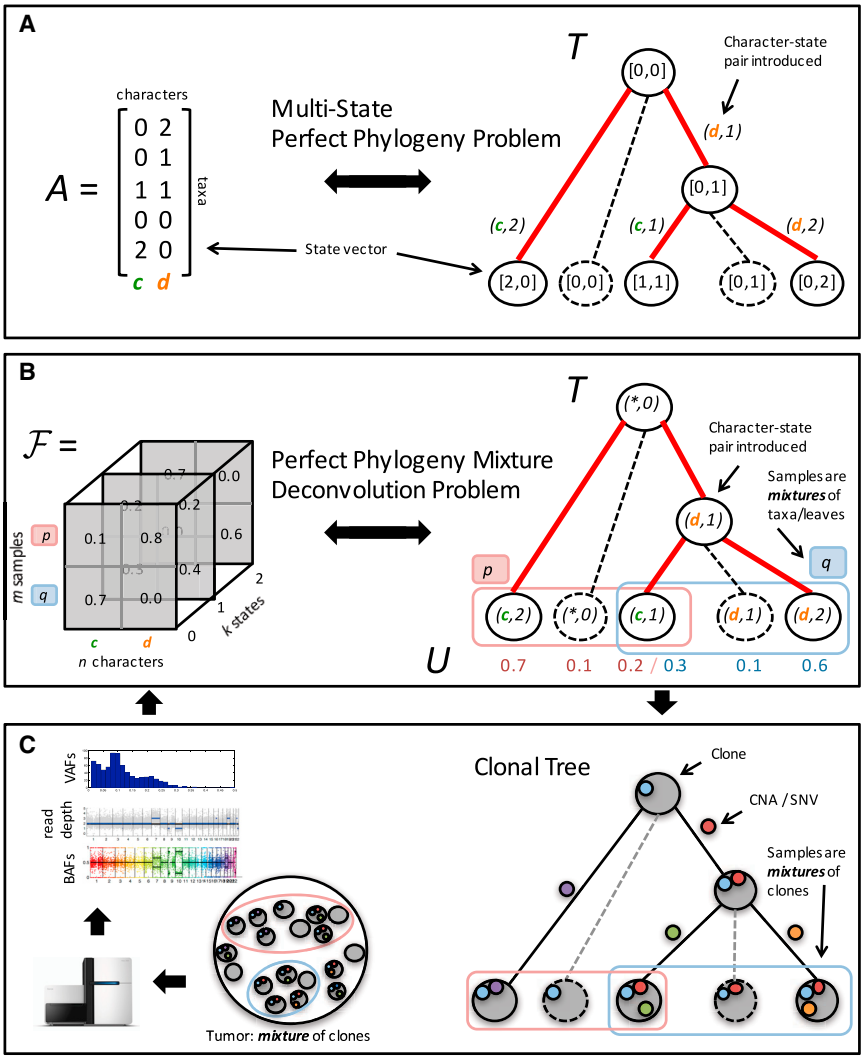


Figure 1. Overview

(A) In the multi-state perfect phylogeny problem, we are given a matrix A whose rows are the state vectors of the taxa. We seek to find a tree T that satisfies the infinite alleles assumption and whose leaves correspond to the taxa.

(B) In the perfect phylogeny mixture deconvolution problem (PPMDP), we do not observe the taxa directly. Instead, our measurements $\mathcal{F}$ correspond to mixtures of the leaves of a tree T according to unknown proportions U. The goal is to infer T and U from the frequencies $\mathcal{F}$ of each character and state in each sample.

(C) Tumors consist of mixtures of distinct cellular populations (clones). Bulk DNA sequencing data of one or more samples of a tumor yields variant allele frequencies (VAFs) of single-nucleotide variants (SNVs) and read-depth ratios and B-allele frequencies (BAFs) from copy number aberrations (CNAs). These measurements are superpositions of the mutations present in the clones in each sample. We show how to derive $\mathcal{F}$ from these measurements; solving the PPMDP yields a clonal tree describing the evolution of the clones in the tumor.

The difficulty of the phylogenetic tree mixture problem depends on the evolutionary model. The simplest such model assumes that characters have only two states and that each character changes state at most once, i.e., the characters evolve with no homoplasy. The latter condition is called the infinite sites assumption and gives rise to a perfect phylogeny. Driven by the application to cancer sequencing, there has been a surge of interest in solving the phylogenetic tree mixture problem for a two-state perfect phylogeny, relying on the idea that the infinite sites assumption is a reasonable model for somatic single-nucleotide variants (SNVs) (Deshwar et al., 2015; El-Kebir et al., 2015; Jiao et al., 2014; Malikic et al., 2015; Nik-Zainal et al., 2012; Popic et al., 2015; Strino et al., 2013; Yuan et al., 2015).

While the two-state perfect phylogeny model might be reasonable for somatic SNVs, it is fairly restrictive for modeling the full somatic mutational process in cancer. For instance, somatic copy number aberrations (CNAs) are ubiquitous in solid tumors (Zack et al., 2013), and these mutations often have more than two states. Thus far, most analysis methods have either excluded genomic regions with CNAs from tree reconstruction or do not address the phylogeny mixture problem. A probabilistic model introduced by Li and Li (2014) considers SNVs and CNAs to infer the genomic composition of tumor populations, but it does so without inferring a phylogenetic tree. In another line of work, Chowdhury et al. (2015) developed rich phylogenetic models for CNAs, but these are for single-cell data. An additional complication is that the observed frequency of an SNV is confounded by CNAs affecting the locus (Eirew et al., 2015). Several authors address this problem by making restrictive assumptions on the mutational process in order to estimate the cancer cell fraction (CCF), the proportion of tumor cells containing an SNV from the observed read counts (Bolli et al., 2014; Gundem et al., 2015; Nik-Zainal et al., 2012). However, these restrictions are not true for all tumors and may result in incomplete reconstructions or rule out alternative valid phylogenies.

An alternative approach is to use a multi-state model for cancer phylogenies. The difficulty of the general multi-state phylogenetic mixture problem is unknown. However, as in the two-state case, we can make the simplifying assumption of no homoplasy. For multi-state characters, the no-homoplasy assumption is referred to as the infinite alleles assumption in population genetics or the multi-state perfect phylogeny (Fernández-Baca, 2001; Gusfield, 2014). In this model, a character may change state more than once on the tree, but it changes *to the same state at most once.*

Here we introduce a perfect phylogeny mixture problem in the case of multi-state characters that evolve without homoplasy under the infinite alleles assumption. We define this problem formally as the multi-state perfect phylogeny mixture deconvolution problem (PPMDP), and we derive a mathematical

characterization of its solutions as a restricted class of spanning trees of a labeled multi-graph. We specialize the PPMDP to analyze bulk sequencing data from multiple samples of a tumor by introducing a multi-state model of SNVs and CNAs. We apply the resulting algorithm, somatic phylogeny reconstruction using combinatorial enumeration (SPRUCE), to simulated and real data. We find that there is considerable ambiguity in the phylogenetic tree with sequencing data from a few samples but that the number of phylogenetic trees decreases dramatically as the number of samples increases. We show that, by explicitly considering the combinatorial structure of the problem, SPRUCE achieves higher recall values on both small and large instances compared to existing methods. On a prostate tumor (Gundem et al., 2015), we infer phylogenetic trees that contain CNAs and SNVs, and we demonstrate the importance of jointly inferring CCFs and phylogenetic trees.

RESULTS

We formulate the PPMDP, a general problem of constructing phylogenetic trees from mixtures, under the infinite alleles model. We show how measurements of SNVs and CNAs in bulk tumor sequencing data yield instances of the PPMDP, which we solve using our algorithm SRUCE. Next we demonstrate the performance of SPRUCE on simulated and real cancer sequencing data. Throughout this section, taxa correspond to distinct tumor populations, characters correspond to genomic loci, and the state of a character encodes the mutational state and copy number of the corresponding locus.

Phylogenetic Mixture Problems

A character-based phylogeny is a rooted tree whose leaves correspond to m taxa and whose vertices are labeled by state vectors, assigning one of k states to each of the n characters. We represent the taxa by a matrix A whose rows are the state vectors of each taxon. In the character-based phylogeny reconstruction problem, we are given A and wish to infer a phylogenetic tree T, whose leaves are labeled by the rows of A and whose internal vertices are labeled by state vectors, such that T maximizes an objective function, e.g., parsimony or likelihood (Figure 1A). However, in the case of bulk sequencing data, we do not measure A, or equivalently the leaves of T, directly but rather we are given m samples that are mixtures of the rows (taxa) of A, or equivalently mixtures of the leaves of T, in unknown proportions (Figure 1B). Formally, for each sample p, character c, and state $i \in \{0, \ldots, k-1\}$, we are given the frequency $f_{p,(c,i)}$, which is the proportion of taxa in sample p that have state i for character c. Note that frequencies $f_{p,(c,i)}$ are non-negative and sum up to 1 across all states. We represent our input measurements by a $k \times m \times n$ frequency tensor $\mathcal{F} = [[f_{p,(c,i)}]]$.

We denote the root vertex of T by $v_{(*,0)}$ and the remaining vertices by $v_{(c,i)}$, where i is the new state of character c that occurs on the incoming edge. We assume without loss of generality that $\mathcal{F}$ is obtained by mixing all the vertices of T instead of just the leaves (Supplemental Experimental Procedures; Figure S1A). For each vertex $v_{(c,i)}$ and each sample p, we define the usage $u_{p,(c,i)}$ to be the proportion of $v_{(c,i)}$ in sample p. The usages $u_{p,(c,i)}$ are non-negative and the usages of all taxa in a sample sum to 1. Note that $f_{p,(c,i)}$ denotes the frequency of character-state pair (c,i) in sample p, whereas $u_{p,(c,i)}$ denotes the proportion of the taxon $v_{(c,i)}$ in sample p. The usage matrix $U = [u_{p,(c,i)}]$ represents the usages of all taxa across all samples. It follows that the frequency of a character-state pair is the sum of the usages of taxa that have state i for character c. That is, $f_{p,(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}$, where $T_{(c,i)}$ is the set of vertices whose corresponding state vectors have state i for character c. Thus, given a tree T and usage matrix U, the frequencies $f_{p,(c,i)}$ for each sample, character, and state are determined. Our problem, the phylogeny mixture deconvolution problem (PMDP), shows that is the converse.

Phylogeny Mixture Deconvolution Problem (PMDP). Given frequencies $f_{p,(c,i)}$ for character-state pairs (c,i) and samples p, find a phylogenetic tree T and a usage matrix $U = [u_{p,(c,i)}]$ such that $f_{p,(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}$.

In the Experimental Procedures, we derive a mathematical characterization for a special case of the above problem called the PPMDP, whose solutions are perfect phylogenies satisfying the infinite alleles assumption. That is, a character may change state multiple times on the tree but changes to the same state at most once. We show that solutions correspond to constrained spanning trees in a directed multi-graph $G_{\mathcal{F}}$ derived from $\mathcal{F}$. We say that a tree T generates $\mathcal{F}$ if there exists a usage matrix U, such that $f_{p,(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}$.

For our application to cancer genomes described below, we consider a special case of the PPMDP, the Cladistic-PPMDP, where, in addition to the frequencies $\mathcal{F}$, we are given a set $\mathcal{S} = \{S_c \mid c \in [n]\}$ of state trees, one for each character. The state trees describe the order of the states of a character; such an order-constrained character is called cladistic (Fernández-Baca, 2001). Formally, the vertex set of a state tree S_c is the set $\{0, \ldots, k-1\}$ of states, and the edges describe the ancestral relationships of the states for character c. A perfect phylogeny tree T is consistent with $\mathcal{S}$ if the states for each character c occur in the same order as prescribed by the state tree $S_c \in \mathcal{S}$. That is, for all characters c and states i, j, (c,i) is ancestral to (c,j) in T if and only if i is ancestral to j in S_c. In the Cladistic-PPMDP, we seek to find a complete perfect phylogeny tree T that generates $\mathcal{F}$ and is consistent with $\mathcal{S}$.

SPRUCE Infers Tumor Phylogenies from SNVs and CNAs

Our motivating application for the PPMDP is to analyze multi-sample cancer sequencing data. Here m distinct samples from a tumor have been sequenced, each such sample being a mixture of cells with different somatic mutations. Our aim is to reconstruct the phylogenetic tree describing the evolutionary relationships among subpopulations of cells (or clones) within the tumor (Figure 1C). Earlier work on this problem (El-Kebir et al., 2015; Hajirasouliha et al., 2014; Jiao et al., 2014; Malikic et al., 2015; Popic et al., 2015) has focused on the problem of SNVs that have changed at most once in the progression from normal to tumor, and thus the phylogenetic tree is a two-state perfect phylogeny. Other approaches have considered SNVs in regions of CNAs by scaling variant allele frequencies (VAFs) into CCFs, which is the proportion of tumor cells in a sample that have the variant. This quantity has been broadly used in analyzing intra-tumor heterogeneity (Andor et al., 2016; Bolli et al., 2014; Gundem et al., 2015; McGranahan et al., 2015). Here we consider CNAs using multi-state characters.

More specifically, the characters in our model are positions in the genome whose states we model with a triple (x, y, z) of integers. Here, x and y are the numbers of maternal and paternal copies of the position, respectively, and z is the number of mutated copies. We call (x,y) the copy number state. In general, we lack information to phase mutations to the maternal or paternal chromosome. Thus, we assume without loss of generality that $x \geq y$. For a character c and sample p, we cannot directly derive the frequencies $f_{p,(c,i)}$ for states (x_i, y_i, z_i) from DNA sequencing data. Instead, alignment of DNA sequence reads from a tumor and matched normal genome gives us measurements of the following three quantities: VAFs for SNVs, which is the proportion of sequence reads that contain the mutant allele; B-allele frequencies for germline SNPs, which reveal loss of heterozygosity events; and read-depth ratios for larger regions, which are the total number of reads of the region in the tumor sample divided by the total number of reads in the same region of a matched normal.

We show in the Experimental Procedures that, with a few simplifying assumptions on the allowed state transitions, we can derive frequencies $\mathcal{F} = [[f_{p,(c,i)}]]$ and the state trees $\mathcal{S}$ for each character directly from DNA sequencing data. There may be several compatible state trees, which has implications for calculating the CCF. Formally, given a sample S with purity ρ (proportion of tumor cells in the sample), the CCF of an SNV c is given by

$$CCF(p,c) = \frac{1}{\rho} \sum_{(x,y,z),z>0} f_{p,(c,(x,y,z))}. \qquad \text{(Equation 1)}$$

The CCF is not uniquely determined from the VAF and the copy number mixing proportions (Figure S2). To overcome this ambiguity, previous work (Bolli et al., 2014; Gundem et al., 2015) makes the simplifying assumption that all tumor cells have a fixed number n_{chr} of copies carrying the SNV. The quantity n_{chr} is computed as the maximum likelihood estimate under a binomial model, resulting in a single CCF choice for each SNV. In contrast, we do not fix the number of mutated copies in distinct tumor populations. By considering all compatible state trees, we allow for multiple alternative explanations with possibly distinct CCFs.

The SPRUCE algorithm (Figure 2) takes as input VAF confidence intervals and copy number mixing proportions for each character per sample, and it outputs multi-state perfect phylogeny trees. The algorithm consists of the following four steps: (1) computing a collection of compatible state trees $\{S_c\}$ for each character c; (2) deriving a compatibility graph whose edges are pairs (S_c, S_d) of pairwise compatible state trees for pairs (c,d) of characters; (3) computing the set of maximal cliques in the compatibility graph; each maximal clique corresponds to an instance $(\mathcal{F}, \mathcal{S})$ of the Cladistic-PPMDP, where $\mathcal{S}$ is the set of state trees encoded by the vertices of the clique and $\mathcal{F}$ is the corresponding frequency tensor; and (4) solving the instance by enumerating all multi-state perfect phylogeny trees on the largest subset of characters. We detail these steps in the Experimental Procedures. SPRUCE is available for download at http://compbio.cs.brown.edu/projects/spruce and as Data S1.

Experimental Evaluation on Simulated and Real Data

We first applied SPRUCE to simulated data, where we analyzed its performance, studied the effect of violations of the infinite alleles assumption, and compared against PhyloWGS (Deshwar et al., 2015). We then analyzed a prostate cancer tumor from Gundem et al. (2015).

Simulated Data

We constructed simulated instances by randomly generating a tree on $n = 5$ characters, from which we obtained $m \in \{2, 5, 10\}$ samples. This process yields VAFs and copy number mixing proportions for each character in each sample. For each value of $m \in \{2, 5, 10\}$, we generated 20 instances with different simulated tree topologies, resulting in a total of 60 instances. In addition to considering error-free VAFs, we introduced noise by drawing variant reads under a binomial model with sequencing coverage values ranging from 50x to 10,000x. We refer the reader to the Supplemental Experimental Procedures for more details on the generation of simulated data.

We ran SPRUCE on each instance and recorded trees that contained the largest subset of characters. Figure 3A shows the number of solutions of the PPMDP, i.e., the number of trees for all 60 instances, as the sequencing coverage is varied. Not surprisingly, we see that increased coverage results in a decrease in the number of solutions, as the uncertainty in the VAFs of each SNV decreases as coverage increases. In fact, a coverage of 10,000x approaches the error-free data. Correspondingly, we see a drop in running time (Figure S3A). However, we found that a more efficient way to deal with ambiguity (and running time) is to increase the number of samples. For instance, a target coverage of 50x with $m = 5$ samples has a similar number of solutions as a coverage of 10,000x with $m = 2$ samples. This demonstrates how the combinatorial constraints (specifically the multi-state sum condition [MSSC] in the Experimental Procedures) become stronger with increasing numbers of samples and constrain the number of possible trees. We note that, in the case of error-free VAFs, the true tree is always contained in the solution space (Figure 3B shows one such solution space, visualized as a graph containing all trees that are solutions to the PPMDP).

Without further prior knowledge about tree topologies, all enumerated trees are equally likely to be the true tree under a uniform prior. We define the representative tree as the tree whose edges occur most frequently across all enumerated trees. Thus, in expectation, the representative tree has the largest set of correct parental relationships given the uniform prior. The recall of a reported tree T is the fraction of edges in the simulated tree T^* that are recovered in T. Note that the precision of a tree T, which is the fraction of edges in the tree T that occur in T^*, is a constant multiple of the recall in this case. We thus only report recall values. The recall of the representative tree increases with increasing number of samples and coverage (Figure S3B), and it is higher than the median recall value (Figure S3C).

Not surprisingly, the recall of the representative tree decreases with an increasing number of violations of the infinite alleles assumption (Figure S3D). However, for small numbers of violations, recall remains high. In moderately rearranged tumors with few or no overlapping CNAs, the number of violations is likely to be small.

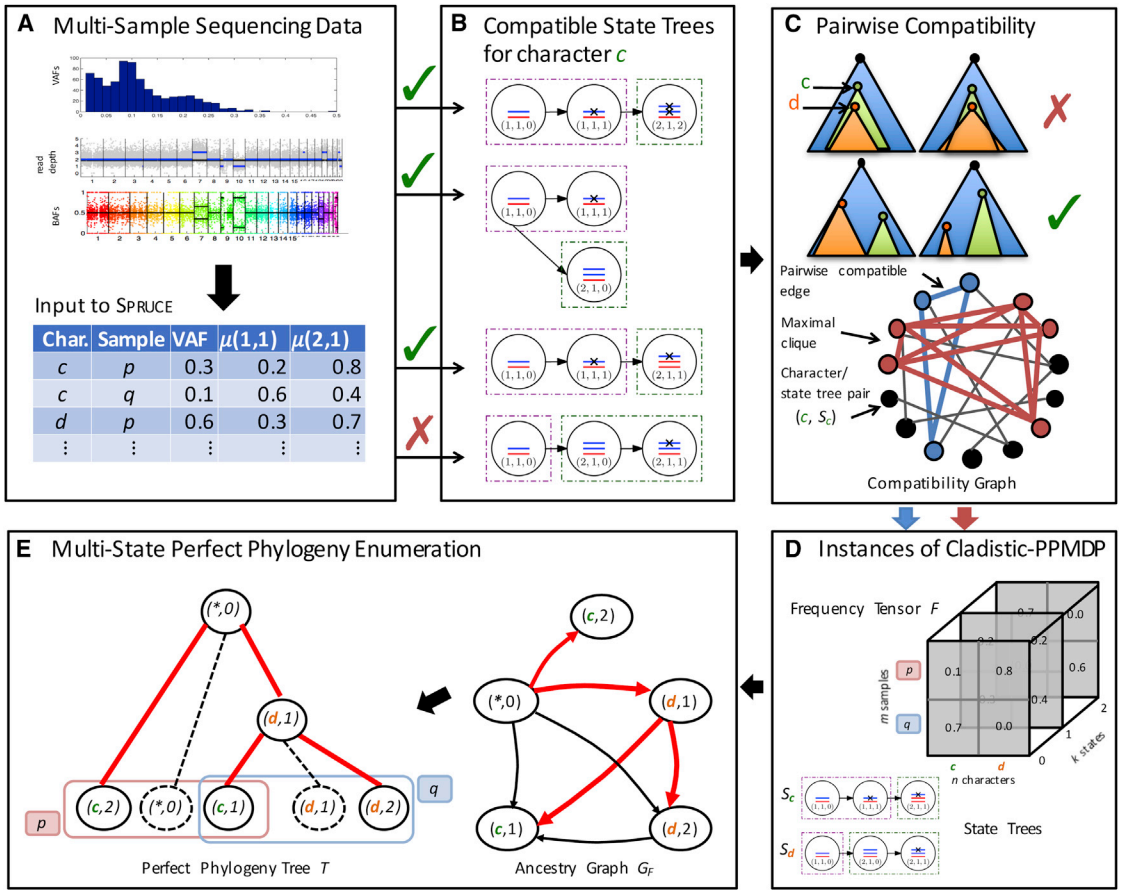


Figure 2. Overview of SPRUCE

(A) Input are the VAFs of SNVs and the copy numbers and mixing proportions of CNAs, which are derived from read depth and B-allele frequencies.

(B) Combining the input with our multi-state model for the somatic mutational process produces a collection of compatible state trees for each character (Figure S2).

(C) Two character-state tree pairs are compatible if there exists a perfect phylogeny tree that contains both. We construct the pairwise compatibility graph by considering all such pairs.

(D) A maximal clique in the compatibility graph yields a frequency tensor $\mathcal{F}$ and collection $\mathcal{S}$ of state trees that are an input to the Cladistic-PPMDP.

(E) For each instance, we construct the multi-state ancestry graph $G_{\mathcal{F}}$, which encodes potential ancestral relationships between character-state pairs. We then enumerate all multi-state perfect phylogeny trees with maximum size in this graph, and we compute the corresponding usage matrices U.

As the number of characters increases, exhaustive enumeration of the space of solutions becomes infeasible. Thus, we analyzed how well we reconstructed the true tree by limiting the running time for each combination of compatible state trees to N seconds. We considered 20 simulated instances with $n = 15$ characters, $m = 10$ samples, and a target coverage of 1,000x. Figure 3D shows that the recall of the representative tree increases with increasing values of N but that the gains are modest. Thus, SPRUCE obtains good solutions within a reasonable amount of computation time.

Comparison to Existing Methods

We compared our method against PhyloWGS (Deshwar et al., 2015), which to our knowledge is currently the only method that jointly infers a phylogenetic tree on SNVs and CNAs from bulk sequencing data. Unlike our multi-state model, PhyloWGS uses a two-state perfect phylogeny model where characters correspond to either CNA events or SNV events. The vertices of the returned trees contain clusters of the two-state characters. To compare tree output by the two methods, we remapped the multi-state characters of the simulated trees to two-state characters as used by PhyloWGS. We used an adjusted measure of recall that accounts for clustering (described in the Supplemental Experimental Procedures). We ran PhyloWGS with default parameters on the 60 simulated $n = 5$ instances and 20 simulated $n = 15$ instances described above, and we compared the recall of the representative tree found by our method to the maximum likelihood tree reported by PhyloWGS. Our method achieved higher recall across varying numbers of samples $m \in \{2, 5, 10\}$, on error-free VAFs with $n = 5$ characters (Figure 3C) and on noisy VAFs with $n = 15$ characters (Figure 3D). We believe that the lower recall values of PhyloWGS, especially those from the larger $n = 15$ instances, indicate difficulties in Markov chain Monte Carlo (MCMC) sampling from the complicated space of possible trees. In contrast, because SPRUCE explicitly models the combinatorial structure of the problem, it is able to effectively find good solutions.

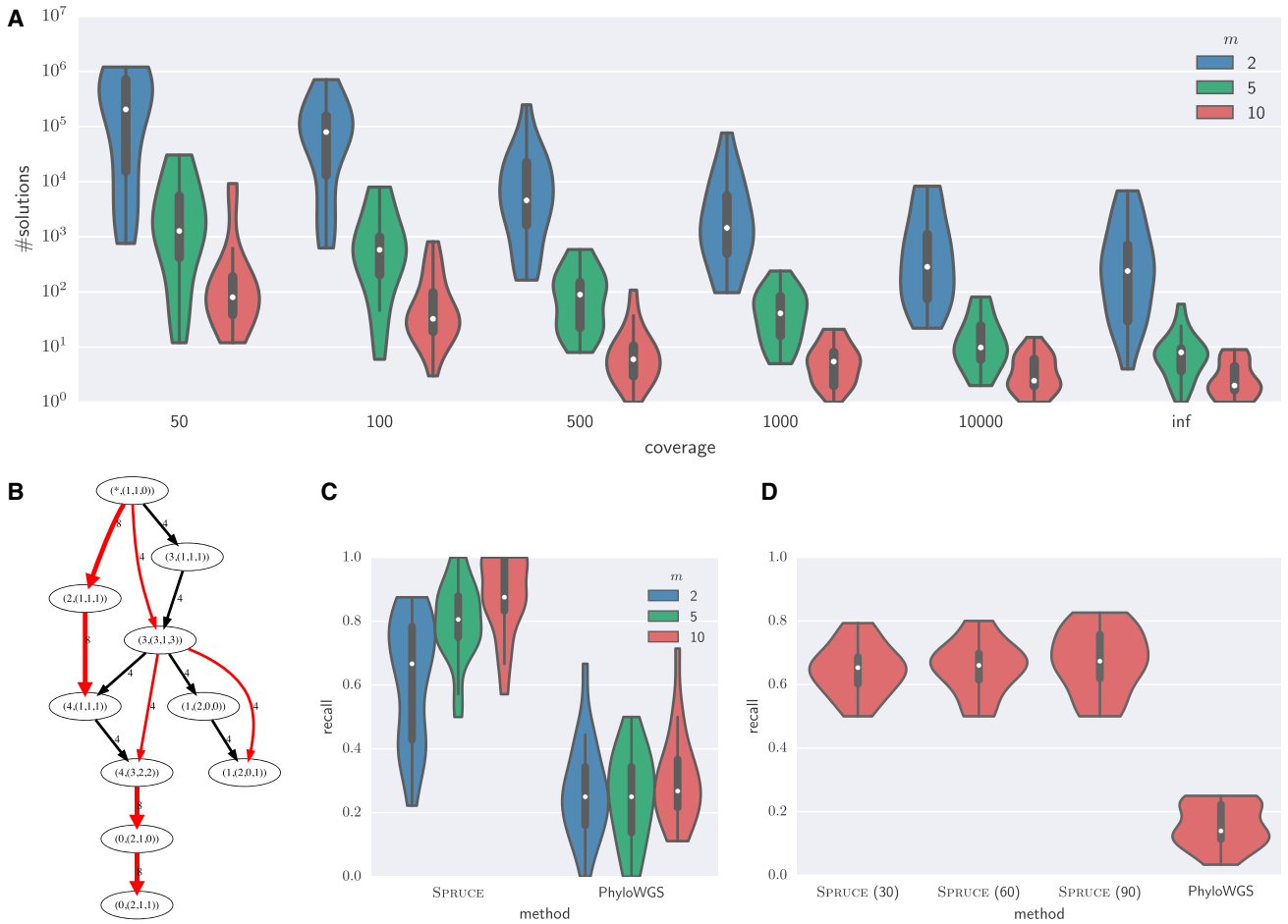


Figure 3. Simulation Results

(A) Number of solutions (log scale) for instances with $n = 5$ characters, $m \in \{2, 5, 10\}$ samples, and varying sequencing coverages. A coverage of inf corresponds to error-free VAFs.

(B) Solution space of an $n = 5$, $m = 5$ instance with error-free VAFs. Vertices are from the solution trees, and each edge is labeled by the number of trees in which it occurs. Red edges indicate the simulated tree.

(C) Recall of SPRUCE's representative trees is compared to PhyloWGS's maximum likelihood trees on 60 instances with $n = 5$ and error-free VAFs (shown in Figure 2A).

(D) Recall values on 20 instances with noisy VAFs (1,000x), $n = 5$, and $m = 10$. We run SPRUCE with varying running time limit (N) per state tree combination in seconds.

Results on Prostate Cancer

Next we analyzed a solid prostate cancer tumor (sample A22 from Gundem et al., 2015) where 11 samples were sequenced, including samples from the primary tumor, multiple metastases, and a matched normal. These samples underwent whole-genome sequencing followed by targeted sequencing resulting in a set of 114 SNVs. The authors constructed a phylogenetic tree for this sample as follows: (1) they assumed that all tumor cells have a fixed number of mutated copies of each variant and rescaled the VAF of each SNV to a CCF accordingly, (2) they clustered SNVs according to the calculated CCFs using a Dirichlet process mixture model, (3) they manually applied the Pigeon-Hole Principle (Nik-Zainal et al., 2012) (a special case of the sum condition described in the Experimental Procedures) to obtain the reported tree (Figure S4A), and (4) they manually placed CNAs on the tree.

We applied SRUCE to the same dataset. We considered only SNVs on autosomes that had a VAF greater than 0.05 in at least one sample. This resulted in a set of 89 SNVs. We used the copy number calls and mixing proportions reported in Gundem et al. (2015), which were obtained using the Battenberg algorithm (Nik-Zainal et al., 2012). For each SNV in each sample, we obtained 99.9% confidence interval on the VAF using a binomial model. Thirteen characters admitted no compatible state tree (Figure S4B). This may have been a result of violations of the infinite alleles model or errors in the input data. We then clustered SNVs that were compatible with a single, shared state tree and whose VAF confidence intervals overlapped in all samples. This yielded 68 characters, four of which corresponded to four clusters containing a total of 12 SNVs (Table S1). The compatibility graph (see the Experimental Procedures) consisted of 5,183 edges, which encode pairwise compatibility between

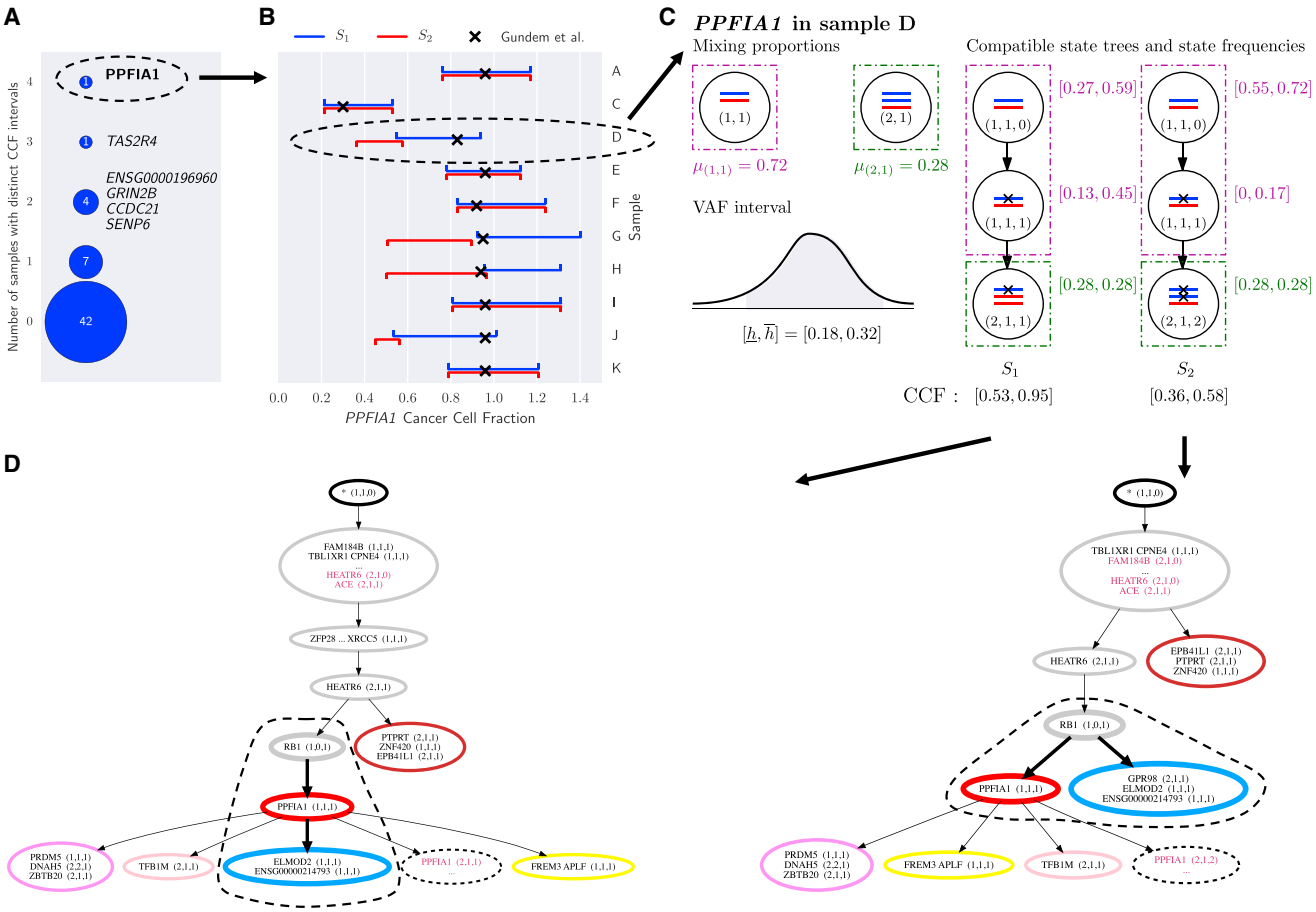


Figure 4. SPRUCE Analysis of Prostate Cancer Sample A22 Reveals Different Phylogenies

(A) The number of characters (SNVs) with distinct CCF intervals in individual samples, computed from the set of compatible state trees of each character. Sizes of circles correspond to the number of characters. *PPFIA1* has distinct CCF intervals in four samples.

(B) *PFFIA1* is compatible with two state trees, S_1 and S_2. We show for each sample the CCF interval for S_1 (blue intervals), S_2 (red intervals), and the cluster CCF (marked ×) reported in Gundem et al. (2015). The Gundem et al. (2015) cluster CCFs are largely consistent with S_1.

(C) In sample D (purity $\rho = 0.772$), the 99.9% VAF confidence interval for *PFFIA1* and the mixing proportions yield two compatible state trees, S_1 and S_2, with distinct CCF intervals. Note that S_1 adheres to the assumption in Gundem et al. (2015) of a fixed number of mutated copies.

(D) Condensed representation of the resulting perfect phylogeny trees using S_1 (left) and S_2 (right). Vertices are labeled by the introduced character-state pairs, and colors correspond to clusters reported in Gundem et al. (2015). In the left tree the blue cluster is a child of *PFFIA1*, whereas in the right tree it is a sibling of *PFFIA1*. See Figure S5 for the complete trees.

character and state tree pairs. There were 310,029 maximal cliques with a maximum size of 30 (Figure S4C). Thus, the largest possible tree we could recover had at most 30 characters. The maximum cliques did not contain all characters for the same reasons stated above (errors in the input data or violations of the infinite alleles model).

The CCF of an SNV depends on the assigned state tree. For each compatible state tree, we computed the CCF interval using the purity values reported in Gundem et al. (2015). Figure 4A shows that there are characters with multiple compatible state trees that yield distinct CCF intervals. Notably, *PPFIA1* admits two compatible state trees with distinct CCF intervals for four samples (Figure 4B). *PPFIA1* is reported as undergoing a single-copy amplification. The two state trees (S_1 and S_2) for this character both have the SNV occur prior to the amplification, but in S_1 the non-mutated copy is amplified whereas in S_2 the mutated copy is amplified (Figure 4C). Gundem et al. (2015) assumed that the number of copies of the mutation is constant across all cells that have the mutation, and they reported a single CCF for this mutation. Since their assumption rules out S_2, it is not surprising that we observed that their reported CCFs are more consistent with the CCFs resulting from S_1 than those from S_2. However, we emphasize that there is no signal in the data that requires one to assume that the number of copies of the mutation is constant across all cells that have the mutation.

We ran SPRUCE on all maximal cliques that contained either S_1 or S_2 for *PPFIA1*. Figure 4D shows two resulting trees: one using state tree S_1 and the other using state tree S_2. The topology of the two trees is largely similar, except for the location of the cluster containing *ELMOD2* (colored blue). In the first tree, *PPFIA1* (colored red) is the parent of the blue cluster. This is the same relationship reported in Gundem et al. (2015) (Figure S4A). In contrast, in the second tree *PPFIA1* and the blue cluster are siblings. As the second tree allows for significantly

lower CCF values across samples (Figure 4B), *RB1* (colored gray) can accommodate both *PPFIA1* and the blue cluster as children. In summary, we have shown that different state trees lead to different CCFs and ultimately different phylogenetic reconstructions of the observed data.

DISCUSSION

We introduce SPRUCE, a method that reconstructs phylogenetic trees using both SNVs and CNAs measured in bulk sequencing data from multiple samples of a tumor. SPRUCE derives from our theoretical results that characterize the solutions of the PPMDP. This theoretical framework enables us to evaluate whether a dataset (in whole or in part) satisfies the no-homoplasy, or infinite alleles, assumption for multi-state characters. We applied SPRUCE to both real and simulated cancer sequencing datasets. In all cases, we found that there was substantial ambiguity in bulk sequencing data, with many alternative phylogenetic reconstructions that were consistent with the data. This observation has several important implications.

First, the ambiguity in phylogenetic reconstructions suggests caution in drawing strong conclusions about the somatic mutation process based on a single reconstruction. For example, on a prostate cancer tumor (Gundem et al., 2015), we found that only a subset of characters was compatible with a multi-state perfect phylogeny. The observed incompatibilities could be a result of inaccuracies in the input data, including uncertainties in VAFs or incorrect copy number calls and mixing proportions, or violation of the infinite alleles model. Notably, methods that analyze copy number data often make assumptions regarding the number of copy states present in the sample. For example, Battenberg (Nik-Zainal et al., 2012) and THetA (Oesper et al., 2014) allow at most two tumor copy number states for each interval in each sample. However, close examination of all samples from the prostate tumor A22 reveals intervals with up to six copy number states (e.g., chromosome 8q), demonstrating the difficulty of obtaining data for accurate phylogenetic analysis of tumors.

The multiple sources of uncertainty in each step of tumor phylogenetic inference, including SNV calling, copy number segmentation, and estimation of CCFs, strongly argue against the current practice that treats these steps independently. One specific example of an issue with independent analyses is the computation of the CCF for an SNV. This quantity has been used extensively in analysis of intra-tumor heterogeneity, such as the inference of developmental patterns of metastases (Brastianos et al., 2015), the timing and fitness effects of cancer driver mutations (Andor et al., 2016; McGranahan et al., 2015), and the reconstruction of tumor phylogenies (Bolli et al., 2014; Gundem et al., 2015; Nik-Zainal et al., 2012). Many of these studies compute CCFs using single values of tumor purity and mutation copy number. However, the presence of multiple clones (i.e., intra-tumor heterogeneity) may confound the computation of a single value for the CCF. Instead, the CCF depends on the composition of each tumor sample, as revealed by the phylogenetic tree and mixing proportions. By jointly deriving CCFs and a phylogeny, SPRUCE reports alternative explanations of the input data. We demonstrated an example of two trees with distinct topologies that include different CCFs for *PPFIA1* mutations in the prostate cancer dataset from Gundem et al. (2015). Extending such ideas further, one could incorporate the determination of copy number states into the process of phylogenetic tree reconstruction.

Another implication of the ambiguity of phylogenetic trees from bulk sequencing data is for the design of tumor sequencing strategies. In any sequencing approach, one is faced with a trade-off in cost versus information gain. In tumor sequencing, two parameters that can be varied are the number of spatial or temporal samples (Gerlinger et al., 2012; Ling et al., 2015; Newburger et al., 2013) and the sequencing depth per sample (Griffith et al., 2015). Several papers have noted that a linear phylogeny can often be fit to data from a single tumor sample (El-Kebir et al., 2015; Hajirasouliha et al., 2014; Jiao et al., 2014; Malikic et al., 2015; Nik-Zainal et al., 2012; Strino et al., 2013) and that multiple samples are needed to accurately recover branching phylogenies. Our simulations show that increasing the number of samples is far more effective than increasing coverage per sample in reducing the number of possible phylogenies as well as increasing the accuracy of the inferred phylogenies. Indeed, low-coverage data with many samples can yield orders of magnitude fewer solutions than error-free data with a small number of samples. Multi-sample sequencing has other benefits, including the ability to assess spatial heterogeneity (Gerlinger et al., 2012; Ling et al., 2015; Sottoriva et al., 2015), and it decreases the likelihood that sample-specific batch effects affect conclusions of studies. While single-cell genome sequencing provides the highest resolution for studies of intra-tumor heterogeneity, technical and financial considerations have limited its use in large-scale studies (Wang et al., 2014; Zhang et al., 2015). In this work, we have shown that the combination of multi-sample sequencing and phylogenetic mixture algorithms can provide high-fidelity phylogenetic trees.

The large number of phylogenetic trees that are consistent with cancer sequencing data can present difficulties for probabilistic approaches that sample from this space. We showed that SPRUCE outperforms one sampling approach, PhyloWGS (Deshwar et al., 2015), on simulated datasets. We attribute this success to the difficulty of sampling the huge solution space directly, as PhyloWGS does, instead of exploiting the combinatorial structure, as SPRUCE does. We envision that the combinatorial characterization of the PPMDP that we derived could be useful for the development of better sampling approaches. For example, rather than considering generic tree-structured priors, one could use priors that are informed by the combinatorial structure of the multi-state ancestry graph.

Importantly, SPRUCE relies on the infinite alleles, or no-homoplasy, assumption. While this is a reasonable simplifying assumption, it may be violated in some tumors, particularly those with extensive CNAs. It is an interesting question whether more permissive phylogenetic models, such as the maximum parsimony model (Fitch, 1977), may more accurately model the evolutionary process. Extending maximum parsimony to the case of phylogenetic mixtures is an intriguing open computational problem. Finally, while we focused here on applications to cancer genome sequencing, the methodology and framework introduced here may be useful in other applications with mixed samples, including metagenomics (Segata et al., 2013) and studying the process of somatic hypermutation in the immune

system, as explored by Strino et al. (2013) using a single-sample, two-state perfect phylogeny model.

In summary, we have derived theoretical results on a phylogenetic tree mixture problem under the multi-state perfect phylogeny, or infinite alleles, model, and we used these results to obtain a practical algorithm, SPRUCE, that gives additional insights into intra-tumor heterogeneity. There is increasing evidence of the importance of intra-tumor heterogeneity in informing cancer treatment (Fisher et al., 2013), including immunotherapy (McGranahan et al., 2016). Thus, it is essential to have methods to reconstruct phylogenetic trees as accurately as possible from cancer sequencing data and to eliminate unnecessary assumptions about the evolutionary process.

EXPERIMENTAL PROCEDURES

In this section, we describe the combinatorial structure of solutions to the PPMDP and Cladistic-PPMDP by generalizing previous results on the two-state case. We then present our algorithm SPRUCE, which is outlined in Figure 2. Further details and proofs are presented in the Supplemental Experimental Procedures.

Mathematical Characterization of Solutions to the PPMDP

We first review results for the perfect phylogeny reconstruction problem with unmixed samples. Let $A \in \{0,\ldots,k-1\}^{m \times n}$ be a matrix whose rows correspond to m taxa and whose columns represent n characters, each of which has at most k states. If $k = 2$, then deciding whether A admits a perfect phylogeny is solvable in polynomial time (Estabrook et al., 1975; Gusfield, 1991). In contrast, if $k > 2$, the reconstruction of a multi-state perfect phylogeny is NP-complete in general (Bodlaender et al., 1992) but is fixed-parameter tractable in the number of states per character (Agarwala and Fernández-Baca, 1994; Kannan and Warnow, 1997). There is an elegant connection between multi-state perfect phylogeny and restricted triangulations of chordal graphs (Buneman, 1974), which recently was exploited by Gusfield and collaborators to obtain combinatorial conditions for the multi-state perfect phylogeny (Gusfield, 2010; Gysel and Gusfield, 2011).

In the case of mixed samples, the perfect phylogeny reconstruction problem becomes more complicated. In a two-state perfect phylogeny, a character changes to state 1 only once and thus the mutated state persists in the tree. Therefore, the subtree $\overline{T}_{(c,1)}$ rooted at vertex $v_{(c,1)}$ is equal to the subtree $T_{(c,1)}$ consisting of all vertices with state 1 for character c (Figure S1F). This observation provides strong constraints on the possible ancestral relationships between any pair of characters, given their observed frequencies across all samples. We write $(c,i) \prec_T (d,j)$ if $v_{(c,i)}$ is ancestral to $v_{(d,j)}$ in T. Note that $\prec_T$ is a reflexive relation, i.e., a vertex $v_{(c,i)}$ is ancestral to itself. Previously, El-Kebir et al. (2015) and Popic et al. (2015) showed that the phylogenetic trees that produce the observed frequencies correspond to constrained spanning trees of a certain graph. In El-Kebir et al. (2015), we defined the ancestry graph, which has a vertex $v_{(c,i)}$ for every character-state pair (c,i) and its edges relate two character-state pairs that are in a potential ancestral relationship, as captured by the ancestry condition (AC)

$$f_{p,(c,1)} \geq f_{p,(d,1)} \text{ for all samples } p \text{ and pairs of characters satisfying } (c,1) \prec_T (d,1).$$

We proved in El-Kebir et al. (2015) that solutions to the two-state problem with mixed samples correspond to constrained spanning trees of the ancestry graph that satisfy the sum condition (SC),

$$f_{p,(c,1)} \geq \sum_{(d,1)\in\delta(c,1)} f_{p,(d,1)} \text{ for all samples } p \text{ and pairs of characters satisfying } (c,1) \prec_T (d,1),$$

where $\delta(c,i)$ denotes the children of $v_{(c,i)}$. Note that SC is a generalization of the Pigeon-Hole Principle (Nik-Zainal et al., 2012), the principle that the sum of CCFs of the children of each vertex in the tree does not exceed 1. The SC also has appeared in other forms elsewhere (Jiao et al., 2014; Malikic et al., 2015; Popic et al., 2015; Strino et al., 2013).

In a multi-state perfect phylogeny, the state of a character c can change more than once on the tree, but never changes back to a previous state. Thus, in general, $T_{(c,i)}$ is not equal to $\overline{T}_{(c,i)}$ as a character can change state again in $\overline{T}_{(c,i)}$ (Figure S1F). This makes the situation much more complicated, since we must consider not only the children of $v_{(c,i)}$ but also the relationships between $T_{(c,i)}$ and subtrees $T_{(c,j)}$ for $j \neq i$. We capture these relationships using the notion of a descendant set for each character-state pair (c,i). Formally, given a complete perfect phylogeny tree T, we define the descendant set $D_{(c,i)} = \{j \mid (c,i) \prec_T (c,j)\}$ as the set of states for character c that are descendants of character-state pair (c,i) in T. The descendant set of a character precisely determines the relationship between $\overline{T}_{(c,i)}$ and $T_{(c,i)}$, namely, we have $\overline{T}_{(c,i)} = \bigcup_{\ell\in D_{(c,i)}} T_{(c,\ell)}$ (Figure S1F). Hence, to obtain the usages of all vertices in $\overline{T}_{(c,i)}$, we must consider cumulative frequencies $f_p^+(D_{(c,i)}) = \sum_{\ell\in D_{(c,i)}} f_{p,(c,\ell)}$, i.e., the sum of the frequencies of all descendant states of the character-state pair. It turns out that the positivity of values $u_{p,(c,i)}$ is a necessary and sufficient condition for T to generate $\mathcal{F}$, as captured by the multi-state sum condition (MSSC) in the following theorem.

Theorem 1

A complete perfect phylogeny tree T generates $\mathcal{F}$ if and only if

$$f_p^+(D_{(c,i)}) \geq \sum_{(d,j)\in\delta(c,i)} f_p^+(D_{(d,j)}) \text{ for all samples } p \text{ and character-state pairs } (c,i).$$

Note that MSSC is a generalization of SC. In the two-state case, $D_{(c,1)} = \{1\}$ for all characters c, and, thus, the cumulative frequencies $f_p^+(D_{(c,1)})$ are equal to the input frequencies $f_{p,(c,1)}$. Hence, in the two-state case, the relative order of the frequencies of the mutated (= 1) states of characters constrain the ancestral relationships between characters, as described above. In contrast, in the multi-state case, we must consider the relative order of cumulative frequencies $f_p^+(D_{(c,i)})$, but these depend on descendant sets that are a priori unknown. Thus, we must consider all combinations of descendant sets $D_{(c,i)}$ and $D_{(d,j)}$ that satisfy the multi-state ancestry condition (MSAC).

Proposition 1

Let T be a complete perfect phylogeny tree that generates $\mathcal{F}$. If $(c,i) \prec_T (d,j)$, then there exists a valid descendant set pair $D_{(c,i)}, D_{(d,j)}$ such that

$$f_p^+(D_{(c,i)}) \geq f_p^+(D_{(d,j)}) \text{ for all samples } p \text{ and character-state pairs } (c,i),(d,j).$$

Note that MSAC is a generalization of AC. For two character-state pairs, there are potentially many descendant set pairs satisfying MSAC. As such, the multi-state ancestry graph $G_{\mathcal{F}}$ is a multi-graph whose vertices correspond to character-state pairs and whose multi-edges are labeled by pairs of descendant sets that respect MSAC. In the Cladistic-PPMDP, the descendant sets $D_{(c,i)}$ are determined by the input set $\mathcal{S}$ of state trees. Thus, the cladistic multi-state ancestry graph $G_{(\mathcal{F},\mathcal{S})}$ is a simple graph, similar to the $k = 2$ case. Solutions to a PPMDP instance $\mathcal{F}$ correspond to threaded spanning trees (defined in the Supplemental Experimental Procedures) of the multi-state ancestry graph $G_{\mathcal{F}}$ that satisfy MSSC and that respect the descendant sets of the edges, as shown in the following theorem.

Theorem 2

A complete perfect phylogeny tree T generates $\mathcal{F}$ if and only if T is a threaded spanning tree of the ancestry graph $G_{\mathcal{F}}$ such that MSSC holds.

Similarly, we show in the Supplemental Experimental Procedures that solutions to a Cladistic-PPMDP instance $(\mathcal{F},\mathcal{S})$ correspond to threaded spanning trees of the cladistic multi-state ancestry graph $G_{(\mathcal{F},\mathcal{S})}$.

We previously have shown a hardness result for the PPMDP in the case with $k = 2$ states and $m = O(n)$ samples (El-Kebir et al., 2015). In the Supplemental Experimental Procedures, we prove a stronger hardness result with $m = 2$ samples (Figure S6).

SPRUCE Algorithm for the Cladistic-PPMDP

SPRUCE (Figure 2) takes as input VAF confidence intervals and copy number mixing proportions for each character per sample, and it outputs multi-state perfect phylogeny trees using the following four steps.

We start in step 1 by computing state trees $\{S_c\}$, for each character c, that are compatible with the input data. Specifically, our input consists of variant allele frequencies $H = [h_{p,c}]$, for each character c and sample p, and copy number calls $\mathcal{M} = [[\mu_{p,c,(x,y)}]]$, where $\mu_{p,c,(x,y)}$ is the mixing proportion of the copy number state (x,y) of character c in sample p. The latter are obtained using a copy number caller on the B-allele frequencies and the read-depth ratios

(Fisher et al., 2013; Ha et al., 2014; Nik-Zainal et al., 2012; Oesper et al., 2014). Using biologically motivated assumptions, we show in the Supplemental Experimental Procedures how to arrive at a set of compatible state trees for each character c given H and $\mathcal{M}$.

To obtain instances $(\mathcal{F},\mathcal{S})$ of the Cladistic-PPMDP, one could naively consider all combinations of state tree to character assignments. Instead in step 2, we restrict the combinations that we need to consider by introducing the compatibility graph. We denote by $\mathcal{F}[X]$ the subtensor of $\mathcal{F}$ that is restricted to characters $X \subseteq [n]$. A set of state trees on characters $X \subseteq [n]$ is compatible if there exists a tree that generates $\mathcal{F}[X]$. Now consider an instance $(\mathcal{F},\mathcal{S})$ and let T be a tree that generates $\mathcal{F}$ and is consistent with $\mathcal{S}$. Then, for any pair (c,d) of characters, it holds that there exists a smaller tree that generates $\mathcal{F}[\{c,d\}]$ and is consistent with S_c and S_d. We say that S_c and S_d are pairwise compatible. We restrict the combinations we consider to those that consist of pairwise-compatible state trees. More precisely, we define a compatibility graph with edges between two vertices (c,S_c) and (d,S_d) if and only if S_c and S_d are pairwise compatible.

In step 3, we use the compatibility graph to obtain instances of the Cladistic-PPMDP. Following the discussion above, each instance $(\mathcal{F},\mathcal{S})$ that admits a consistent tree T and generates $\mathcal{F}$ must be composed of a set of pairwise-compatible state tree and character pairs. Such sets correspond to cliques in the compatibility graph. In our application, we are not guaranteed to find trees that contain all the characters. Instead, we aim to find trees on the largest subset of characters that form a maximal clique in the compatibility graph. We enumerate all maximal cliques using the Bron-Kerbosch algorithm (Bron and Kerbosch, 1973).

Finally, for step 4, we find all threaded spanning trees that satisfy MSSC in $G_{(\mathcal{F},\mathcal{S})}$ by adapting the Gabow-Myers algorithm (Gabow and Myers, 1978) for enumerating spanning trees. This approach was applied previously to the two-state problem in Popic et al. (2015). In the Supplemental Experimental Procedures, we describe an enumeration algorithm that deals with multi-state characters and noisy frequencies.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, one table, and one data file and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2016.07.004.

AUTHOR CONTRIBUTIONS

M.E.-K. and B.J.R. conceived the project. M.E.-K., G.S., and B.J.R. developed the theory and algorithms. M.E.-K. implemented the algorithm. M.E.-K., G.S., and L.O. performed simulations and runs on real data. M.E.-K., G.S., L.O., and B.J.R. wrote the manuscript.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation (NSF) grant IIS-1016648 and NIH grants R01HG005690, R01HG007069, and R01CA180776 to B.J.R. B.J.R. is supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, an Alfred P. Sloan Research Fellowship, and an NSF CAREER Award (CCF-1053753). An early version of this paper was submitted to and peer reviewed at the 2016 Annual International Conference on Research in Computational Molecular Biology (RECOMB). The manuscript was revised and then independently further reviewed at *Cell Systems*.

Received: March 16, 2016
Revised: June 29, 2016
Accepted: July 6, 2016
Published: July 27, 2016

REFERENCES

Agarwala, R., and Fernández-Baca, D. (1994). A polynomial-time algorithm for the perfect phylogeny problem when the number of character states is fixed. SIAM J. Comput. *23*, 1216–1224.

Andor, N., Graham, T.A., Jansen, M., Xia, L.C., Aktipis, C.A., Petritsch, C., Ji, H.P., and Maley, C.C. (2016). Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. Nat. Med. *22*, 105–113.

Bodlaender, H.L., Fellows, M.R., and Warnow, T.J. (1992). Two strikes against perfect phylogeny. In Automata, Languages and Programming, W. Kuich, ed. (Springer), pp. 273–283.

Bolli, N., Avet-Loiseau, H., Wedge, D.C., Van Loo, P., Alexandrov, L.B., Martincorena, I., Dawson, K.J., Iorio, F., Nik-Zainal, S., Bignell, G.R., et al. (2014). Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. Nat. Commun. *5*, 2997.

Brastianos, P.K., Carter, S.L., Santagata, S., Cahill, D.P., Taylor-Weiner, A., Jones, R.T., Van Allen, E.M., Lawrence, M.S., Horowitz, P.M., Cibulskis, K., et al. (2015). Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets. Cancer Discov. *5*, 1164–1177.

Bron, C., and Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. Commun. ACM *16*, 575–577.

Buneman, P. (1974). A note on the metric properties of trees. J. Comb. Theory B *17*, 48–50.

Chowdhury, S.A., Gertz, E.M., Wangsa, D., Heselmeyer-Haddad, K., Ried, T., Schäffer, A.A., and Schwartz, R. (2015). Inferring models of multiscale copy number evolution for single-tumor phylogenetics. Bioinformatics *31*, i258–i267.

Deshwar, A.G., Vembu, S., Yung, C.K., Jang, G.H., Stein, L., and Morris, Q. (2015). PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol. *16*, 35.

Eirew, P., Steif, A., Khattra, J., Ha, G., Yap, D., Farahani, H., Gelmon, K., Chia, S., Mar, C., Wan, A., et al. (2015). Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. Nature *518*, 422–426.

El-Kebir, M., Oesper, L., Acheson-Field, H., and Raphael, B.J. (2015). Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. Bioinformatics *31*, i62–i70.

Estabrook, G.F., Johnson, C.S., Jr., and Mc Morris, F.R. (1975). An idealized concept of the true cladistic character. Math. Biosci. *23*, 263–272.

Fernández-Baca, D. (2001). The perfect phylogeny problem. In Steiner Trees in Industry, X. Cheng and D.Z. Du, eds. (Springer), pp. 203–234.

Fisher, R., Pusztai, L., and Swanton, C. (2013). Cancer heterogeneity: implications for targeted therapeutics. Br. J. Cancer *108*, 479–485.

Fitch, W.M. (1977). On the problem of discovering the most parsimonious tree. Am. Nat. *111*, 223–257.

Gabow, H.N., and Myers, E.W. (1978). Finding all spanning trees of directed and undirected graphs. SIAM J. Comput. *7*, 280–287.

Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N. Engl. J. Med. *366*, 883–892.

Griffith, M., Miller, C.A., Griffith, O.L., Krysiak, K., Skidmore, Z.L., Ramu, A., Walker, J.R., Dang, H.X., Trani, L., Larson, D.E., et al. (2015). Optimizing cancer genome sequencing and analysis. Cell Syst. *1*, 210–223.

Gundem, G., Van Loo, P., Kremeyer, B., Alexandrov, L.B., Tubio, J.M., Papaemmanuil, E., Brewer, D.S., Kallio, H.M., Högnäs, G., Annala, M., et al.; ICGC Prostate UK Group (2015). The evolutionary history of lethal metastatic prostate cancer. Nature *520*, 353–357.

Gusfield, D. (1991). Efficient algorithms for inferring evolutionary trees. Networks *21*, 19–28.

Gusfield, D. (2010). The multi-state perfect phylogeny problem with missing and removable data: solutions via integer-programming and chordal graph theory. J. Comput. Biol. *17*, 383–399.

Gusfield, D. (2014). ReCombinatorics: The Algorithmics of Ancestral Recombination Graphs and Explicit Phylogenetic Networks (MIT Press).

Gysel, R., and Gusfield, D. (2011). Extensions and improvements to the chordal graph approach to the multistate perfect phylogeny problem. IEEE/ACM Trans. Comput. Biol. Bioinform. *8*, 912–917.

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L.M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., et al. (2014). TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome Res. *24*, 1881–1893.

Hajirasouliha, I., Mahmoody, A., and Raphael, B.J. (2014). A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. Bioinformatics *30*, i78–i86.

Jiao, W., Vembu, S., Deshwar, A.G., Stein, L., and Morris, Q. (2014). Inferring clonal evolution of tumors from single nucleotide somatic mutations. BMC Bioinformatics *15*, 35.

Kannan, S., and Warnow, T. (1997). A fast algorithm for the computation and enumeration of perfect phylogenies. SIAM J. Comput. *26*, 1749–1763.

Li, B., and Li, J.Z. (2014). A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. Genome Biol. *15*, 473.

Ling, S., Hu, Z., Yang, Z., Yang, F., Li, Y., Lin, P., Chen, K., Dong, L., Cao, L., Tao, Y., et al. (2015). Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. Proc. Natl. Acad. Sci. USA *112*, E6496–E6505.

Malikic, S., McPherson, A.W., Donmez, N., and Sahinalp, C.S. (2015). Clonality inference in multiple tumor samples using phylogeny. Bioinformatics *31*, 1349–1356.

McGranahan, N., Favero, F., de Bruin, E.C., Birkbak, N.J., Szallasi, Z., and Swanton, C. (2015). Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. Sci. Transl. Med. *7*, 283ra54.

McGranahan, N., Furness, A.J., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S.K., Jamal-Hanjani, M., Wilson, G.A., Birkbak, N.J., Hiley, C.T., et al. (2016). Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. Science *351*, 1463–1469.

Newburger, D.E., Kashef-Haghighi, D., Weng, Z., Salari, R., Sweeney, R.T., Brunner, A.L., Zhu, S.X., Guo, X., Varma, S., Troxell, M.L., et al. (2013). Genome evolution during progression to breast cancer. Genome Res. *23*, 1097–1108.

Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., et al.; Breast Cancer Working Group of the International Cancer Genome Consortium (2012). The life history of 21 breast cancers. Cell *149*, 994–1007.

Nowell, P.C. (1976). The clonal evolution of tumor cell populations. Science *194*, 23–28.

Oesper, L., Satas, G., and Raphael, B.J. (2014). Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. Bioinformatics *30*, 3532–3540.

Popic, V., Salari, R., Hajirasouliha, I., Kashef-Haghighi, D., West, R.B., and Batzoglou, S. (2015). Fast and scalable inference of multi-sample cancer lineages. Genome Biol. *16*, 91.

Sanborn, J.Z., Chung, J., Purdom, E., Wang, N.J., Kakavand, H., Wilmott, J.S., Butler, T., Thompson, J.F., Mann, G.J., Haydu, L.E., et al. (2015). Phylogenetic analyses of melanoma reveal complex patterns of metastatic dissemination. Proc. Natl. Acad. Sci. USA *112*, 10995–11000.

Schuh, A., Becq, J., Humphray, S., Alexa, A., Burns, A., Clifford, R., Feller, S.M., Grocock, R., Henderson, S., Khrebtukova, I., et al. (2012). Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. Blood *120*, 4191–4196.

Segata, N., Börnigen, D., Morgan, X.C., and Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat. Commun. *4*, 2304.

Sottoriva, A., Kang, H., Ma, Z., Graham, T.A., Salomon, M.P., Zhao, J., Marjoram, P., Siegmund, K., Press, M.F., Shibata, D., and Curtis, C. (2015). A Big Bang model of human colorectal tumor growth. Nat. Genet. *47*, 209–216.

Strino, F., Parisi, F., Micsinai, M., and Kluger, Y. (2013). TrAp: a tree approach for fingerprinting subclonal tumor composition. Nucleic Acids Res. *41*, e165.

Venkatesan, S., and Swanton, C. (2016). Tumor evolutionary principles: how intratumor heterogeneity influences cancer treatment and outcome. Am. Soc. Clin. Oncol. Educ. Book *35*, e141–e149.

Wang, Y., Waters, J., Leung, M.L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., et al. (2014). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature *512*, 155–160.

Yuan, K., Sakoparnig, T., Markowetz, F., and Beerenwinkel, N. (2015). BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. Genome Biol. *16*, 36.

Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhsng, C.Z., Wala, J., Mermel, C.H., et al. (2013). Pan-cancer patterns of somatic copy number alteration. Nat. Genet. *45*, 1134–1140.

Zhang, C.Z., Adalsteinsson, V.A., Francis, J., Cornils, H., Jung, J., Maire, C., Ligon, K.L., Meyerson, M., and Love, J.C. (2015). Calibrating genomic and allelic coverage bias in single-cell sequencing. Nat. Commun. *6*, 6822.

Cell Systems, Volume *3*

Supplemental Information

Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures

Mohammed El-Kebir, Gryte Satas, Layla Oesper, and Benjamin J. Raphael

Supplemental Material: Inferring the Mutational History of a Tumor using Multi-State Perfect Phylogeny Mixtures

Mohammed El-Kebir[1,*], Gryte Satas[1,*], Layla Oesper[1,2], and Benjamin J. Raphael[1,†]

[1]Department of Computer Science and Center for Computational Molecular Biology, Brown University, Providence, RI 02912.

[2]Department of Computer Science, Carleton College, Northfield, MN 55057.

[*]The authors wish it to be known that the first two authors should be considered joint first authors

[†]Correspondence: braphael@brown.edu

Contents

A Supplemental Figures

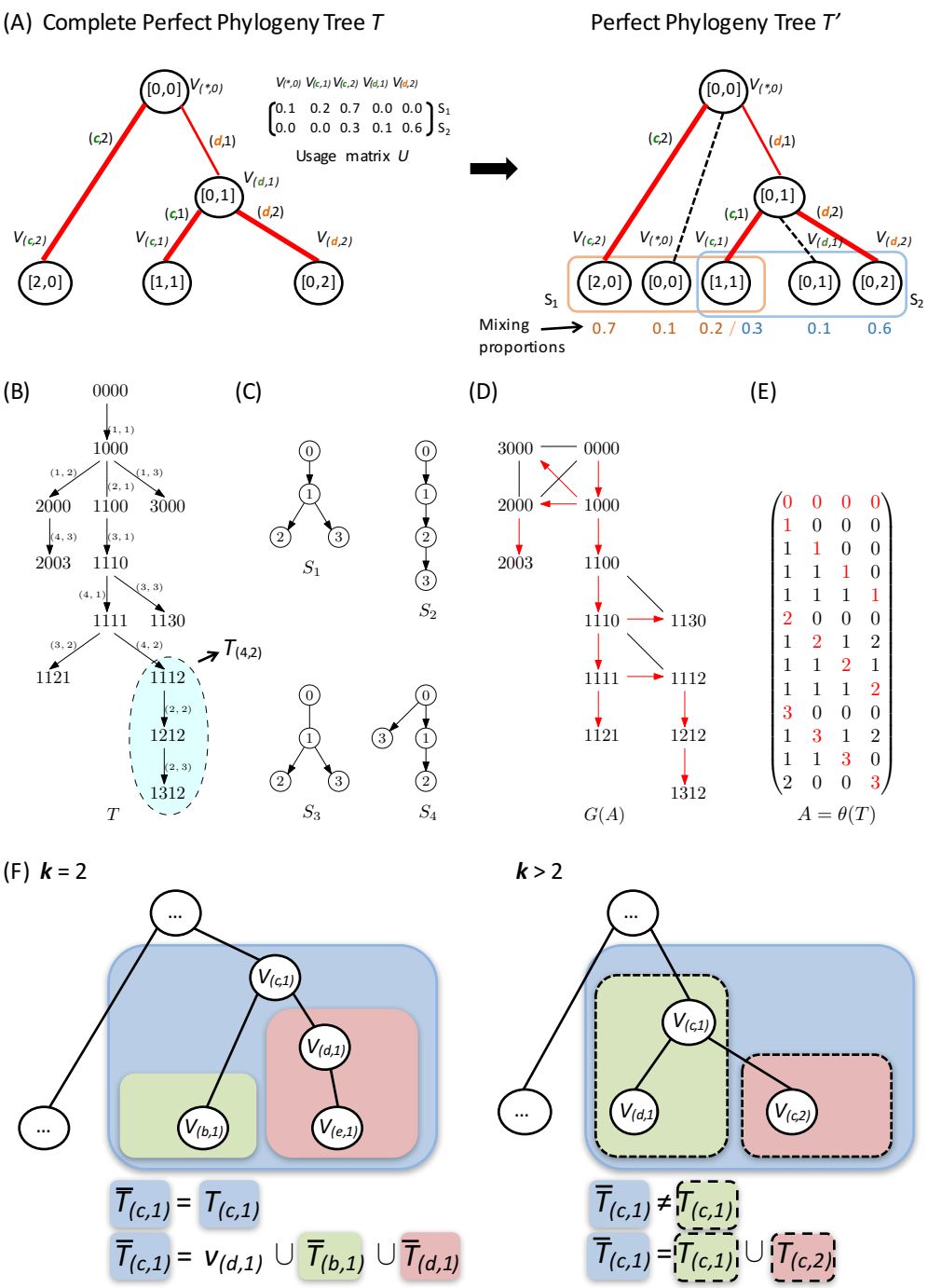


Figure S1: **Concepts of the Perfect Phylogeny Mixture Deconvolution Problem. Related to Figure 1.** (A) Each complete perfect phylogeny tree T corresponds to a perfect phylogeny tree T'. (B) A $4,4$-complete perfect phylogeny tree T. (C) State trees $\mathcal{S}$ determined by T. (D) Red edges denote a spanning tree rooted at 0000 that corresponds to T. (E) $4,4$-complete perfect phylogeny matrix $A = \theta(T)$. Note that entries in red correspond to the first two conditions of Definition 5. (F) Relationship between $T_{(c,i)}$ and $\overline{T}_{(c,i)}$. In the case of $k = 2$ states, we have that $T_{(c,1)}) = \overline{T}_{(c,1)}$. In the case of $k > 2$ states, $T_{(c,i)} \neq \overline{T}_{(c,i)}$. Instead, $\overline{T}_{(c,i)} = \bigcup_{l \in D_{(c,i)}} T_{(c,l)}$ where $D_{(c,i)}$ is the descendant set of (c, i). Here, $D_{(c,1)} = \{1, 2\}$.

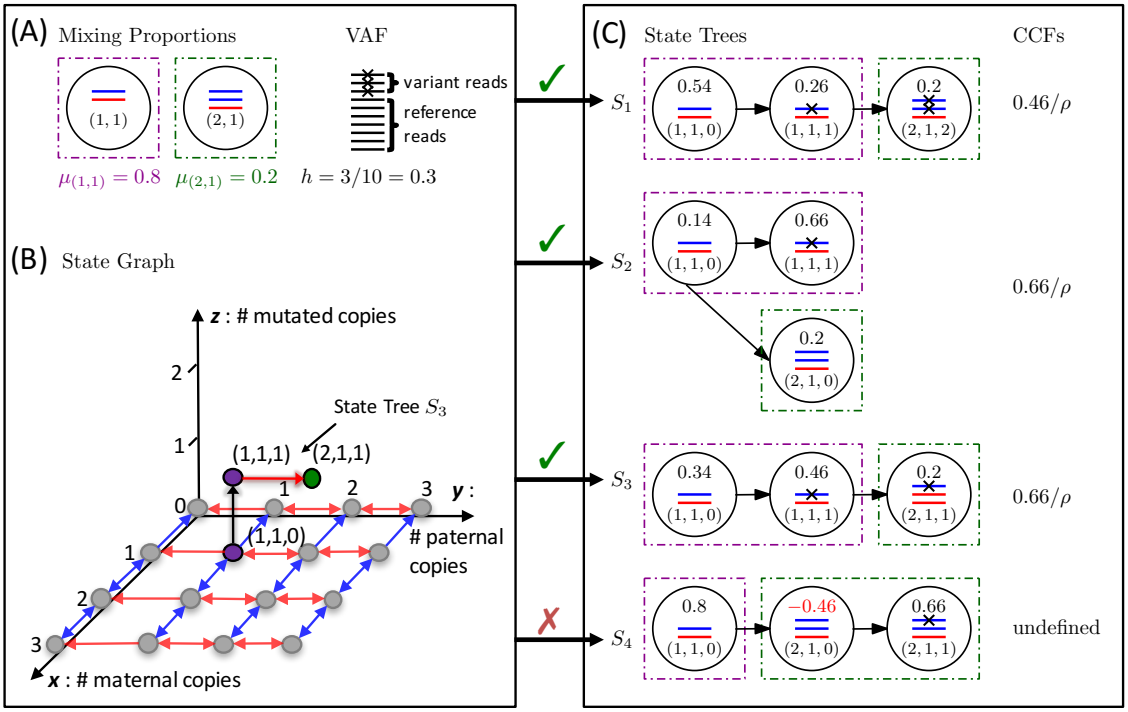
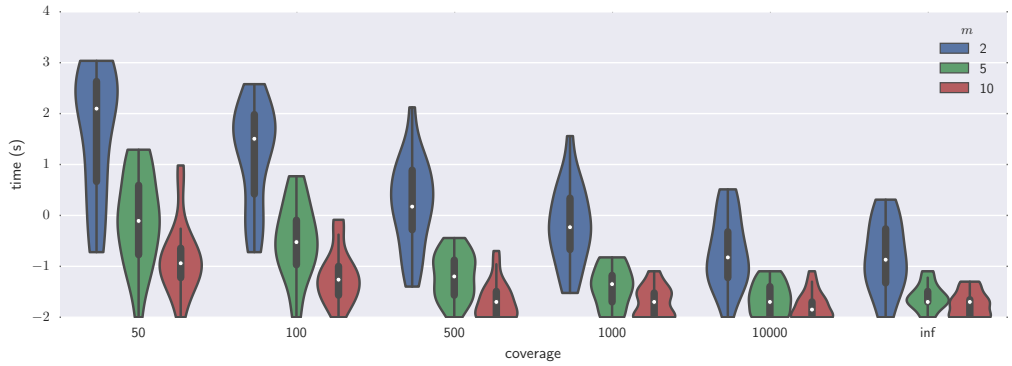
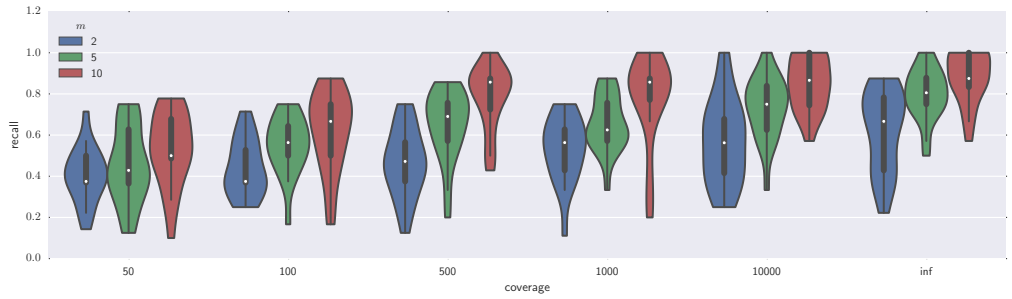


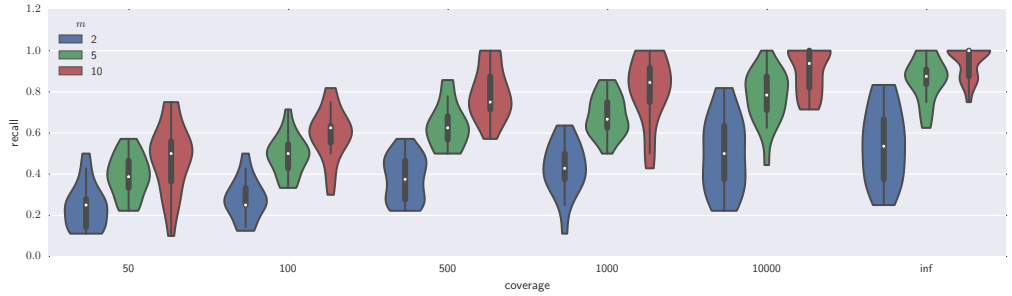
Figure S2: **The cancer cell fraction (CCF) of an SNV depends on its copy-number states, variant allele frequency and state tree. Related to Figure 2B.** (A) For each character in each sample, we observe copy-number states (x, y) with mixing proportions $\mu_{(x,y)}$ as well as a VAF h. (B) Our multi-state model encodes the somatic mutational process in cancer in the form of a state graph whose vertices (x, y, z) denote x maternal copies, y paternal copies and z mutated copies of the locus. The edges of the state graph correspond to mutation events, amplifications and deletions. A state tree of a character models the evolutionary history of its mutational states and corresponds to a constrained subtree of the state graph. (C) The observed data of an SNV and the state graph determine a set of compatible state trees, which have non-negative frequencies for each state. Here, states trees S_1, S_2 and S_3 are compatible with the input, whereas state tree S_4 is incompatible because it has a negative frequency for state $(2, 1, 0)$. The CCF of state trees S_2 and S_3 is $0.66/\rho$, where ρ is the purity of the sample. State tree S_1 has different numbers of copies of the mutation resulting in a distinct CCF of $0.46/\rho$.



(A)



(B)



(C)

(D)

Figure S3: **Simulated data results for $n = 5$ instances with noisy VAFs. Related to Figure 3.** A coverage of 'inf' corresponds to error-free VAFs. (A) Running time (seconds, log-scale). (B) Recall of representative tree. (C) Median recall. (D) Effect of violations of the infinite alleles assumption on recall. Shown are recall values for 20 instances with $n = 5$, $m = 10$ and a target coverage of 1,000x; x-axis denotes the number of violations.

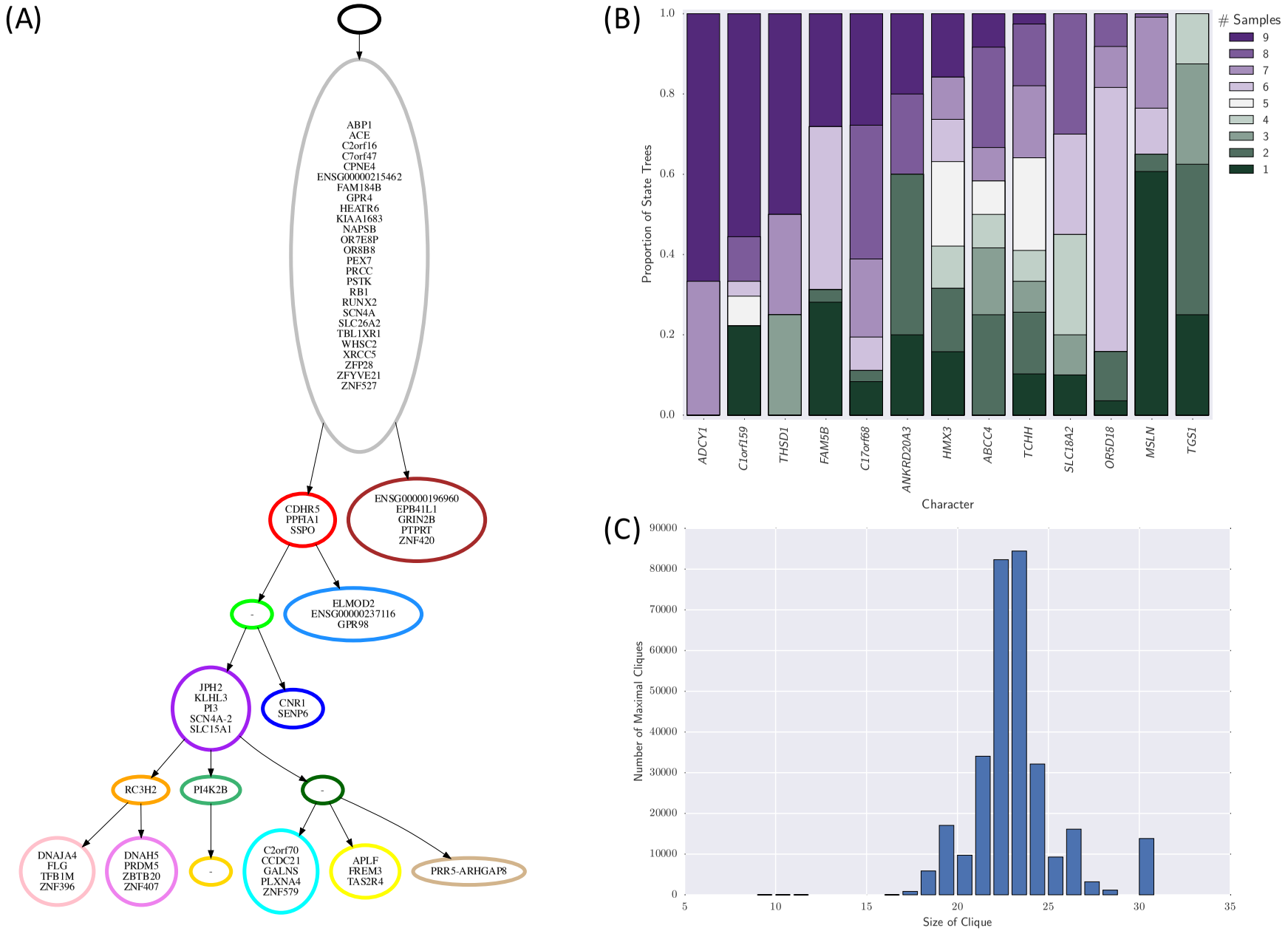


Figure S4: **Additional results for A22. Related to Figure 4.** (A) Reported tree for A22 by Gundem et al. [7]. Vertices and coloring correspond to the reported clusters on SNVs. Note that we only show the subset of SNVs that pass the filtering criteria described in Results. As such, some of the reported clusters are empty. (B) Number of compatible state trees for A22. Thirteen characters had no state tree that was compatible across all samples. Each column corresponds to a character, with stacked bars denoting the fraction of state trees compatible with a specified number of samples indicated by the color. (C) Distribution of sizes of maximal cliques in the compatibility graph of A22.

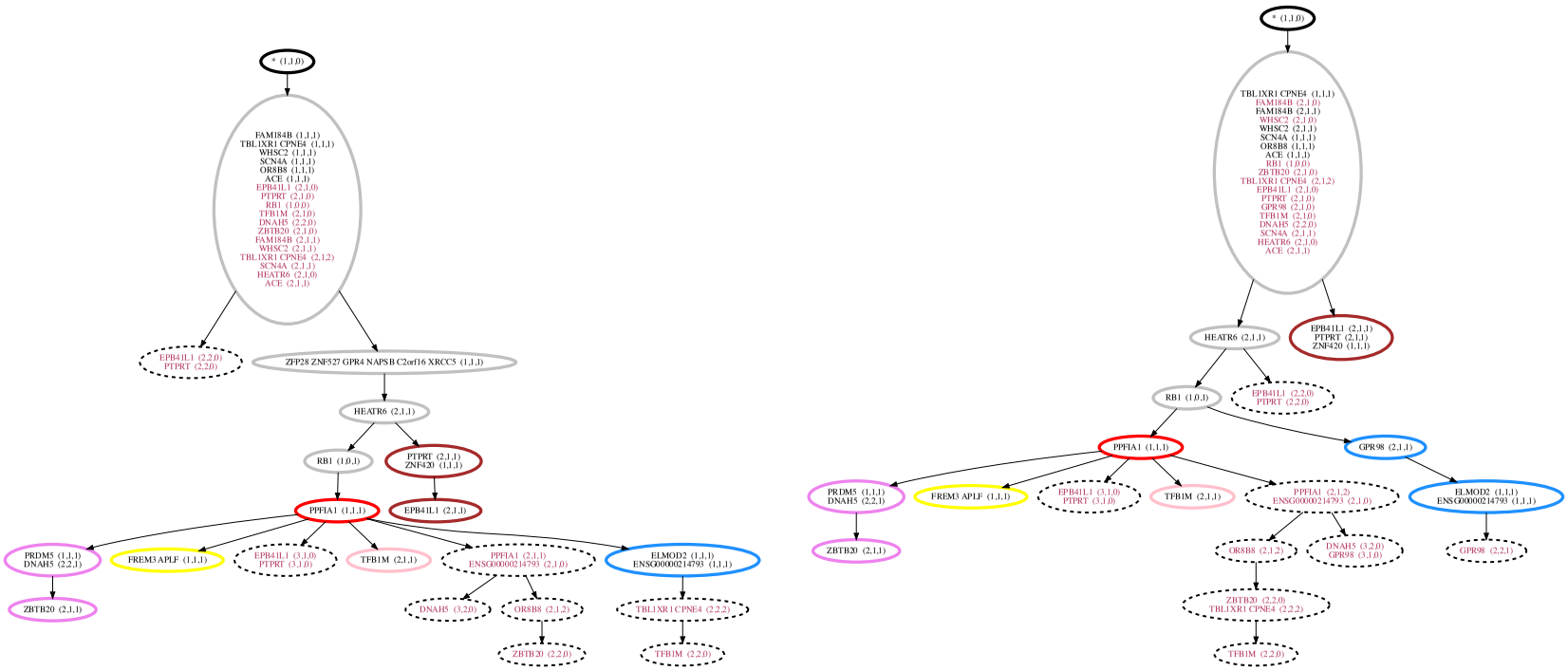
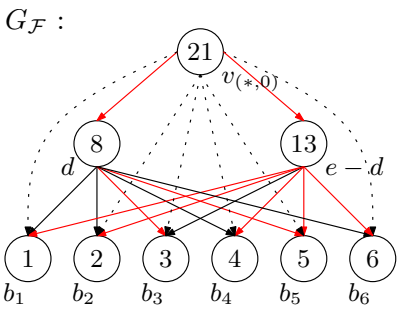


Figure S5: **Phylogenetic trees for A22. Related to Figure 4D.** (A) State tree S_1 for *PPFIA1* results in a tree where *PPFIA1* is a parent of the blue cluster. This relationship is consistent with the tree reported in [7]. (B) State tree S_2 for *PPFIA1* results in a tree in which the blue and the red clusters are siblings.



$B = \{1, 2, 3, 4, 5, 6\}$

$d = 8$

$e = 21$

$$F_0 = \tfrac{1}{21}\begin{pmatrix} 13 & 8 & 20 & 19 & 18 & 17 & 16 & 15 \\ 8 & 13 & 21-6\epsilon & 21-5\epsilon & 21-4\epsilon & 21-3\epsilon & 21-2\epsilon & 21-\epsilon \end{pmatrix}$$

$$F_1 = \tfrac{1}{21}\begin{pmatrix} 8 & 13 & 1 & 2 & 3 & 4 & 5 & 6 \\ 13 & 8 & 6\epsilon & 5\epsilon & 4\epsilon & 3\epsilon & 2\epsilon & \epsilon \end{pmatrix}$$

Figure S6: **Reduction from SUBSET SUM. Related to Figure 1.**

B Supplemental Experimental Procedures

In Section B.1, we describe the procedure used to generate simulated instances. We provide additional details on the comparison of Spruce to PhyloWGS [2] in Section B.2. We introduce the Perfect Phylogeny Mixture Deconvolution Problem (PPMDP) in Section B.3. We then consider a restricted version with cladistic characters in Section B.4. Finally, we tie things back to our cancer application in Section B.5, where we introduce a multi-state model that captures the somatic mutational process in cancer.

B.1 Generation of Simulated Instances

We generated simulated instances using the following procedure. We create 20 multi-state perfect phylogeny trees T^* containing $n = 5$ characters, using randomly generated state trees that have at most two copy-number states per character with $x \leq 3$ and $y \leq 3$. For each simulated tree T^*, we generate a frequency tensor $\mathcal{F}$ containing $m \in \{2, 5, 10\}$ samples, by mixing vertices from T^*. This results in 60 simulated mixtures. Next, we use the entries of the simulated $\mathcal{F}$ to compute the input by calculating for each character c in sample p, the VAF $h_{p,c}$ and the mixing proportions $\mu_{p,c,(x,y)}$ as defined in the main text. For each character c and sample p, we draw its total read count from a Poisson distribution parameterized by a target coverage. We then draw the number of variant reads from a binomial parameterized by the previously drawn total read count and the true VAF. Finally, we consider the posterior distribution of observing the drawn total and variant read counts from which we compute 0.99 confidence intervals on the VAF [3].

B.2 Comparison against PhyloWGS

We run PhyloWGS with default parameters (2500 MCMC samples, 5000 Metropolis-Hastings iterations) on the simulated instances with $n = 5$ characters and $m \in \{2, 5, 10\}$ samples. As PhyloWGS does not have a "perfect data" mode, we provide correct VAF and copy number mixing proportions. We indicate that these are high-confidence by specifying a coverage of $1,000,000$x. We also run PhyloWGS on larger instances with $n = 15$ characters and noisy VAFs (1,000x coverage).

PhyloWGS has a two-state perfect phylogeny model where characters correspond to either CNA events or SNV events, and expected SNV frequencies are adjusted based on associated CNAs. As the method does not report the number of copies that contain each mutation, we could not directly map the results onto (x, y, z) states. Instead, for the comparison, we map the states in the true trees to an (x, y) copy number state, and an indicator w_c for the presence of a mutation in character c. We calculate recall as the fraction of edges (characterized by an ordered pair of states) from the simulated tree that are recovered in the solution. Some vertices from the output PhyloWGS trees contained clusters of mutations. In those instances, we include all combinations of pairs from the parent cluster to a child cluster.

B.3 Perfect Phylogeny Mixture Deconvolution Problem

In this section we introduce the Perfect Phylogeny Mixture Deconvolution Problem (PPMDP). We show in Section B.3.3 that there exists a unique matrix U relating a given frequency tensor $\mathcal{F}$ and complete perfect phylogeny tree T. Next, we derive in Section B.3.4 that solutions to the PPMDP correspond to constrained spanning trees in a directed, edge-labeled multi-graph. In Section B.3.5 we prove that the problem is NP-complete.

Recall that m is the number of samples, n is the number of characters and k is the number of states of each character. Our input measurements are given by the $k \times m \times n$ *frequency tensor* $\mathcal{F} = [F_i] = [[f_{p,(c,i)}]]$ where $f_{p,(c,i)}$ is the proportion of taxa of sample p that have state i for character c. We denote by F_i the slice of $\mathcal{F}$ where the state of each character is i. Formally, $\mathcal{F}$ is defined as follows.

Definition 1. *An $k \times m \times n$ tensor $\mathcal{F} = [F_i] = [[f_{p,(c,i)}]]$ is a* frequency tensor *provided $f_{p,(c,i)} \geq 0$ and $\sum_{i=0}^{k-1} f_{p,(c,i)} = 1$ for all characters c and samples p.*

As mentioned in the main text, the goal is to explain the observed frequencies $\mathcal{F}$ as m mixtures of the leaves of a perfect phylogeny tree T, where each mixture corresponds to one sample. We recall the definition of a perfect phylogeny [4, 8].

Definition 2. *A rooted tree T is a* perfect phylogeny tree *provided that (1) each vertex is labeled by a* state vector *in $\{0, \ldots, k-1\}^n$, which denotes the state for each character; (2) the root vertex of T has state 0 for each character; (3) vertices labeled with state i for character c form a connected subtree $T_{(c,i)}$ of T.*

Rather than explaining $\mathcal{F}$ as mixtures of the leaves of a perfect phylogeny tree, we aim to explain $\mathcal{F}$ as m mixtures of *all* vertices of an n, k-complete perfect phylogeny tree, which is defined as follows.

Definition 3. *An edge-labeled rooted tree T on $n(k-1)+1$ vertices is a n, k-*complete perfect phylogeny tree *provided each of the $n(k-1)$ edges is labeled with exactly one character-state pair from $[n] \times [k-1]$ and no character-state pair appears more than once in T. Let $\mathcal{T}_{n,k}$ be the set of all n, k-complete perfect phylogeny trees.*

We may do this without loss of generality, as each n, k-complete perfect phylogeny T can be mapped to a perfect phylogeny tree T' by extending inner vertices of T that have non-zero mixing proportions to leaves of T'. See Figure S1 for an example. In the following we denote by T an n, k-complete perfect phylogeny tree T, by $v_{(c,i)}$ the vertex of T whose incoming edge is labeled by (c, i), and by $v_{(*,i)}$ the root of T. Alternatively, $v_{(1,0)}, \ldots, v_{(n,0)}$ all refer to the root vertex $v_{(*,0)}$.

The observed frequencies $\mathcal{F}$ are related to the vertices of T by an $m \times (n(k-1)+1)$ *usage matrix* $U = [u_{p,(c,i)}]$ whose rows correspond to samples and columns to vertices of T such that each entry $u_{p,(c,i)}$ indicates the mixing proportion of the vertex $v_{(c,i)}$ of T in sample p. More specifically, each frequency $f_{p,(c,i)}$ is the sum of mixing proportions of all vertices of T that possess state i for character c, i.e. $f_{p,(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}$ (Figure S1B). Formally, we define a usage matrix U as follows.

Definition 4. *An $m \times (n(k-1)+1)$ matrix $U = [u_{p,(c,i)}]$ is an* m, n, k-usage matrix *provided $u_{p,(c,i)} \geq 0$, and $u_{p,(*,0)} + \sum_{c=1}^{n} \sum_{i=1}^{k-1} u_{p,(c,i)} = 1$ for all samples p. Let $\mathcal{U}_{m,n,k}$ be the set of all m, n, k-usage matrices U.*

Given $\mathcal{F}$, the goal is to infer a n, k-complete perfect phylogeny T and a usage matrix U such that mixing the vertices of T according to U results in $\mathcal{F}$, which was stated as Problem 1 in the main text.

Problem 1 (Perfect Phylogeny Mixture Deconvolution Problem (PPMDP))**.** *Given an $k \times m \times n$ frequency tensor $\mathcal{F} = [[f_{p,(c,i)}]]$, find a n, k-complete perfect phylogeny tree T and a m, n, k-usage matrix $U = [u_{p,(c,i)}]$ such that $f_{p,(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}$ for all character-state pairs (c, i) and all samples p.*

B.3.1 Relation to Two-State Perfect Phylogeny Mixtures

We now review and recast the main results for mixtures of a two-state ($k = 2$) perfect phylogeny with all-zero root. Here, a character changes state from 0 to 1 at most once. The key insight is that the relative frequencies of the mutated ($= 1$) states for a subset of characters constrain their potential ancestral relationships. This is because the mutated state persists in the tree. In particular, if $(c, 1) \prec_T (d, 1)$ then all vertices in T that have state 1 for character d must also have state 1 for character c. A consequence is the following condition, called the *Ancestry Condition* (AC) in [3]:

$$f_{p,(c,1)} \geq f_{p,(d,1)} \text{ for all samples } p \text{ and characters } (c, 1) \prec_T (d, 1). \tag{AC}$$

In fact, a stronger condition than the ancestry condition can be derived by considering the relationships between subtrees of T. Specifically, for each character c, the subtree $T_{(c,1)}$, consisting of all vertices with state 1 for character c, *is identical to* the subtree $\overline{T}_{(c,1)}$ rooted at a vertex $v_{(c,1)}$. Moreover, $\overline{T}_{(c,1)}$ is the disjoint union of $v_{(c,1)}$ and the subtrees rooted at its children: $\overline{T}_{(c,1)} = v_{(c,1)} \cup \left(\bigcup_{(d,1) \in \delta(c,1)} \overline{T}_{(d,1)} \right)$ (Figure S1F). Combining this fact together with the equation $f_{p,(c,1)} = \sum_{(d,1) \in T_{(c,1)}} u_{p,(d,1)}$ yields the key equation $f_{p,(c,1)} = u_{p,(c,1)} + \sum_{(d,1) \in \delta(c,1)} f_{p,(d,1)}$. Recalling that $u_{p,(c,1)} \geq 0$, we can relax this equation to the following inequality, referred to as the *Sum Condition* (SC) in [3],

$$f_{p,(c,1)} \geq \sum_{(d,1) \in \delta(c,1)} f_{p,(d,1)} \qquad \text{for all samples } p \text{ and characters } c. \tag{SC}$$

The sum condition is both necessary and sufficient for $\mathcal{F}$ to be a mixture of T. The sum and ancestry conditions provide a combinatorial characterization of solutions as constrained spanning trees of a directed acyclic graph, which was called the *ancestry graph* in [3]. This derivation of the sum condition from the fact that $\overline{T}_{(c,1)} = T_{(c,1)}$ in the two-state case is not explicitly stated in previous work, but this turns out to be the key ingredient in the generalization to the multi-state case.

B.3.2 Reformulating the PPMDP as k Matrix Factorization Problems

In the remainder of this section we show how Problem 1 can be restated as a linear algebra problem. We start by observing that each vertex $v_{(c,i)}$ of T defines a *state vector* $\mathbf{a}_{(c,i)} \in \{0, \ldots, k-1\}^n$ indicating the state

of each character at that vertex. The root vertex $v_{(*,0)}$ has state vector $\mathbf{a}_{(*,0)} = (0, \ldots, 0)$, i.e. $a_{(*,0),d} = 0$ for each character $d \in [n]$. The state vector $\mathbf{a}_{(c,i)}$ of the remaining vertices $v_{(c,i)} \neq v_{(*,0)}$ is the same as the state vector $\mathbf{a}_{\pi(c,i)}$ of the parent vertex $v_{\pi(c,i)}$ except at character c where the state is i. The state vectors of all the vertices of an n, k-complete perfect phylogeny tree T correspond to an $(n(k-1)+1) \times n$ matrix A (Figure S1D).

We now define a subset of matrices $A \in \{0, \ldots, k-1\}^{(n(k-1)+1)\times n}$ that we call *n, k-complete perfect phylogeny matrices* whose rows encode the state vectors of the vertices of an n, k-complete perfect phylogeny tree T. Let $A \in \{0, \ldots, k-1\}^{(n(k-1)+1)\times n}$. We define $G(A)$ as the undirected graph whose vertices correspond to the rows of A, and whose edges set consists of all pairs of vertices whose corresponding state vectors differ at exactly one position, i.e. have Hamming distance 1. We require that $G(A)$ is connected (Figure S1C), that every row $\mathbf{a}_{(c,i)}$ of A (where $i \in [k]$) introduces the character-state pair (c, i) and that there is a row $\mathbf{a}_{(*,0)}$ that contains only 0-s (Figure **??**). Formally, we say that A is an n, k-complete perfect phylogeny matrix if the following holds.

Definition 5. *Matrix $A = [a_{(c,i),d}] \in \{0, \ldots, k-1\}^{(n(k-1)+1)\times n}$ is a n, k-*complete perfect phylogeny matrix *provided $a_{(*,0),d} = 0$ for all characters d, $a_{(c,i),c} = i$ for all character-state pairs (c, i) and $G(A)$ is connected. Let $\mathcal{A}_{n,k}$ be the set of all n, k-complete perfect phylogeny matrices.*

Unlike the general multi-state perfect phylogeny problem [1], we can recognize complete perfect phylogeny matrices in polynomial time, as these matrices form a restricted subset of multi-state perfect phylogeny matrices whose rows unambiguously encode all the vertices of a corresponding tree. We relate complete perfect phylogeny trees to complete perfect phylogeny matrices by defining the following function.

Definition 6. *The function $\theta : \mathcal{T}_{n,k} \to \mathcal{A}_{n,k}$ maps a complete perfect phylogeny tree $T \in \mathcal{T}_{n,k}$ to the complete perfect phylogeny matrix $\theta(T) = A = [a_{(c,i),d}]$ where*

$$a_{(c,i),d} = \begin{cases} 0, & \text{if } (c,i) = (*,0), \\ i, & \text{if } d = c, \\ a_{\pi(c,i),d}, & \text{if } d \neq c. \end{cases} \tag{1}$$

Lemma 1. *The function $\theta : \mathcal{T}_{n,k} \to \mathcal{A}_{n,k}$ is a surjection.*

Proof. The set of complete perfect phylogeny trees corresponding to a matrix $A \in \mathcal{A}_{n,k}$ is exactly the set of spanning trees of $G(A)$ rooted at $v_{(*,0)}$. This set is nonempty as by Definition 5, $G(A)$ is connected and thus has at least one spanning tree for any $A \in \mathcal{A}_{n,k}$. □

We now have the following convenient parameterization of the problem. We define matrix $A_i = [a^i_{(d,j),c}]$ as

$$a^i_{(d,j),c} = \begin{cases} 1, & \text{if } a_{(d,j),c} = i, \\ 0, & \text{otherwise.} \end{cases}$$

Note that $\sum_{i=0}^{k-1} iA_i = A$. Since each sample is a mixture of the vertices of T, captured by the complete perfect phylogeny matrix A, with proportions defined in the usage matrix U, the observed frequency tensor $\mathcal{F} = [F_i]$ satisfies

$$F_i = UA_i \tag{2}$$

for all states $i \in \{0, \ldots, k-1\}$. Assuming no errors in $\mathcal{F}$, our goal is thus to find $U \in \mathcal{U}_{m,n,k}$ and $A \in \mathcal{A}_{n,k}$ satisfying (2). We thus may restate Problem 1 as follows.

Problem 2 (Perfect Phylogeny Mixture Deconvolution Problem (PPMDP))**.** *Given an $k \times m \times n$ frequency tensor $\mathcal{F} = [F_i] = [[f_{p,(c,i)}]]$, find a n, k-complete perfect phylogeny tree T and a m, n, k-usage matrix $U = [u_{p,(c,i)}]$ such that $F_i = UA_i$ for all $i \in \{0, \ldots, k-1\}$ where $A = \theta(T)$.*

B.3.3 Uniqueness of U given $\mathcal{F}$ and T

Remarkably, $\mathcal{F} = [F_i]$ and $A \in \mathcal{A}_{n,k}$ *uniquely* define the matrix U such that $F_i = UA_i$ for all states i as we prove in the following.

We start by defining a set of $(n(k-1)+1) \times nk$ binary matrices $B_{n,k}$ that are in 1-1 correspondence to $A_{n,k}$. We do so by defining the undirected graph $H(B)$ for a matrix $B \in \{0,1\}^{(n(k-1)+1)\times nk}$. The vertices of $H(B)$ correspond to the rows of B and there is an edge in $H(B)$ if and only if the two corresponding state vectors differ at exactly two positions, i.e. have Hamming distance 2. Formally, we define a *binary n, k-complete perfect phylogeny matrix* as follows.

Definition 7. *A matrix $B = [b_{(d,j),(c,i)}] \in \{0,1\}^{(n(k-1)+1)\times nk}$ matrix is a* binary n, k-complete perfect phylogeny matrix *provided*

- $\sum_{c=1}^{n} \sum_{i=0}^{k-1} b_{(d,j),(c,i)} = n$ *where* $(d, j) \in [n] \times [k-1]$,
- $\sum_{c=1}^{n} b_{(d,j),(c,i)} = 1$ *where* $(d, j) \in [n] \times [k-1]$ *and* $i \in \{0, \ldots, k-1\}$,
- $b_{(*,0),(c,0)} = 1$ *where* $c \in [n]$,
- $b_{(c,i),(c,i)} = 1$ *for all* $(c, i) \in [k] \times [n-1]$ *and*
- $H(B)$ *is connected.*

Let $\mathcal{B}_{n,k}$ be the set of all binary n, k-complete perfect phylogeny matrices.

We now define the following function ψ that maps a complete perfect phylogeny matrix A to a binary matrix.

Definition 8. *The function ψ maps a complete perfect phylogeny matrix $A \in \mathcal{A}_{n,k}$ to the binary matrix $\psi(A) = B = [A_0 \ldots A_{k-1}]$.*

We show that $\psi(A)$ is a binary n, k-complete perfect perfect phylogeny matrix for each $A \in \mathcal{A}_{n,k}$, and that ψ is in fact a bijection.

Lemma 2. *The function ψ is a bijection between $\mathcal{A}_{n,k} \to \mathcal{B}_{n,k}$.*

Proof. Let $A \in \mathcal{A}_{n,k}$. We claim that $B = \psi(A) = [A_0 \dots A_{k-1}] \in \mathcal{B}_{n,k}$. Recall that $A_i = [a^i_{(d,j),c}]$ where

$$a^i_{(d,j),c} = \begin{cases} 1, & \text{if } a_{(d,j),c} = i, \\ 0, & \text{otherwise.} \end{cases}$$

Therefore $a^i_{(d,j),c} = b_{(d,j),(c,i)}$. We thus have that $\sum_{c=1}^{n} \sum_{i=0}^{k-1} b_{(d,j),(c,i)} = n$ and $\sum_{c=1}^{n} b_{(d,j),(c,i)} = 1$ where $i \in \{0, \dots, k-1\}$. Moreover, because $a_{(*,0),d} = 0$ for all $d \in [n]$, we have that $b_{(*,0),(d,0)} = 1$. Also, as $a_{(c,i),c} = i$, we have $b_{(c,i),(c,i)} = 1$. Furthermore, $G(A)$ and $H(B)$ are isomorphic with $u_{(c,i)} \leftrightarrow v_{(c,i)}$ where $u_{(c,i)} \in V(G)$, $v_{(c,i)} \in V(H)$. Thus, $H(B)$ is connected as $G(A)$ is connected. Hence, $A \in \mathcal{A}_{n,k}$.

Let $B = [b_{(d,j),(c,i)}] \in \mathcal{B}_{n,k}$. We claim that $A = [a_{(d,j),c}] \in \mathcal{A}_{n,k}$ where $a_{(d,j),c} = i$ such that $b_{(d,j),(c,i)} = 1$. Since $b_{(*,0),(c,0)} = 1$ where $c \in [n]$, we have that $a_{(*,0),c} = 0$. Moreover, as $b_{(c,i),(c,i)} = 1$ for all $(c,i) \in [k] \times [n-1]$, we have that $a_{(c,i),c} = i$. Furthermore, $H(B)$ and $G(A)$ are isomorphic with $u_{(c,i)} \leftrightarrow v_{(c,i)}$ for $u_{(c,i)} \in V(G)$, $v_{(c,i)} \in V(H)$. Thus, $G(A)$ is connected as $H(B)$ is connected. Hence, $B \in \mathcal{B}_{n,k}$. □

We flatten frequency tensor $\mathcal{F} = [F_i]$ into the $m \times nk$ matrix $F = [F_0 \dots F_{k-1}]$ and prove that the problem is equivalent to factorizing F into U and B.

Lemma 3. *Let $\mathcal{F} = [F_i]$ be frequency tensor and let $U \in \mathcal{U}_{m,n,k}$ be a usage matrix. There exists a matrix $A \in \mathcal{A}_{n,k}$ such that $F_i = UA_i$ for all states $i \in \{0, \dots, k-1\}$ if and only if there exists a matrix $B \in \mathcal{B}_{n,k}$ such that*

$$F = [F_0 \dots F_{k-1}] = UB. \tag{3}$$

Proof. By Lemma 2, let $A \in \mathcal{A}_{n,k}$ and $B \in \mathcal{B}_{n,k}$ be corresponding matrices. Note that $B = \psi(A) = [A_0 \dots A_{k-1}]$ where $A_i = [a^i_{(d,j),c}]$ such that

$$a^i_{(d,j),c} = \begin{cases} 1, & \text{if } a_{(d,j),c} = i, \\ 0, & \text{otherwise.} \end{cases}$$

Since $a^i_{(d,j),c} = b_{(d,j),(c,i)}$, we have

$$f_{p,(c,i)} = u_{p,(*,0)} a^i_{(*,0),c)} + \sum_{d=1}^{n} \sum_{j=1}^{k-1} u_{p,(d,j)} a^i_{(d,j),c} = u_{p,(*,0)} b_{(*,0),(c,i)} + \sum_{d=1}^{n} \sum_{j=1}^{k-1} u_{p,(d,j)} b_{(d,j),(c,i)}$$

for all $p \in [m]$, $c \in [n]$ and $i \in \{0, \dots, k-1\}$. □

Hence, we may restate Problem 1 as a single matrix factorization problem.

Problem 3 (Perfect Phylogeny Mixture Deconvolution Problem (PPMDP))**.** *Given an $k \times m \times n$ frequency tensor $\mathcal{F} = [F_i] = [[f_{p,(c,i)}]]$, find a n,k-complete perfect phylogeny tree T and a m,n,k-usage matrix $U = [u_{p,(c,i)}]$ such that $F = UB$ where $F = [F_0 \dots F_{k-1}]$ and $B = \psi(\theta(T))$.*

We now have all the ingredients to show that $\mathcal{F}$ and T uniquely define a matrix U. We first show that any matrix $B \in \mathcal{B}_{n,k}$ has full row rank.

Lemma 4. *Any matrix $B \in \mathcal{B}_{n,k}$ has row rank $n(k-1)+1$.*

Proof. By Definition 7, we have that $B = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ C & D \end{pmatrix}$ where C has dimensions $n(k-1) \times n$, D has dimensions $n(k-1) \times n(k-1)$, $\mathbf{1}$ is the $1 \times n$ matrix of all 1-s and $\mathbf{0}$ is the $1 \times n(k-1)$ matrix of all 0-s. We show that the square submatrix $D = [b_{(d,j),(c,i)}]$, where $(c,i),(d,j) \in [n] \times [k-1]$, has full rank by performing Gaussian elimination according to a breadth-first search on $H(B)$, starting from the all-zero ancestor $v_{(*,0)}$. Let $\ell(v)$ denote the breadth-first search (BFS) level of vertex $v \in V(H(B))$. Note that $\ell(v_{(*,0)}) = 0$. We claim that this process results in the $n(k-1) \times n(k-1)$ identity matrix I where row $\mathbf{i}_{(c,i)}$ corresponds to vertex $v_{(c,i)} \in V(H(B)) \setminus \{v_{(*,0)}\}$.

We show this constructively by induction on the BFS level l. The claim is that at BFS level l all rows $\mathbf{i}_{(c,i)}$ where $\ell(v_{(c,i)}) \leq l$ have been generated from D using elementary row operations. Initially, at $l = 1$, for each vertex $v_{(d,j)}$ with BFS level $\ell(v_{(d,j)}) = l = 1$ it holds that $b_{(d,j),(d,j)} = 1$ and $b_{(d,j),(c,i)} = 0$ for all $(c,i) \in [n] \times [k-1] \setminus \{(d,j)\}$. Therefore the vertices $v_{(d,j)}$ at BFS level 1 correspond directly to rows $\mathbf{i}_{(d,j)}$ of I. At every iteration $l > 1$, we generate, using elementary row operations, the rows $\mathbf{i}_{(d,j)}$ of I that correspond to vertices $v_{(d,j)}$ at BFS level $\ell(v_{(d,j)}) = l$. In order to obtain row $\mathbf{i}_{(d,j)}$ of I, we subtract from $\mathbf{b}_{(d,j)}$ the rows $\mathbf{i}_{(c,i)}$ where $b_{(d,j),(c,i)} = 1$ and $(c,i) \neq (d,j)$. Observe that by Definition 7, $H(B)$ is connected and that every character-state pair (c,i) must have been introduced by vertex $v_{(c,i)}$, which must on a path to the root $v_{(*,0)}$ from $v_{(d,j)}$. Hence, $\ell(v_{(c,i)}) < l$ for every $(c,i) \in [n] \times [k-1] \setminus \{(d,j)\}$ where $b_{(d,j),(c,i)} = 1$. Since the corresponding vertices $v_{(c,i)}$ are at a BFS level strictly smaller than l, the corresponding rows $\mathbf{i}_{(c,i)}$ have already been generated by the induction hypothesis. Therefore at the final iteration, we obtain the identity matrix I from D using elementary row operations. It thus follows that D is full rank.

Since D is full rank, the row rank of $\begin{pmatrix} C & D \end{pmatrix}$ equals the rank of D, which is $n(k-1)$. Furthermore, the first row $\begin{pmatrix} \mathbf{1} & \mathbf{0} \end{pmatrix}$ of B cannot be expressed as a linear combination of $\begin{pmatrix} C & D \end{pmatrix}$. This implies that the row rank of B is $n(k-1)+1$. □

This means that given the $m \times nk$ frequency matrix F and the $(n(k-1)+1) \times nk$ binary complete perfect phylogeny matrix B, there exists a *unique* $m \times (n(k-1)+1)$ matrix U such that (3) holds, i.e. $U = FB^{-1}$ where the $nk \times (n(k-1)+1)$ matrix B^{-1} is the right inverse of B such that $BB^{-1} = I$ where I is the $(n(k-1)+1) \times (n(k-1)+1)$ identity matrix. Given F and T, we now define the unique matrix $U = [u_{p,(c,i)}]$ without explicitly computing the right inverse of B. We do this using the notion of a *descendant set* of character-state pairs (c,i). The descendant set $D_{(c,i)}$ is the set of states for character c that are descendants of $v_{(c,i)}$ in T. Formally, $D_{(c,i)} = \{j \mid (c,i) \prec_T (c,j)\}$. Recall that $T_{(c,i)}$ is the subtree of T consisting of all vertices that have state i for character c, and that $\overline{T}_{(c,i)}$ the subtree rooted at vertex $v_{(c,i)}$. The descendant set of a character precisely determines the relationship between $\overline{T}_{(c,i)}$ and $T_{(c,i)}$; namely we have

$\overline{T}_{(c,i)} = \bigcup_{l \in D_{(c,i)}} T_{(c,l)}$. In the two-state ($k = 2$) case, we have that $T_{(c,1)} = \overline{T}_{(c,1)}$ (see Figure S1F). Recall that $f_{p,(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}$. We define the *cumulative frequency* $f_p^+(D_{(c,i)}) = \sum_{l \in D_{(c,i)}} f_{p,(c,l)}$. In the following lemma we show that given $T \in \mathcal{T}_{n,k}$, the cumulative frequencies $f_p^+(D_{(c,i)})$ for the descendant sets defined by T *uniquely* determine the usage matrix U.

Lemma 5. *Let $T \in \mathcal{T}_{n,k}$ and $\mathcal{F} = [[f_{p,(c,i)}]]$ be a frequency tensor. For a character-state pair (c,i) and sample p, let*

$$u_{p,(c,i)} = f_p^+(D_{(c,i)}) - \sum_{(d,j) \in \delta(c,i)} f_p^+(D_{(d,j)}). \tag{4}$$

Then $U = [u_{p,(c,i)}]$ is the unique matrix whose entries satisfy $f_{p,(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}$.

Proof. Let (c,i) be a character-state pair and p be a sample. Let $A = \theta(T)$ and $B = \psi(A)$. Recall that $T_{(c,i)}$ is the set of vertices $\{v_{(d,j)} \mid b_{(d,j),(c,i)} = 1\}$. Note that by definition, the vertices of $T_{(c,i)}$ induce a connected subtree in T. We thus need to show that

$$f_{p,(c,i)} = u_{p,(*,0)} b_{(d,j),(c,i)} + \sum_{d=1}^{n} \sum_{j=1}^{k-1} u_{p,(d,j)} b_{(d,j),(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}.$$

We introduce the following shorthand $\Delta(c,i) = \bigcup_{(d,j) \in T_{(c,i)}} \delta(d,j) \setminus T_{(c,i)}$, which is the set of vertices $\{v_{(d,j)}\}$ that are not in $T_{(c,i)}$ but whose parent $v_{\pi(d,j)}$ is in $T_{(c,i)}$. Thus,

$$\begin{aligned} f_{p,(c,i)} &= \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)} \\ &= \sum_{(d,j) \in T_{(c,i)}} \left(f_p^+(D_{(d,j)}) - \sum_{(e,l) \in \delta(d,j)} f_p^+(D_{(e,l)}) \right). \end{aligned}$$

Observe that in the equation above, for every $(d,j) \in T_{(c,i)} \setminus \{(c,i)\}$, there are two terms $f_p^+(D_{(d,j)})$ and $-f_p^+(D_{(d,j)})$, which consequently cancel out. The remaining terms are $f_p^+(D_{(c,i)})$, and $-f_p(D_{(c,l)})$ for each $(c,l) \in \Delta(c,i)$. In addition, we have that the state trees $\{D_{(c,l)}\}$ corresponding to the elements of $(c,l) \in \Delta(c,i)$ are pairwise disjoint. Moreover, $D_{(c,i)} \setminus \{i\} = \bigcup_{(c,l) \in \Delta(c,i)} D_{(c,l)}$. Thus,

$$\begin{aligned} f_{p,(c,i)} &= f_p^+(D_{(c,i)}) - \sum_{(c,l) \in \Delta(c,i)} f_p^+(D_{(c,l)}) \\ &= f_{p,(c,i)} + \sum_{(c,l) \in \Delta(c,i)} f_p^+(D_{(c,l)}) - \sum_{(c,l) \in \Delta(c,i)} f_p^+(D_{(c,l)}) \\ &= f_{p,(c,i)}. \end{aligned}$$

By Lemma 4, the equation $F = UB$ has only one solution given F and B. Thus $U = [u_{p,(c,i)}]$ is the unique matrix such that $f_{p,(c,i)} = \sum_{(d,j) \in T_{(c,i)}} u_{p,(d,j)}$ for all samples p and character-state pairs (c,i). □

B.3.4 Combinatorial Characterization of the PPMDP

We say that T *generates* $\mathcal{F}$ if the corresponding matrix U, defined by (4), is a usage matrix. It turns out that positivity of the values $u_{p,(c,i)}$ is a necessary and sufficient condition for T to generate $\mathcal{F}$. We show this in the following theorem.

(Main Text) Theorem 1. *A complete perfect phylogeny tree T generates $\mathcal{F}$ if and only if*

$$f_p^+(D_{(c,i)}) - \sum_{(d,j)\in\delta(c,i)} f_p^+(D_{(d,j)}) \geq 0 \tag{MSSC}$$

for all character-state pairs (c,i) and samples p.

Proof. ($\Rightarrow$) We start by proving the forward direction. Let $F = [F_0 \dots F_{k-1}]$, T be a tree that generates F, $B = \psi(\theta(T))$ be the corresponding binary matrix of T and let $U = [u_{p,(c,i)}]$ be the corresponding usage matrix. Since T generates F, we have that $u_{p,(c,i)} \geq 0$ for all character-state pairs (c,i) and samples p. By Lemma 5, (MSSC) thus holds.

($\Leftarrow$) As for the reverse direction, we need to show that if (MSSC) is met, the matrix U as defined in Lemma 5 is a usage matrix. Let $p \in [m]$. We prove this direction by showing that (i) $u_{p,(c,i)} \geq 0$ for all character-state pairs (c,i), and (ii) $u_{p,(*,0)} + \sum_{c=1}^{n}\sum_{i=1}^{k-1} u_{p,(c,i)} = 1$.

(i) By Lemma 5 and (MSSC), we have that $u_{p,(c,i)} \geq 0$.

(ii) We now have

$$u_{p,(*,0)} + \sum_{c=1}^{n}\sum_{i=1}^{k-1} u_{p,(c,i)} = f_p^+(D_{(*,0)}) - \sum_{(d,j)\in\delta(*,0)} f_p^+(D_{(d,j)}) + \sum_{c=1}^{n}\sum_{i=1}^{k-1}\left(f_p^+(D_{(c,i)}) - \sum_{(d,j)\in\delta(c,i)} f_p^+(D_{(d,j)})\right).$$

Observe that for each $(c,i) \in [n]\times[k-1]$ there are exactly two terms in the above equation: $+f_p^+(D_{(c,i)})$ when $v_{(c,i)}$ is considered as a parent and $-f_p^+(D_{(c,i)})$ when $v_{(c,i)}$ is considered as a child. Hence, all these terms cancel out. Since $D_{(*,0)} = \{0,\dots,k-1\}$ and $\sum_{i=0}^{k-1} f_{p,(c,i)} = 1$, we have that $f_p^+(D_{(*,0)}) = 1$. Thus, $u_{p,(*,0)} + \sum_{c=1}^{n}\sum_{i=1}^{k-1} u_{p,(c,i)} = 1$.

□

Although the sets $D_{(c,i)}$ are *a priori* unknown, we show that the following condition (MSAC) must hold for any tree T that generates $\mathcal{F}$ where $(c,i) \prec_T (d,j)$.

Definition 9. *Let (c,i) and (d,j) be distinct character-state pairs and let $D_{(c,i)}, D_{(d,j)} \subseteq \{0,\dots,k-1\}$. A pair $(D_{(c,i)}, D_{(d,j)})$ is a* valid descendant set pair *provided*

$$f_p^+(D_{(c,i)}) - f_p^+(D_{(d,j)}) \geq 0 \tag{MSAC}$$

for all samples p; and additionally if $c = d$ then $D_{(c,j)} \subsetneq D_{(c,i)}$.

It turns out that there are potentially many valid descendant set pairs as shown by the following lemma.

Lemma 6. *Let $(D_{(c,i)}, D_{(d,j)})$ be a valid descendant set pair. If $D_{(c,i)} \subseteq D'_{(c,i)}$ and $D'_{(d,j)} \subseteq D_{(d,j)}$ then $(D'_{(c,i)}, D'_{(d,j)})$ is a valid descendant set pair.*

Proof. Let $D_{(c,i)} \subseteq D'_{(c,i)}$ and $D'_{(d,j)} \subseteq D_{(d,j)}$. Let $p \in [m]$. Since $f_{p,(c,i)} \geq 0$ (by Definition 1), $D_{(c,i)} \subseteq D'_{(c,i)}$ and $D'_{(d,j)} \subseteq D_{(d,j)}$, we have that $f_p^+(D'_{(c,i)}) \geq f_p^+(D_{(c,i)})$ and $f_p^+(D_{(d,j)} \geq f_p^+(D'_{(d,l)})$. Moreover, if $c = d$ then $D_{(d,j)} \subsetneq D_{(c,i)}$ and thus $D'_{(d,j)} \subsetneq D'_{(c,i)}$. Hence, $(D'_{(c,i)}, D'_{(d,j)})$ is a valid descendant set pair. □

Since $v_{(*,0)}$ is the all-zero ancestor, we have the following corollary that describes the *extreme* valid descendant set pair.

Corollary 1. *Let T be a complete perfect phylogeny tree that generates $\mathcal{F}$. If $(c,i) \prec_T (d,j)$ and $(c,i) \neq (c,0)$ then $D_{(c,i)} = [k-1]$ and $D_{(d,j)} = \{j\}$ form a valid descendant set pair $(D_{(c,i)}, D_{(d,j)})$.*

The following proposition shows that (MSAC) is a necessary condition to solutions of the PPMDP.

(Main Text) Proposition 1. *Let T be a complete perfect phylogeny tree that generates $\mathcal{F}$. If $(c,i) \prec_T (d,j)$ then there exist a valid descendant set pair $(D_{(c,i)}, D_{(d,j)})$.*

Proof. Let $(c,i) \prec_T (d,j)$. Let $P = v_{(c_1,i_1)}, \ldots, v_{(c_t,i_t)}$ where $v_{(c_1,i_1)} = v_{(c,i)}$ and $v_{(c_t,i_t)} = v_{(d,j)}$ be the unique path from $v_{(c,i)}$ to $v_{(d,j)}$ in T. Assume for a contradiction that there exists no valid descendant set pair $(D_{(c,i)}, D_{(d,j)})$. By Corollary 1, we thus have that $D_{(c,i)} = [k-1]$ and $D_{(d,j)} = \{j\}$ do not form a valid descendant set pair $(D_{(c,i)}, D_{(d,j)})$. Now, if $c = d$, we would have a contradiction as $\{j\} \subseteq [k-1]$. Therefore, $c \neq d$ and $f_p^+(D_{(c,i)}) - f_p^+(D_{(d,j)}) < 0$. By Theorem 1, we have that $f_p^+(D_{(c_l,i_l)}) - f_p^+(D_{(c_{l+1},i_{l+1})}) \geq 0$ for all $1 \leq l < t$. Therefore, $f_p^+(D_{(c_1,i_1)}) - f_p^+(D_{(c_t,i_t)}) \geq 0$, which leads to a contradiction. □

We now define the multi-state ancestry graph $G_{\mathcal{F}}$ whose vertices correspond to character-state pairs and whose edges correspond to valid descendant state pairs.

Definition 10. *The* multi-state ancestry graph *$G_{\mathcal{F}}$ of the frequency tensor $\mathcal{F}$ is an edge-labeled, directed multi-graph $G_{\mathcal{F}} = (V, E)$ whose vertices $v_{(c,i)}$ correspond to character-state pairs (c,i) and whose multi-edges are $(v_{(c,i)}, v_{(d,j)})$ for all valid descendant set pairs $(D_{(c,i)}, D_{(d,j)})$.*

Note that $v_{(1,0)}, \ldots, v_{(n,0)}$ all refer to the same root vertex $v_{(*,0)}$. We use the labels of the multi-edges to restrict the set of allowed spanning trees by defining a threading as follows.

Definition 11. *A rooted subtree T of $G_{\mathcal{F}} = (V, E)$ is a* threaded tree *provided (1) for every pair of adjacent edges $(v_{(c,i)}, v_{(d,j)}), (v_{(d,j)}, v_{(e,l)}) \in E(T)$ with corresponding labels $(D_{(c,i)}, D_{(d,j)})$ and $(D'_{(d,j)}, D'_{(e,l)})$, it holds that $D_{(d,j)} = D'_{(d,j)}$, and (2) for every pair of vertices $v_{(c,i)}, v_{(c,j)} \in V(T)$ it holds that $D_{(c,j)} \subseteq D_{(c,i)}$ if and only if $(c,i) \prec_T (c,j)$.*

We now prove that solutions of an PPMDP instance $\mathcal{F}$ correspond to threaded spanning trees of the ancestry graph $G_{\mathcal{F}}$.

(Main Text) Theorem 2. *A complete perfect phylogeny tree T generates $\mathcal{F}$ if and only if T is a threaded spanning tree of the ancestry graph $G_{\mathcal{F}}$ such that* (MSSC) *holds.*

Proof. ($\Rightarrow$) Let T be a complete perfect phylogeny tree generating $\mathcal{F}$. We claim that T is a threaded spanning tree of the ancestry graph $G_{\mathcal{F}}$. We start by showing that every edge $(v_{(c,i)}, v_{(d,j)}) \in E(T)$ is an edge of $G_{\mathcal{F}}$ labeled by $(D_{(c,i)}, D_{(d,j)})$. By Theorem 1, we have that $f_p^+(D_{(c,i)}) - \sum_{(d,j)\in\delta(c,i)} f_p^+(D_{(d,j)}) \geq 0$ for all character-state pairs (c,i) and samples p. Let (c,i) be a character-state pair. By definition, we have

that $D_{(c,j)} \subsetneq D_{(c,i)}$ for all $(c,j) \in \delta(c,i)$. Moreover, $i \in D_{(c,i)}$ and $j \in D_{(d,j)}$ for all character-state pairs $(d,j) \in \delta(c,i)$. Thus, $(D_{(c,i)}, D_{(d,j)})$ is a valid descendant set pair for all character-state pairs $(d,j) \in \delta(c,i)$. Hence, every edge $(v_{(c,i)}, v_{(d,j)})$ of T is an edge of $G_{\mathcal{F}}$ labeled by the valid descendant set pair $(D_{(c,i)}, D_{(d,j)})$. Thus, T is a tree of G. Next, we show that T is a *threaded* spanning tree.

1. By definition of D, we have that every pair of adjacent edges $(v_{(c,i)}, v_{(d,j)}), (v_{(d,j)}, v_{(e,l)}) \in E(T)$ is labeled by $(D_{(c,i)}, D_{(d,j)})$ and $(D_{(d,j)}, D_{(e,l)})$, respectively.

2. By definition of D, we have that for every edge $(v_{(c,i)}, v_{(d,j)}) \in E(T)$ labeled by $(D_{(c,i)}, D_{(d,j)})$, it holds that $D_{(c,i)} = \{l \mid (c,i) \prec_T (c,l)\}$ and $D_{(d,j)} = \{l \mid (d,j) \prec_T (d,l)\}$. Hence, $(c,i) \prec_T (c,j)$ if and only if $D_{(c,j)} \subseteq D_{(c,i)}$.

The conditions of Definition 9 are thus met. Therefore, T is a threaded spanning tree of $G_{\mathcal{F}}$.

($\Leftarrow$) Let T be a threaded spanning tree of the ancestry graph $G_{\mathcal{F}}$ such that $f_p^+(D_{(c,i)}) - \sum_{(d,j)\in\delta(c,i)} f_p^+(D_{(d,j)}) \geq 0$ for all character-state pairs (c,i) and samples p. Observe that T is a complete perfect phylogeny tree. By condition (1) of Definition 9, we have that for all character-state pairs $(c,0) \neq (*,0)$ and $(d,j) \in \delta(c,i)$, adjacent edges $(v_{\pi(c,i)}, v_{(c,i)})$, $(v_{(c,i)}, v_{(d,j)})$ are labeled by $(D_{\pi(c,i)}, D_{(c,i)})$ and $(D_{(c,i)}, D_{(d,j)})$—where $\pi(c,i)$ is the parent character-state pair of (c,i). Hence, we may use $D_{(c,i)}$ to unambiguously denote the descendant state set of the character-state pair (c,i). Moreover, by condition (2) of Definition 9 and the fact that T is a spanning tree of $G_{\mathcal{F}}$, we have that D is defined in the same way as in Theorem 1. By Theorem 1, we thus have that T generates $\mathcal{F}$. □

B.3.5 Complexity

In previous work, we have shown that the problem is NP-complete for general m [3]. An open question was the hardness for constant m, which we resolve with the following lemma.

Theorem 1. *The PPMDP is NP-complete even for $m = 2$ and $k = 2$.*

Proof. Clearly, the problem is in NP—given matrices U and A we can check in polynomial time whether $F_i = UA_i$ for all $i \in \{0, \ldots, k-1\}$. We show NP-hardness by reduction from SUBSET SUM, which, given non-negative integers $B = \{b_1, \ldots, b_t\}$ and d, asks whether there exists a subset $B' \subseteq B$ whose sum equals d. This problem is NP-complete [6].

Let $e = \sum_{\ell=1}^{t} c_\ell$. Without loss of generality assume that $e > 0$, $b_\ell < d$ for all $\ell \in [t]$ and that $b_\ell \leq b_{\ell+1}$ for all $\ell \in [t-1]$. The corresponding frequency tensor $\mathcal{F} = [F_0\, F_1]$ is then defined as follows.

$$F_0 = \frac{1}{e}\begin{pmatrix} e-d & d & e-b_1 & e-b_2 & \ldots & e-b_t \\ d & e-d & e-t\epsilon & e-(t-1)\epsilon & \ldots & e-\epsilon \end{pmatrix}$$

$$F_1 = \frac{1}{e}\begin{pmatrix} d & e-d & b_1 & b_2 & \ldots & b_t \\ e-d & d & t\epsilon & (t-1)\epsilon & \ldots & \epsilon \end{pmatrix}.$$

Note that $\mathcal{F}$ is a $k \times m \times n$ tensor where $k = 2$, $m = 2$ and $n = t + 2$. Also note that the normalization factor $\frac{1}{e}$ ensures that each $f_{p,(c,i)} \in [0, 1]$ and that $f_{p,(c,0)} + f_{p,(c,1)} = 1$ for all $p \in [m]$ and $c \in [n]$. Clearly $\mathcal{F}$ can be obtained in polynomial time from a SUBSET SUM instance. By construction, the ancestry graph $G_{\mathcal{F}}$ consists of a root vertex $v_{(*,0)}$ with outgoing edges to all other vertices $\{d, e-d, b_1, \ldots, b_t\}$. In addition, these vertices induce a complete bipartite graph with vertex sets $\{d, e-d\}$ and $\{b_1, \ldots, b_t\}$. See Figure S6 for an illustration.

We claim that there exist $U \in \mathcal{U}_{2,t+2,2}$ and $A \in \mathcal{A}_{t+2,2}$ such that $F_0 = UA_0$ and $F_1 = UA_1$ if and only if there exists a subset $B' \subseteq B$ whose sum is d. Equivalently, by Theorem 2, we claim that $G_{\mathcal{F}}$ admits a threaded spanning tree T satisfying (MSSC) if and only if B has a subset B' whose sum is d.

We start by proving the forward direction. Since $e = \sum_{\ell=1}^{t} b_\ell$ and T is a threaded spanning tree, the sum of the children $\delta(d)$ equals d and the sum of the children $\delta(e-d)$ equals $e-d$. Therefore $B' = \delta(d)$ is a subset of B whose sum is d. As for the reverse direction, we have that the sum of B' equals d and thus that the sum of $B \setminus B'$ equals $e-d$. The corresponding threaded spanning tree T where $\delta(*, 0) = \{d, e-d\}$, $\delta(d) = B'$ and $\delta(e-d) = B \setminus B'$ is therefore a threaded spanning tree of $G_{\mathcal{F}}$ that satisfies (MSSC). □

The result above settles the outstanding question posed in [3] of fixed-parameter tractability in m.

Corollary 2. *PPMDP is not fixed-parameter tractable in m.*

B.4 Cladistic Perfect Phylogeny Mixture Deconvolution Problem

In this section, we consider a restricted version with cladistic characters. We present an enumeration algorithm in Section B.4.1.

In the case of cladistic characters we are given a set $\mathcal{S} = \{S_c \mid c \in [n]\}$ of state trees for each character. The vertex set of a state tree S_c is $\{0, \ldots, k-1\}$, and the edges describe the relationships of the states of character c. A perfect phylogeny T is *consistent* with $\mathcal{S}$ provided $i \prec_{S_c} j$ if and only if $(c, i) \prec_T (c, j)$ for all characters c and states i, j. In the cladistic PPMDP we seek to find a complete perfect phylogeny tree T that generates $\mathcal{F}$ and is consistent with $\mathcal{S}$. The cladistic multi-state perfect phylogeny problem reduces to the binary case and is polynomial-time decidable [4]. Thus, it is not surprising that the PPMDP also simplifies in the cladistic case. In particular, the set $\mathcal{S}$ of states trees determines the set of descendant sets as $D_{(c,i)} = \{j \mid i \prec_{S_c} j\}$. Therefore the cladistic multi-state ancestry graph $G_{(\mathcal{F},\mathcal{S})}$ is a simple graph with edges $(v_{(c,i)}, v_{(d,j)})$ labeled by $(D_{(c,i)}, D_{(d,j)})$ provided $c \neq d$ and (MSAC) holds. Moreover, for each character c there is an edge $(v_{(c,i)}, v_{(c,j)})$ provided state i is the parent of state j in S_c. Solutions of the cladistic PPMDP correspond to threaded spanning trees in $G_{(\mathcal{F},\mathcal{S})}$ as shown by the following proposition.

Proposition 1. *A complete perfect phylogeny tree T generates $\mathcal{F}$ and is consistent with $\mathcal{S}$ if and only if T is a threaded spanning tree of the cladistic ancestry graph $G_{(\mathcal{F},\mathcal{S})}$ such that* (MSSC) *holds.*

Proof. Let T be a threaded spanning tree of the ancestry graph $G_{(\mathcal{F},\mathcal{S})}$. We claim that T is consistent with $\mathcal{S}$, i.e. $i \prec_{S_c} j$ if and only if $(c, i) \prec_T (c, j)$ for all characters c and states i, j. By condition (1) of

Definition 9, we have that for all character-state pairs $(c,i) \neq (*,0)$ and $(d,j) \in \delta(c,i)$, adjacent edges $(v_{\pi(c,i)}, v_{(c,i)})$, $(v_{(c,i)}, v_{(d,j)})$ are labeled by $(D_{\pi(c,i)}, D_{(c,i)})$ and $(D_{(c,i)}, D_{(d,j)})$, respectively—where $\pi(c,i)$ is the parent character-state pair of (c,i). By definition, we have $D_{(c,i)} = \{l \mid i \prec_{S_c} l\}$ for each character-state pair (c,i). By condition (2) of Definition 9 and the fact that T is a spanning tree of $G_{(\mathcal{F},\mathcal{S})}$, we have $D_{(c,i)} = \{l \mid (c,i) \prec_T (c,l)\}$. The lemma now follows. □

B.4.1 Enumeration Algorithm for the Cladistic-PPMDP

We now describe an algorithm for enumerating all trees T that are consistent with the given state trees $\mathcal{S}$ and generate the given frequencies $\mathcal{F}$. The crucial observation is that any subtree of a consistent, threaded spanning tree T that satisfies (MSSC) must itself be satisfy (MSSC) and be consistent and threaded. A subtree T' is *consistent* if it is rooted at $v_{(*,0)}$, and for each character c the set of states $\{i \mid v_{(c,i)} \in V(T')\}$ induces a connected subtree in S_c. We can thus constructively grow consistent, threaded trees that satisfy (MSSC) by maintaining the following invariant.

Invariant 1. *Let tree T be the partially constructed tree. It holds that (1) for each $v_{(c,i)} \in V(T) \setminus \{v_{(*,0)}\}$ and parent $\pi(i)$ of i in S_c, the vertex $v_{(c,\pi(i))}$ is the first vertex labeled by character c on the unique path from $v_{(*,0)}$ to $v_{(c,i)}$; and (2) for each vertex $v_{(c,i)} \in V(T)$,* (MSSC) *holds for T and $\mathcal{F}$.*

We maintain a subset of edges $H \subseteq E(G_{(\mathcal{F},\mathcal{S})})$ called the *frontier* that can be used to extend a partial tree T such that the following invariant holds.

Invariant 2. *Let tree T be the partially constructed tree. For every edge $(v_{(c,i)}, v_{(d,j)}) \in H$, (1) $v_{(c,i)} \in V(T)$, (2) $v_{(d,j)} \notin V(T)$, and (3) Invariant 1 holds for T' where $E(T') = E(T) \cup \{(v_{(c,i)}, v_{(d,j)})\}$.*

Algorithm 1 gives the pseudo code of ENUMERATE described in the main text. The initial call is ENUMERATE$(G, \{v_{(*,0)}\}, \delta(*,0))$. The partial tree containing just the vertex $v_{(*,0)}$ satisfies Invariant 1. The set $\delta(*,0)$ corresponds to the set of outgoing edges from vertex $v_{(*,0)}$ of $G_{(\mathcal{F},\mathcal{S})}$, which by definition satisfies Invariant 2. Upon the addition of an edge $(v_{(c,i)}, v_{(d,j)}) \in H$ (line 5), Invariant 2 is restored by adding all outgoing edges from $v_{(d,j)}$ whose addition results in a consistent partial tree that satisfies (MSSC) (lines 8-9) and by removing all edges from H that introduce a cycle (lines 11-12) or violate (MSSC) (lines 13-14). The running time is the same as the original Gabow-Myers algorithm: $O(|V| + |E| + |E| \cdot K)$ where K is the number of spanning trees in $G_{(\mathcal{F},\mathcal{S})} = (V,E)$ (disregarding (MSSC)).

In order to extend this algorithm to the general PPMDP, we need to update the descendant sets $D_{(c,i)}$ as we grow the tree. This in turn has implications for how we maintain the frontier: for each potential frontier edge, we need to consider how its addition to T affects the descendant sets of the existing vertices of T and thereby (MSSC). We leave this extension as future work.

Noisy VAFs. We account for errors by taking as input nonempty intervals $[\underline{f}_{p,(c,i)}, \bar{f}_{p,(c,i)}]$ that contain the true frequency $f_{p,(c,i)}$ for character-state pairs (c,i) in samples p. A tree T is *valid* if there exists a

Algorithm 1: ENUMERATE(G, T, H)

Input: Ancestry graph $G_{(\mathcal{F},\mathcal{S})}$, perfect phylogeny tree T, frontier H

Output: All complete perfect phylogeny trees that generate $\mathcal{F}$ and are consistent with $\mathcal{S}$

if $H = \emptyset$ **and** $|V(T)| = |V(G)|$ **then**
    Return T
else
    while $H \neq \emptyset$ **do**
        $(v_{(c,i)}, v_{(d,j)}) \leftarrow$ POP(H)
        $E(T) \leftarrow E(T) \cup \{(v_{(c,i)}, v_{(d,j)})\}$
        $H' \leftarrow H$
        foreach $(v_{(d,j)}, v_{(e,l)}) \in E(G)$ **do**
            if $v_{(e,l)} \notin V(T)$ **and** $v_{(e,\pi(l))}$ *is the first vertex with character* e *on path from* $v_{(*,0)}$ *to* $v_{(d,j)}$ **and** $f_p^+(D_{(d,j)}) \geq f_p^+(D_{(e,l)}) + \sum_{(f,s)\in\delta(d,j)} f_p^+(D_{(f,s)})$ **then**
                PUSH$(H', (v_{(d,j)}, v_{(e,l)}))$
        foreach $(v_{(e,l)}, v_{(f,s)}) \in H'$ **do**
            if $v_{(f,s)} = v_{(d,j)}$ **then**
                Remove $(v_{(e,l)}, v_{(f,s)})$ from H'
            else if
            $v_{(e,l)} = v_{(c,i)}$ **and** $\exists p \in [m]$ *such that* $f_p^+(D_{(c,i)}) < f_p^+(D_{(f,s)}) + \sum_{(h,t)\in\delta(c,i)} f_p^+(D_{(h,t)})$ **then**
                Remove $(v_{(e,l)}, v_{(f,s)})$ from H'
        ENUMERATE(G, T, H')
        $E(T) \leftarrow E(T) \setminus \{(v_{(c,i)}, v_{(d,j)})\}$

frequency tensor $\mathcal{F}' = [[f'_{p,(c,i)}]]$ such that $\underline{f}_{p,(c,i)} \leq f'_{p,(c,i)} \leq \bar{f}_{p,(c,i)}$ and $\mathcal{F}'$ generates T—i.e. (MSSC) holds for T and $\mathcal{F}'$. A valid tree T is *maximal* if there exists no valid supertree T' of T, i.e. $E(T) \subsetneq E(T')$. The task now becomes to find the set of all maximal valid trees. We recursively define $\hat{\mathcal{F}} = [[\hat{f}_{p,(c,i)}]]$ where

$$\hat{f}_{p,(c,i)} = \max\Big\{\underline{f}_{p,(c,i)}, \sum_{(d,j)\in\delta(c,i)} \hat{f}^+_p(D_{(d,j)}) - \sum_{j\in D(c,i)\setminus\{i\}} \hat{f}_{p,(c,j)}\Big\}. \tag{5}$$

The intuition here is to satisfy (MSSC) by assigning the smallest possible values to the children. We do this bottom-up from the leaves and set $\hat{f}_{p,(c,i)} = \underline{f}_{p,(c,i)}$ for each leaf vertex $v_{(c,i)}$. It turns out that $\underline{f}_{p,(c,i)} \leq \hat{f}_{p,(c,i)} \leq \overline{f}_{p,(c,i)}$ is a necessary condition for T to be valid as shown in the following lemma.

Lemma 7. *If tree T is valid then $\underline{f}_{p,(c,i)} \leq \hat{f}_{p,(c,i)} \leq \bar{f}_{p,(c,i)}$ for all p and (c,i).*

Proof. Let T be a valid tree and let $\mathcal{F} = [[f_{p,(c,i)}]]$ be a frequency tensor that generates T such that $\underline{f}_{p,(c,i)} \leq f_{p,(c,i)} \leq \bar{f}_{p,(c,i)}$. We claim that $\underline{f}_{p,(c,i)} \leq \hat{f}_{p,(c,i)} \leq f_{p,(c,i)} \leq \bar{f}_{p,(c,i)}$. We proof this by structural induction on T by working our way up to the root.

For the base case, let $v_{(c,i)}$ be a leaf of T. That is, $\delta(c,i) = \emptyset$. Thus by definition, $\hat{f}_{p,(c,i)} = \underline{f}_{p,(c,i)}$. Therefore, $\underline{f}_{p,(c,i)} \leq \hat{f}_{p,(c,i)} \leq f_{p,(c,i)} \leq \bar{f}_{p,(c,i)}$. For the step, let $v_{(c,i)}$ be an inner vertex of T. Thus, $\delta(c,i) \neq \emptyset$. The induction hypothesis (IH) states that $\underline{f}_{p,(d,j)} \leq \hat{f}_{p,(d,j)} \leq f_{p,(d,i)} \leq \bar{f}_{p,(d,j)}$ for all descendants (d,j) of (c,i) in T. We distinguish two cases:

1. In case $\hat{f}_{p,(c,i)} = \underline{f}_{p,(c,i)}$, we have $\underline{f}_{p,(c,i)} \leq \hat{f}_{p,(c,i)} \leq f_{p,(c,i)} \leq \bar{f}_{p,(c,i)}$.
2. In case $\hat{f}_{p,(c,i)} = \sum_{(d,j)\in\delta(c,i)} \hat{f}^+_p(D_{(d,j)}) - \sum_{j\in D(c,i)\setminus\{i\}} \hat{f}_{p,(c,i)}$, we have $\hat{f}_{p,(c,i)} \geq \underline{f}_{p,(c,i)}$ by definition. By (MSSC), we have that $f_{p,(c,i)} \geq \sum_{(d,j)\in\delta(c,i)} f^+_p(D_{(d,j)}) - \sum_{j\in D(c,i)\setminus\{i\}}$. By the IH, we have $\underline{f}_{p,(d,j)} \leq \hat{f}_{p,(d,j)} \leq f_{p,(d,j)} \leq \bar{f}_{p,(d,j)}$ for all $(d,j) \in \delta(c,i)$ and $\underline{f}_{p,(c,l)} \leq \hat{f}_{p,(c,l)} \leq f_{p,(c,l)} \leq \bar{f}_{p,(c,l)}$ for all $l \in D_{(c,i)} \setminus \{i\}$. Therefore, $\hat{f}_{p,(c,i)} = \sum_{(d,j)\in\delta(c,i)} \hat{f}^+_p(D_{(d,j)}) - \sum_{j\in D(c,i)\setminus\{i\}} \hat{f}_{p,(c,i)} \leq \sum_{(d,j)\in\delta(c,i)} f^+_p(D_{(d,j)}) - \sum_{j\in D(c,i)\setminus\{i\}} f_{p,(c,i)} \leq f_{p,(c,i)}$. Hence, $\underline{f}_{p,(c,i)} \leq \hat{f}_{p,(c,i)} \leq f_{p,(c,i)} \leq \bar{f}_{p,(c,i)}$.

□

It may be the case that the frequencies $\hat{f}_{p,(c,i)}$ of a character c in a sample p do not sum to 1. Therefore, $\underline{f}_{p,(c,i)} \leq \hat{f}_{p,(c,i)} \leq \bar{f}_{p,(c,i)}$ is not a sufficient condition for T to be valid. However, if $1 - \sum_{i=1}^{k-1} \hat{f}_{p,(c,i)} \leq \bar{f}_{p,(c,0)}$ holds, T is valid as it is generated by frequency tensor $\mathcal{F}' = [[f'_{p,(c,i)}]]$ where

$$f'_{p,(c,i)} = \begin{cases} 1 - \sum_{i=1}^{k-1} \hat{f}_{p,(c,i)}, & \text{if } i = 0, \\ \hat{f}_{p,(c,i)}, & \text{otherwise.} \end{cases}$$

This is the case if $\bar{f}_{p,(c,0)} = 1$. By assuming the latter, we are able to enumerate all maximal valid trees by updating condition (2) of Invariant 1 so that (MSSC) holds for T and $\hat{\mathcal{F}}$.

Algorithm 2 gives the pseudo code of an enumeration procedure of all maximal valid trees given intervals $[l_{p,(c,i)}, u_{p,(c,i)}]$ for each character-state pair (c,i) in each sample p. The initial call is NoisyEnumerate$(G, \{v_{(*,0)}\}, \delta(*,0))$. The partial tree containing just the vertex $v_{(*,0)}$ satisfies Invariant 1. The set $\delta(*,0)$

corresponds to the set of outgoing edges from vertex $v_{(*,0)}$ of $G_{(\mathcal{F},\mathcal{S})}$, which by definition satisfies Invariant 2. Upon the addition of an edge $(v_{(c,i)}, v_{(d,j)}) \in H$ (line 5), Invariant 2 is restored by adding all outgoing edges from $v_{(d,j)}$ whose addition results in a consistent partial tree T' that satisfies (MSSC) for $\hat{\mathcal{F}}$ (lines 9-10) and by removing all edges from H that introduce a cycle (lines 12-13) or violate (MSSC) for $\hat{\mathcal{F}}$ (lines 14-15). Note that in line 13 we dropped the condition $v_{(e,l)} = v_{(c,i)}$ as the newly added edge $(v_{(c,i)}, v_{(d,j)})$ may affect the frequencies $\hat{\mathcal{F}}$ of the vertices of the current partial tree T.

Since a maximal valid tree T does not necessarily span all the vertices, it may happen that for a character c not all states in S_c are present. We say that a maximal valid tree T is *state complete* if for each vertex $v_{(c,i)}$ of T, all vertices $v_{(c,j)}$ where $j \in V(S_c)$ are also in $V(T)$. Our goal is to report all maximal valid and state-complete trees. Therefore, we post-process each maximal valid tree T and remove all vertices $v_{(c,i)}$ where there is a $j \in V(S_c)$ such that $v_{(c,j)} \notin V(T)$. The tree that we report corresponds to the connected component rooted at $v_{(*,0)}$. Since each maximal valid and state-complete tree is a partial valid tree rooted at $v_{(*,0)}$, our enumeration procedure reports all maximal valid and state-complete trees.

Algorithm 2: NoisyEnumerate(G, T, H)

Input: Ancestry graph $G_{(\mathcal{F},\mathcal{S})}$, perfect phylogeny tree T, frontier H

Output: All maximal valid perfect phylogenies that are consistent with $\mathcal{S}$

1 **if** $H = \emptyset$ **then**
2 Let T' be the subtree of T that only contains state-complete characters
3 Return T'
4 **else**
5 **while** $H \neq \emptyset$ **do**
6 $(v_{(c,i)}, v_{(d,j)}) \leftarrow$ Pop(H)
7 $E(T) \leftarrow E(T) \cup \{(v_{(c,i)}, v_{(d,j)})\}$
8 $H' \leftarrow H$
9 **foreach** $(v_{(d,j)}, v_{(e,l)}) \in E(G)$ **do**
10 **if** $v_{(e,l)} \notin V(T)$ **and** $v_{(e,\pi(l))}$ *is the first vertex with character* e *on path from* $v_{(*,0)}$ *to* $v_{(d,j)}$ **and** $\hat{f}_p^+(D_{(d,j)}) \geq \hat{f}_p^+(D_{(e,l)}) + \sum_{(f,s)\in\delta(d,j)} \hat{f}_p^+(D_{(f,s)})$ **then**
11 Push$(H', (v_{(d,j)}, v_{(e,l)}))$
12 **foreach** $(v_{(e,l)}, v_{(f,s)}) \in H'$ **do**
13 **if** $v_{(f,s)} = v_{(d,j)}$ **then**
14 Remove $(v_{(e,l)}, v_{(f,s)})$ from H'
15 **else if** $\exists p \in [m], \hat{f}_p^+(D_{(c,i)}) < \hat{f}_p^+(D_{(f,s)}) + \sum_{(h,t)\in\delta(c,i)} \hat{f}_p^+(D_{(h,t)})$ **then**
16 Remove $(v_{(e,l)}, v_{(f,s)})$ from H'
17 NoisyEnumerate(G, T, H')
18 $E(T) \leftarrow E(T) \setminus \{(v_{(c,i)}, v_{(d,j)})\}$

B.5 Multi-State Model for the Somatic Mutational Process in Cancer

The characters in our model are positions in the genome whose states we model with a 4-tuple $(x, y, \bar{x}, \bar{y})$ where x and y are the number of maternal and paternal copies of the position, respectively, and $\bar{x}$ and $\bar{y}$ are the number of mutated maternal and paternal copies, respectively. We define $z := \max\{\bar{x}, \bar{y}\}$. As input we are given $H = [h_{p,c}]$ where $h_{p,c}$ is the VAF of character c in sample p. In addition, we are given $\mathcal{M} = [[\mu_{p,c,(x,y)}]]$ where $\mu_{p,c,(x,y)}$ is the *mixing proportion* of the copy-number state (x, y) of character c in sample p. The tensor $\mathcal{M}$ is obtained by running a copy-number caller [5, 9–12] on the B-allele frequencies and the read-depth ratios. The following linear system relates mixing proportions $\mathcal{M}$ and variant allele frequencies H to state frequencies $\mathcal{F} = [[f_{p,(c,(x,y,z))}]]$.

$$\sum_{z} f_{p,(c,(x,y,z))} = \mu_{p,c,(x,y)} \qquad \text{for all copy-number states } (x, y) \tag{6}$$

$$\sum_{(x,y,z)} z \cdot f_{p,(c,(x,y,z))} = h_{p,c} \cdot \sum_{(x,y)} (x + y) \cdot \mu_{p,c,(x,y)} \tag{7}$$

Importantly, this system of equations is under-determined, i.e. given $\mathcal{M}$ and H there are many possible $\mathcal{F} = [[f_{p,(c,(x,y,z))}]]$ that satisfy (6) and (7). We resolve the ambiguity in $\mathcal{F}$ by imposing biologically-motivated restrictions on the set of state trees $\mathcal{S}$. We define a *state graph* to be a directed graph whose nodes are character states, and whose edges are allowed transitions between states. Figure S2B shows a partial representation of this graph including the three types of edges: mutation edges, amplification edges and deletion edges that are defined as follows.

1. Mutation edges:
   - If $\bar{x} = \bar{y} = 0$ and $x > 0$ then $(x, y, \bar{x}, \bar{y}) \to (x, y, \bar{x} + 1, \bar{y}) \in E(G)$.
   - If $\bar{x} = \bar{y} = 0$ and $y > 0$ then $(x, y, \bar{x}, \bar{y}) \to (x, y, \bar{x}, \bar{y} + 1) \in E(G)$.
2. Amplification edges:
   - If $x > 0$ and $\bar{x} < x$ then $(x, y, \bar{x}, \bar{y}) \to (x + 1, y, \bar{x}, \bar{y}) \in E(G)$.
   - If $x > 0$ and $\bar{x} > 0$ then $(x, y, \bar{x}, \bar{y}) \to (x + 1, y, \bar{x} + 1, \bar{y}) \in E(G)$.
   - If $x > y$, $y > 0$ and $\bar{y} < y$ then $(x, y, \bar{x}, \bar{y}) \to (x, y + 1, \bar{x}, \bar{y}) \in E(G)$.
   - If $x > y$, $y > 0$ and $\bar{y} > 0$ then $(x, y, \bar{x}, \bar{y}) \to (x, y + 1, \bar{x}, \bar{y} + 1) \in E(G)$
   - If $x = y$, $y > 0$ and $\bar{y} < y$ then $(x, y, \bar{x}, \bar{y}) \to (y + 1, x, \bar{y}, \bar{x}) \in E(G)$.
   - If $x = y$, $y > 0$ and $\bar{y} > 0$ then $(x, y, \bar{x}, \bar{y}) \to (y + 1, x, \bar{y} + 1, \bar{x}) \in E(G)$.
3. Deletion edges:
   - If $x > \bar{x}$ and $x > y$ then $(x, y, \bar{x}, \bar{y}) \to (x - 1, y, \bar{x}, \bar{y}) \in E(G)$.
   - If $\bar{x} > 0$ and $x > y$ then $(x, y, \bar{x}, \bar{y}) \to (x - 1, y, \bar{x} - 1, \bar{y}) \in E(G)$.

- If $y > \bar{y}$ then $(x, y, \bar{x}, \bar{y}) \to (x, y-1, \bar{x}, \bar{y}) \in E(G)$.
- If $\bar{y} > 0$ then $(x, y, \bar{x}, \bar{y}) \to (x, y-1, \bar{x}, \bar{y}-1) \in E(G)$.
- If $x = y$ and $x > \bar{x}$ then $(x, y, \bar{x}, \bar{y}) \to (y, x-1, \bar{y}, \bar{x}) \in E(G)$.
- If $x = y$ and $\bar{x} > 0$ then $(x, y, \bar{x}, \bar{y}) \to (y, x-1, \bar{y}, \bar{x}-1) \in E(G)$.

We define a *valid state tree* as follows.

Definition 12. *A subtree S of G is a* valid state tree *provided that (1) S is rooted at $(1, 1, 0, 0)$, (2) there is at most one mutation edge in $E(S)$, and (3) vertices $V_{(x,y)} \subseteq V(S)$ with identical copy-number states (x, y) form a connected subtree of S.*

Intuitively, the first condition requires that the root state is the non-mutated heterozygous diploid state (assuming autosomes), the second condition is the infinite sites assumption on SNVs and the third condition is the infinite alleles assumption on copy-number states. Using these constraints, the system (6) and (7), restricted to the states in S, is fully determined. Note that the conditions above do allow for SNVs to be affected by multiple copy-number events as long as the same copy-number state does not recur. To show this, we remap each vertex $v_{(x,y,\bar{x},\bar{y})} \in V(G)$ to $v_{(x,y,z)}$ where $z = \max\{\bar{x}, \bar{y}\}$. We denote by $S_{(x,y)}$ the subset of vertices that have copy-number state (x, y). We now have the following lemma.

Lemma 8. *Let S be a valid state tree and let $X = \{(x, y) \mid v_{(x,y,z)} \in V(S)\}$. There exists at most one pair $(x^*, y^*) \in X$ such that $|V(S_{(x^*,y^*)})| = 2$ and $|V(S_{(x,y)})| = 1$ for all other $(x, y) \in X \setminus \{(x^*, y^*)\}$.*

Proof. Since S is a valid state tree, there is at most one mutation edge in $E(S)$. If there is no mutation edge in S then, by Definition 12, $S_{(x,y)} = \{v_{(x,y,0)}\}$ for all $(x, y) \in X$. Now consider the case where there is a mutation edge in $E(S)$. There must only be one such edge, which we denote by $(x^*, y^*, 0) \to (x^*, y^*, 1)$. By Definition 12, it holds that $S_{(x^*,y^*)} = \{v_{(x^*,y^*,0)}, v_{(x^*,y^*,1)}\}$ and that $|V(S_{(x,y)})| = 1$ for all $(x, y) \in X \setminus \{(x^*, y^*)\}$. □

Corollary 3. *The system composed of* (6) *and* (7) *is fully determined.*

Let S be a valid state tree and let $(x^*, y^*, 1)$ be a vertex of S such that $|V(S_{(x^*,y^*)})| = 2$. From (6) and (7) it follows that

$$f_{p,c,(x^*,y^*,1)} = h_{p,c} \cdot \sum_{(x,y,z) \in V(S)} (x + y) \cdot \mu_{p,c,(x,y)} - \sum_{\substack{(x,y,z) \in V(S) \\ (x,y) \neq (x^*,y^*)}} z \cdot \mu_{p,c,(x,y)} \tag{8}$$

$$= h_{p,c} \cdot \sum_{(x,y,z) \in V(S)} (x + y) \cdot \mu_{p,c,(x,y)} - \sum_{(x,y,z) \in V(S)} z \cdot \mu_{p,c,(x,y)} + \mu_{p,c,(x^*,y^*)}. \tag{9}$$

If $f_{p,c,(x^*,y^*,1)} \geq 0$ for all samples p then S_c is *compatible* with character c. Given $\mathcal{M}$ and H, the goal is to find all compatible valid state trees for each character c.

It may be the case that there exists two distinct compatible state trees S and S' such that for each distinct vertex $(x, y, z) \notin V(S) \cap V(S')$ it holds that $f_{p,c,(x,y,z)} = 0$. We do not wish to distinguish between such state

trees. Therefore, to deal with these degenerate cases, we remove all non-root vertices $(x, y, z) \neq (1, 1, 0)$ where $\mu_{p,c,(x,y)} = 0$ across all samples p through the following operation. We remove all incoming and outgoing edges of (x, y, z) and introduce edges $\pi(x, y, z) \to (x', y', z')$ for every child $(x', y', z') \in \delta(x, y, z)$. The order in which we consider the vertices does not matter.

Noisy VAFs. We now show how to derive frequency intervals $[\underline{f}_{p,(c,i)}, \bar{f}_{p,(c,i)}]$ from a valid state tree S_c, the VAF confidence interval $[\underline{h}_{p,c}, \bar{h}_{p,c}]$, and the mixing proportions $\mu_{p,c,(x,y)}$. We define

$$[\underline{h}'_{p,c}, \bar{h}'_{p,c}] = \left[\frac{\sum_{(x,y,z)} z \cdot \mu_{p,c,(x,y)} - \mu_{p,c,(x^*,y^*)}}{\sum_{(x,y)} (x+y) \cdot \mu_{p,c,(x,y)}}, \frac{\sum_{(x,y,z)} z \cdot \mu_{p,c,(x,y)}}{\sum_{(x,y)} (x+y) \cdot \mu_{p,c,(x,y)}} \right] \cap [\underline{h}_{p,c}, \bar{h}_{p,c}].$$

That is, $[\underline{h}'_{p,c}, \bar{h}'_{p,c}]$ is the VAF interval for which state tree S_c is compatible with character c in sample p. Character c is *compatible* with state tree S_c (in sample p) only if $[\underline{h}'_{p,c}, \bar{h}'_{p,c}]$ is nonempty. In that case, we have

$$[\underline{f}_{p,(c,(x^*,y^*,1))}, \bar{f}_{p,(c,(x^*,y^*,1))}] = \Bigg[\underline{h}'_{p,c} \sum_{(x,y)} (x+y) \cdot \mu_{p,c,(x,y)} - \sum_{(x,y,z)} z \cdot \mu_{p,c,(x,y)} + \mu_{p,c,(x^*,y^*)}, \\ \bar{h}'_{p,c} \sum_{(x,y)} (x+y) \cdot \mu_{p,c,(x,y)} - \sum_{(x,y,z)} z \cdot \mu_{p,c,(x,y)} + \mu_{p,c,(x^*,y^*)} \Bigg]$$

$$[\underline{f}_{p,(c,(x^*,y^*,0))}, \bar{f}_{p,(c,(x^*,y^*,0)} = [\mu_{p,c,(x^*,y^*)} - \bar{f}_{p,(c,(x^*,y^*,1))}, \mu_{p,c,(x^*,y^*)} - \underline{f}_{p,(c,(x^*,y^*,1))}]$$

and

$$[\underline{f}_{p,(c,(x,y,z))}, \bar{f}_{p,(c,(x,y,z)} = [\mu_{p,c,(x,y)}, \mu_{p,c,(x,y)}] \qquad \text{for all } (x, y, z) \text{ where } (x, y) \neq (x^*, y^*).$$

C Supplemental References

[1] Hans L. Bodlaender, Michael R. Fellows, and Tandy Warnow. Two strikes against perfect phylogeny. In Werner Kuich, editor, *Automata, Languages and Programming, 19th International Colloquium, ICALP92, Vienna, Austria, July 13-17, 1992, Proceedings*, volume 623 of *Lecture Notes in Computer Science*, pages 273–283. Springer, 1992.

[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] David Fernández-Baca. The perfect phylogeny problem. In *Steiner Trees in Industry*, pages 203–234. Springer, 2001.

[5] 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.

[6] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1979.

[7] Gunes Gundem, Peter Van Loo, Barbara Kremeyer, Ludmil B Alexandrov, Jose M C Tubio, Elli Papaemmanuil, Daniel S Brewer, Heini M L Kallio, Gunilla Högnäs, Matti Annala, Kati Kivinummi, Victoria Goody, Calli Latimer, Sarah O'Meara, Kevin J Dawson, William Isaacs, Michael R Emmert-Buck, Matti Nykter, Christopher Foster, Zsofia Kote-Jarai, Douglas Easton, Hayley C Whitaker, ICGC Prostate UK Group, David E Neal, Colin S Cooper, Rosalind A Eeles, Tapio Visakorpi, Peter J Campbell, Ultan McDermott, David C Wedge, and G Steven Bova. The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547):353–357, April 2015.

[8] D. Gusfield. *ReCombinatorics: The Algorithmics of Ancestral Recombination Graphs and Explicit Phylogenetic Networks*. MIT Press, 2014.

[9] Gavin Ha, Andrew Roth, Jaswinder Khattra, Julie Ho, Damian Yap, Leah M Prentice, Nataliya Melnyk, Andrew McPherson, Ali Bashashati, Emma Laks, et al. Titan: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Research*, 24(11):1881–1893, 2014.

[10] Serena Nik-Zainal et al. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, May 2012.

[11] Layla Oesper, Ahmad Mahmoody, and Benjamin J Raphael. THetA: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome Biology*, 14(7):R80, 2013.

[12] Layla Oesper, Gryte Satas, and Benjamin J Raphael. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics*, 30(24):3532–40, Dec 2014.