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

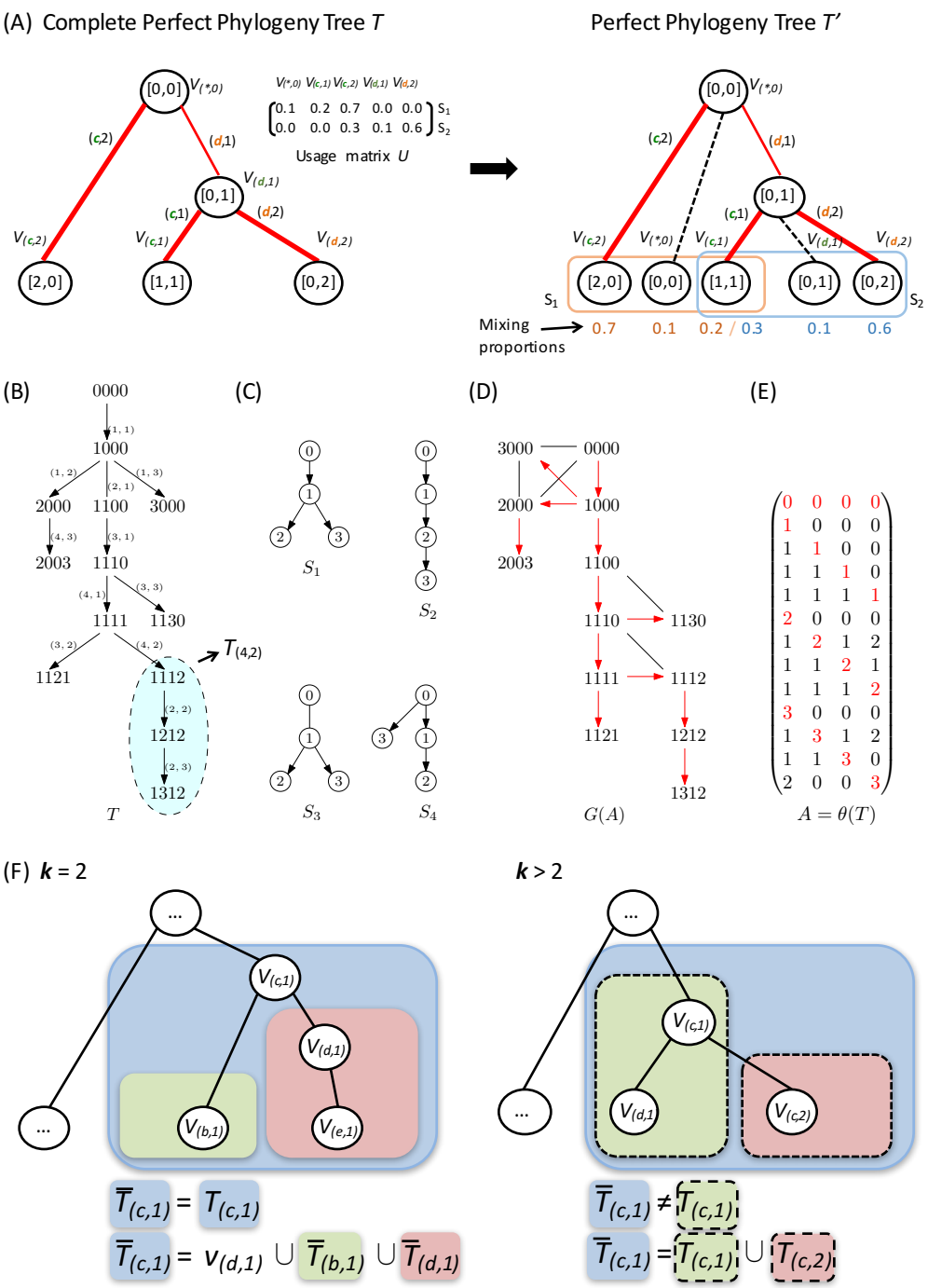


Figure S1: **Concepts of the Perfect Phylogeny Mixture Deconvolution Problem. Related to Figure 1.** (A) Each complete perfect phylogeny tree T corresponds to a perfect phylogeny tree T'. (B) A $4,4$-complete perfect phylogeny tree T. (C) State trees $\mathcal{S}$ determined by T. (D) Red edges denote 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, $D_{(c,1)} = \{1, 2\}$.

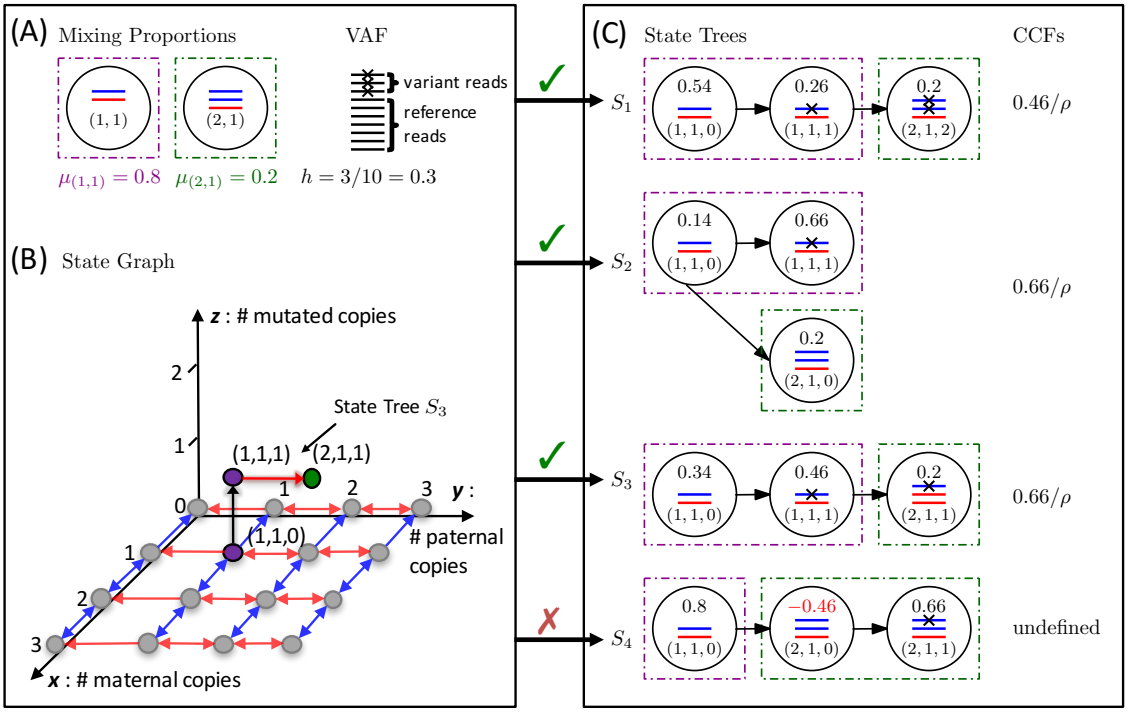
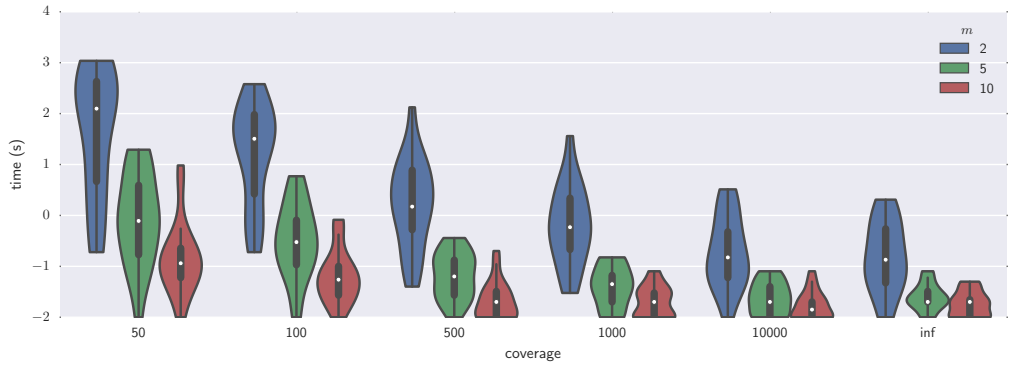
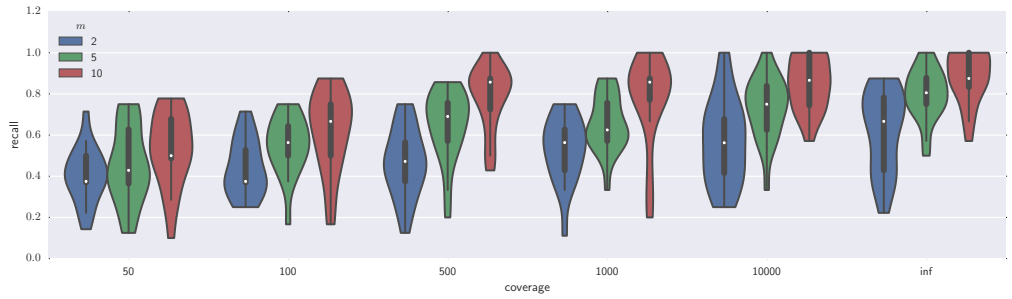


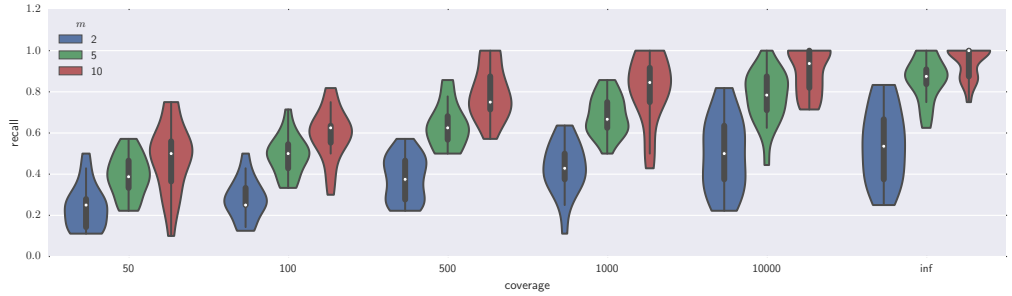
Figure S2: **The cancer cell fraction (CCF) of an SNV depends on its copy-number states, variant allele frequency and state tree. Related to Figure 2B.** (A) For each character in each sample, we observe copy-number states (x, y) with mixing proportions $\mu_{(x,y)}$ as well as a VAF h. (B) Our multi-state 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.

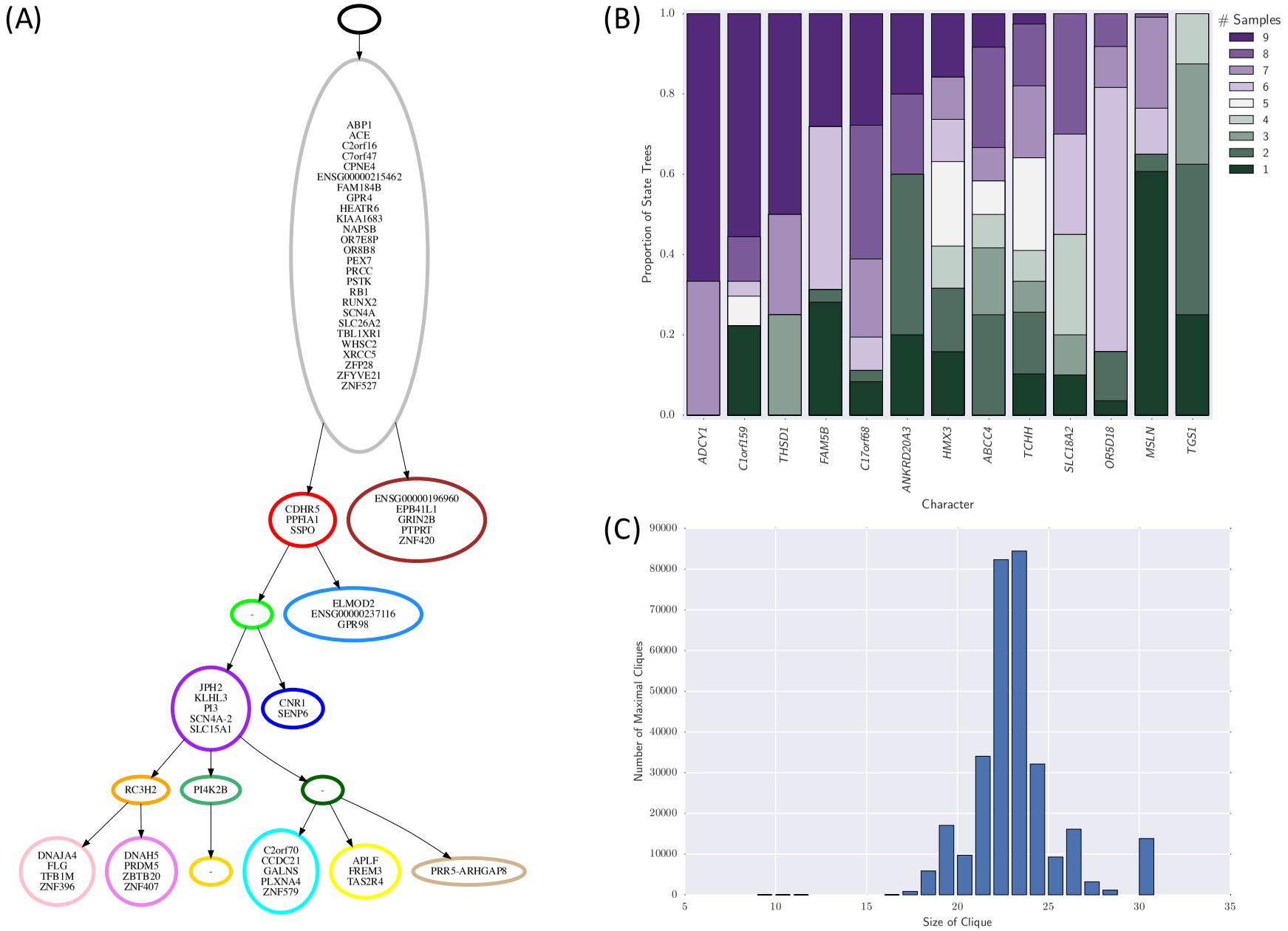


Figure S4: **Additional results for A22. Related to Figure 4.** (A) Reported tree for A22 by Gundem et al. [7]. Vertices and coloring correspond to the reported clusters on SNVs. Note that we only show the subset of SNVs that pass the filtering criteria described in Results. As such, some of the reported 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.

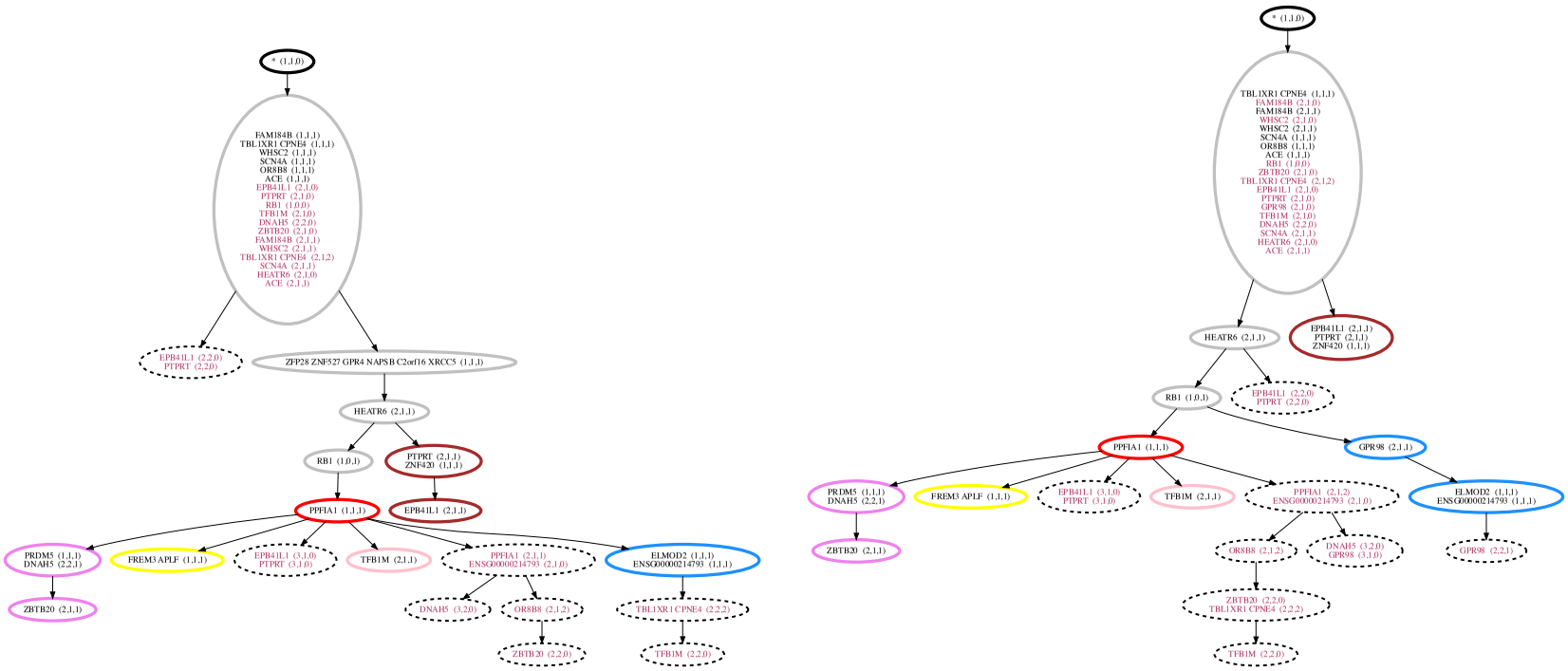
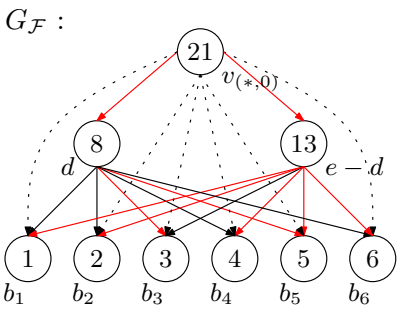


Figure S5: **Phylogenetic trees for A22. Related to Figure 4D.** (A) State tree S_1 for *PPFIA1* results in a tree where *PPFIA1* is a parent of the blue cluster. This relationship is consistent with the tree reported in [7]. (B) State tree S_2 for *PPFIA1* results in a tree in which the blue and the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