Supplementary Methods for CALDER: Inferring phylogenetic trees from longitudinal tumor samples

Contents

1 Proofs, related to STAR Methods

Lemma 1. *The following conditions are necessary and sufficient for a clone proportion matrix U and mutation tree T to determine a longitudinally-observed clone tree P:*

1. ***Permanent extinction**. For all clones v, $u_{t,v} = 0$ for all $t \geq t_v^{\max}$.*
2. ***Lineage continuity**. For each edge $(v, w) \in E_T$, $t_w^{\min} \leq t_v^{\max}$.*

Proof. First, we will show that the clone proportion matrix U and mutation tree $T = (V_T, E_T)$ that correspond to a longitudinally-observed clone tree $P_{U,B} = (V, E, C)$ necessarily meet these conditions.

Each colored vertex in P corresponds to the presence of a clone in a sample, so the nonzero entries of U are determined by the colored vertices in P: if $v_t \in V$, then $u_{t,v} > 0$. Let v_α where $\alpha = \min\{t; v_t \in P\}$ represent the first vertex corresponding to clone v. Let Δ_v^X represent the vertices in the subtree of tree X rooted at v. Observe that $t_v^{\min} = \min\{t; u_{t,v} + \sum_{w \in \Delta_v^T} u_{t,w} > 0\} = \min\{t; w_t \in \Delta_{v_\alpha}^P, w_t \in V\}$ and $t_v^{\max} = \min\{t; t > t_v^{\min}, v_t \notin V\}$. Additionally, the vertices and edges of T are determined by $P_{U,B}$: for each clone v represented as one or more vertices in P, there is a vertex $v \in V_T$; and for each edge $(v_a, w_b) \in E$ (for some colors a and b), there is an edge $(v, w) \in E_T$. By definition of $t_v^{\max}$, $u_{t,v} = 0$ for all $t > t_v^{\max}$, which is permanent extinction. To obtain lineage continuity, we first observe that each edge $(v, w) \in E_T$ is also an edge (v_a, w_b) in T. Then, by definition, $a < t_v^{\max}$ and $b \geq t_w^{\min}$, and because each path encounters colors in order, either $a = b = 0$ or $a = b - 1$, so $a \geq b - 1$. As a result, $t_v^{\max} > a \geq b - 1 \geq t_w^{\min} - 1$, so $t_v^{\max} > t_w^{\min} - 1$ and thus $t_w^{\min} \leq t_v^{\max}$.

Now, we will show that for a clone proportion matrix U and mutation tree T, where U meets permanent extinction and lineage continuity with respect to T, the corresponding observed clone tree $P_{U,B} = (V, E, C)$ is longitudinally-observed. Each vertex in this tree is colored by construction (see Section), so the first part of the definition is met. The second condition for $P_{U,B}$ to be longitudinally-observed

is that for any path from the root r to a leaf vertex, the colored vertices on this path are encountered in order of color.

Consider the path corresponding to a single clone v, $\pi(v) = v_0 \to v_{i_1} \to \ldots \to v_{i_k}$. From permanent extinction, we have that for all t such that $t_v^{\min} < t < t_v^{\max}$, the edge $(v_{t-1}, v_t) \in E$, and we also have that $(v_0, v_{t_v^{\min}}) \in E$. By construction, all edges between two vertices corresponding to the same clone are of this form, and thus all paths $\pi(v)$ (which consist entirely of such edges) adhere to the sequential coloring constraint.

Each remaining edge $(v', w') \in E$ represents an ancestral relationship between distinct clones, i.e., $(v, w) \in E_T$. By construction, each edge connects $v' = \text{argmax}_{x \in \pi(v)}\{c(x); c(x) < t_w^{\min}\}$ to w', where $c(w') = t_w^{\min}$ if $t_w^{\min} < t_w^{\max}$ (i.e., clone w is observed) or $c(w') = 0$ otherwise (i.e., if clone w is not observed). We proceed by cases to show that any path containing this edge adheres to the sequential coloring constraint:

- $c(v') = 0, c(w') = 0$. This case does not impact the coloring along any path.
- $c(v' > 0), c(w') > 0$. By permanent extinction, clone v is present in all samples t such that $t_v^{\min} \leq t < t_v^{\max}$. By lineage continuity, $t_w^{\min} \leq t_v^{\max}$, so clone v must be present in sample $t_w^{\min} - 1$. Thus, because $c(w') = t_w^{\min}$ and $c(v')$ is the latest sample strictly before $t_w^{\min}$ by construction, we have that $c(v') = c(w') - 1$ as required.
- $c(v') = 0, c(w') > 0$. In this case, let x be the most recent colored vertex ancestor of v', i.e., $x = \text{argmax}_{u \in r, \ldots, v'}\{c(u)\}$. If there is no such vertex, then $t_v^{\max} = 1$ by permanent extinction and therefore $c(w') = 1$ by lineage continuity. If there is such a vertex x, then $c(x) = c(w') - 1$ by the same logic as the previous case.
- $c(v') > 0, c(w') = 0$. Observe that this case is equivalent to the previous case, i.e., $v' = x$ for some edge (v'', w''), $c(v'') = 0, c(w'') > 0$: because there must be some observed descendant of w' in order for it to be included in the clone tree, there must be some edge (v'', w''), $c(v'') = 0, c(w'') > 0$ such that $c(u) = 0$ for all $u \in w' \to \ldots \to v''$, so v' is the most recent colored vertex ancestral to w'' and thus $c(v') = c(w'') - 1$ by the previous case.

Thus, because all edges on paths π between vertices corresponding to the same clone adhere to the sequential coloring constraint, and all edges corresponding to edges in T between distinct clones adhere to the sequential coloring constraint, the clone tree $P_{U,B}$ is longitudinally-observed. □

Lemma 2. *If $F = UB$ for some perfect phylogeny matrix B and clone proportion matrix U, and U is strictly positive (i.e., $u_{t,p} > 0$ for all time points t and clones p), then the observed clone tree $P_{U,B}$ is longitudinally observed.*

Proof. For all clones v, $t_v^{\min} = 1$ because clone v is present in the first sample, and $t_v^{\max} = \infty$ because its extinction is not observed. Thus, all clones satisfy permanent extinction because there are no samples t such that $t > t_v^{\max}$ for any clone v, and all edges $(v, w) \in T$ satisfy lineage continuity because $t_w^{\min} = 1 \leq t_v^{\max} = \infty$. Thus, because these conditions are sufficient for $P_{U,B}$ to be longitudinally observed, U and B correspond to a longitudinally-observed clone tree. □

Definition 1 (Rooted Subtree Consistency)**.** *A boolean function $\Phi : T \to \{true, false\}$ on rooted trees is* rooted subtree consistent *provided that if $\Phi(T)$ is true for a rooted tree, then $\Phi(T')$ is true for any subtree of T with the same root.*

Lemma 3. *Longitudinal constraints (sum condition, permanent extinction, and lineage continuity) are Rooted Subtree Consistent.*

Proof. We will proceed by showing that, if perfect phylogeny tree T meets longitudinal constraints with respect to a given frequency matrix F, then any $T' = (V', E')$ where $V' = V \backslash \{q\}$, $E' = E \backslash \{(p, q)\}$ for some leaf vertex q also meets longitudinal constraints with respect to the same F. This is sufficient for Rooted Subtree Consistency.

Let $t_k^{\min\prime}$ and $t_k^{\max\prime}$ represent the new values of $t_k^{\min}$ and $t_k^{\max}$, respectively, after the removal of edge (p, q). Because F does not change, $t_k^{\min\prime} = t_k^{\min}$ for all vertices. Additionally, because the only vertex whose children change is p, $t_k^{\max\prime} = t_k^{\max}$ for all vertices $k \in V', k \neq p$. Thus, permanent extinction holds for all vertices $k \in V', k \neq p$, and lineage continuity holds for all edges $(a, b) \in E', a \neq p$.

For the remaining conditions, we must consider how the mixture proportions of clone p change with the removal of edge (p, q). Let $u'_{t,p}$ correspond to the entries in U after the edge removal. From the sum condition, we have $u_{t,p} = f_{t,p} - \sum_{r \in \delta(p)} f_{t,r}$, and similarly, $u'_{t,p} = f_{t,p} - \sum_{r \in \delta(p)} f_{t,r} + f_{t,q} = u_{t,p} + f_{t,q}$. Also, because q is a leaf and has no children, $u_{t,q} = f_{t,q}$. Then, we consider two cases separately:

- $t_q^{\max} \leq t_p^{\max}$. In this case, $f_{t,q} = 0$ for all $t \geq t_p^{\max}$, so $u'_{t,p} = u_{t,p} = 0$ for all $t \geq t_p^{\max}$, and $t_p^{\max\prime} = t_p^{\max}$. Thus, lineage continuity and permanent extinction hold in T' as they held in the original T.
- $t_q^{\max} > t_p^{\max}$. In this case, $u'_{t,p} = f_{t,\ell(q)}$ for all $t \in [t_p^{\max}, t_q^{\max}]$, so $t_p^{\max\prime} = t_q^{\max}$. Thus, because T meets both conditions with respect to F, all edges $(v_p, v_s) \in E'$ meet lineage continuity because $t_q^{\max} > t_p^{\max} \geq t_s^{\min}$, and v_p meets permanent extinction by definition of $t_q^{\max}$.

□

Fixed-Precision Variant Allele Frequency Factorization Problem. *Given an $m \times n$ frequency matrix $F_d = [f_{t,i}]$ where each entry $f_{t,i} = k/d$ for some integer k and some integer d divisible by 20, determine whether or not there exists a clone proportion matrix U and perfect phylogeny matrix B such that $F_d = UB$.*

Lemma 4. *The fixed-precision VAFFP is NP-complete.*

The proof that the fixed-precision VAFFP is NP-complete is omitted, as it is identical to the proof by El-Kebir et al. (2015) that the original VAFFP is NP-complete – this proof only required the frequency values $\{0, 0.1, 0.15, 0.25, 0.5\}$, which can all be represented as fixed-precision frequency values with any such denominator d.

Lemma 5. *The longitudinal VAF factorization problem (LVAFFP) is NP-complete.*

Proof. Determining whether an instance F of the LVAFFP admits a solution is equivalent to determining whether there exists a factorization $F = UB$ such that the clone tree $P_{U,B}$ is longitudinally observed. We claim that the latter problem is NP-complete.

First, we show that the problem is in NP by describing how to check, in polynomial time, whether or not a solution is correct. A solution to the fixed-precision LVAFFP is determined by a mutation tree T. Given T, we can check in polynomial time whether it indeed factorizes the given frequency matrix F (by simply multiplying U and B), and whether or not it meets the three necessary and sufficient conditions listed in Section of the main text. A putative solution T to the LVAFFP is verifiable in time polynomial in n, the number of mutations, and m, the number of samples. T can be immediately invalidated if it is not a spanning tree with n vertices or if any mutation does not appear exactly once. Each sum condition involves at most $n \cdot m$ of the given frequency values, and is applied to each of the n vertices in T. The application of the sum condition to vertices in T yields the clone proportion matrix U. If U is a valid clone proportion matrix, $t_i^{\min}$ and $t_i^{\max}$ can be evaluated in polynomial time for each of n vertices by simply examining the relevant

column of the clone proportion matrix, i.e., at most m entries each for a total of up to $n \cdot m$. Then, the permanent extinction condition for each of n mutations i depends on $t_i^{\max}$ and at most m entries of the clone proportion matrix. Each lineage continuity condition depends on a single $t^{\min}$ value and a single $t^{\max}$ value, and because T is a spanning tree, there are exactly $n-1$ such conditions to check. Thus, the necessary and sufficient conditions can be checked in polynomial time, verifying whether or not T is a solution to the LVAFFP.

Next, we show that the decision problem version of the LVAFFP is NP-hard by reduction from the fixed-precision VAFFP. In order to relate these two problems, we transform the input fixed-precision frequency matrix F_d into frequency matrix $\hat{F}_d$ such that the multiplicity of factorizations $\hat{F}_d = UB$ remains the same, but all such factorizations become trivially longitudinal. We do this by adding a small quantity to each value in the frequency matrix to ensure that the sum condition (Equation 1) is always a strict inequality, and thus all possible entries in the clone proportion matrix U are strictly positive and all factorizations $\hat{F}_d = UB$ are trivially longitudinal as described in Lemma 2.

First, we construct the ancestry graph as previously described and remove strongly connected components. The ancestry graph is a directed graph $G = (V, E)$ with vertices $V = \{1, \ldots, n\}$ corresponding to mutations and edges $E = \{(i,j) | f_{t,i} \geq f_{t,j} \text{ for all } t \in [1,m]\}$ corresponding to potential ancestral relationships. Observe that a strongly connected component S in the ancestry graph must be a set of vertices such that, for all $p \in S, q \in S$, $f_{t,p} = f_{t,q}$ for all $t \in [1,m]$. Each of these strongly connected components can thus be represented as a single vertex v_S, corresponding to the cluster of mutations with identical frequencies that make up S. Then, each edge of the ancestry graph $(i,j), j \in S$ is replaced by an edge (i, v_S), and each edge $(i,j), i \in S$ similarly is replaced by an edge (v_S, j) (multiple edges between the same pair of vertices are collapsed to a single edge). By compressing all strongly connected components in this manner, we remove all cycles from the ancestry graph.

Then, we transform F_d into $\hat{F}$, such that $\hat{F}$ admits the same number of longitudinal factorizations as F_d admits factorizations. Let ϵ be a very small value such that $\epsilon < d$. Let δ_i represent the set of outgoing neighbors of vertex $i \in G$. We define entries of $\hat{F}$ as follows:

$$\hat{f}_{t,i} = f_{t,i} + \epsilon + \sum_{j \in \delta_i} \left(\hat{f}_{t,j} - f_{t,j} \right)$$

In other words, for all time points, each frequency value is increased by ϵ plus the frequency increments of its children. Note that the increment does not depend on t, and because G is directed and acyclic, this increment value can be computed recursively beginning with those vertices in G that have no outgoing edges.

Next, we show that if there is a solution to VAFFP(F_d), then there is a solution to LVAFFP($\hat{F}_d$). This follows from the fact that all factorizations $\hat{F}_d = UB$ are trivially longitudinal - particularly, every possible entry in U where the sum condition is met is at least ϵ:

$$\begin{aligned}
u &= \hat{f}_{t,i} - \sum_{j \in \delta_i} \hat{f}_{t,j} \\
&= f_{t,i} + \epsilon + \sum_{j \in \delta_i} \left(\hat{f}_{t,j} - f_{t,j} \right) - \sum_{j \in \delta_i} \left(f_{t,j} + \epsilon + \sum_{k \in \delta_j} \Gamma_k \right) \\
&= f_{t,i} + \epsilon + \sum_{j \in \delta_i} \left(\epsilon + \sum_{k \in \delta_j} \Gamma_k \right) - \sum_{j \in \delta_i} f_{t,j} - \sum_{j \in \delta_i} \left(\epsilon + \sum_{k \in \delta_j} \Gamma_k \right) \\
&= f_{t,i} + \epsilon - \sum_{j \in \delta_i} f_{t,j}
\end{aligned}$$

Thus, if $f_{t,i} \geq \sum_{j \in \delta_i} f_{t,j}$ then $u_{t,i} > 0$, so all factorizations UB are trivially longitudinal by Lemma 2.

Finally, we show that if there is a solution to LVAFFP($\hat{F}$), then there is a solution to VAFFP(F_d). This follows from the fact that the sum condition for each mutation i and its children δ_i has the same truth value in terms of $\hat{F}$.

$$\begin{aligned} \hat{f}_{t,i} &\geq \sum_{j \in \delta_i} \hat{f}_{t,j} \\ f_{t,i} + \epsilon + \sum_{j \in \delta_i} \left(\hat{f}_{t,j} - f_{t,j} \right) &\geq \sum_{j \in \delta_i} \hat{f}_{t,j} \\ f_{t,i} + \epsilon &\geq \sum_{j \in \delta_i} f_{t,j} \end{aligned}$$

Because $\epsilon < d$ and all frequencies are fixed-precision values, $f_{t,i} + \epsilon \geq \sum_{j \in \delta_i} f_{t,j}$ if and only if $f_{t,i} \geq \sum_{j \in \delta_i} f_{t,j}$. Thus, if there is a solution to LVAFFP($\hat{F}$), then there is a solution to VAFFP(F_d). □

2 ILP formulation for the LVAFFP, related to Figure 1 and STAR Methods

We formulate an integer linear program (ILP) to find the largest tree in an ancestry graph G that adheres to the sum, lineage continuity, and permanent extinction conditions. If this is a spanning tree, then we have a solution to the LVAFFP. First, we construct the ancestry graph $G_F = (V, E)$, which is the directed graph with vertices $V = \{1, \ldots, n\}$ and edges $E = \{(p, q) | f_{t,p} \geq f_{t,q} \text{ for all } t \in [1, m]\}$. Let r be an artificial root vertex with an outgoing edge to every other vertex in V. Let $E' = E \cup \{(r, w); w \in V\}$ denote this extended edge set. For $v \in V \cup \{r\}$, let $\delta^+(v) = \{w \in V; (v, w) \in E'\}$ be the set of outgoing neighbors of v, and let $\delta^-(v) = \{w \in V; (w, v) \in E'\}$ be the set of incoming neighbors of v. Let variables $\mathbf{x} \in \{0, 1\}^{|E'|}$ be binary variables indicating the presence or absence of edges in a solution. Given F, let $t_i^{\min} = \min\{t; f_{t,i} > 0\}$.

$$\max_{\mathbf{x}} \sum_{(i,j) \in E} x_{i,j}$$

$$\text{s.t.} \sum_{v_i \in \delta^+(r)} x_{r,i} = 1 \tag{1}$$

$$\sum_{i \in \delta^-(j)} x_{i,j} \geq x_{j,k} \qquad \text{for all } (j,k) \in E \tag{2}$$

$$\sum_{i \in \delta^-(j)} x_{i,j} \leq 1 \qquad \text{for all } j \in V \tag{3}$$

$$f_{t,i} - \sum_{v \in \delta^+(i)} f_{t,j} \cdot x_{i,j} \geq 0 \qquad \text{for all } i \in V, \quad t \in [1, m] \tag{4}$$

$$f_{t,i} - \sum_{j \in \delta^+(i)} f_{t,j} \cdot x_{i,j} \geq c \left(f_{t+1,i} - \sum_{j \in \delta^+(i)} f_{t+1,j} \cdot x_{i,j} \right) \qquad \text{for all } i \in V, \quad t \in [t_i^{\min}, m) \tag{5}$$

$$f_{t,i} - \sum_{j \in \delta^+(i)} f_{t,j} \cdot x_{i,j} \geq c \cdot x_{i,k} \qquad \text{for all } (i,k) \in E, \quad t \in [t_i^{\min}, t_k^{\min}) \tag{6}$$

$$x_{i,j} \in \{0, 1\} \qquad \text{for all } (i,j) \in E' \tag{7}$$

$$c > 0 \tag{8}$$

Constraints (1) and (2) enforce that T has exactly one root vertex. Constraint (3) ensures that the resulting graph is a tree. We introduce non-negative dummy variable c. Constraints (4-6) enforce the sum, permanent extinction and lineage continuity conditions respectively.

3 Enumeration algorithm for the LVAFFP, related to STAR Methods

We adapt the Gabow-Myers (Gabow and Myers, 1978) enumeration algorithm to solve the LVAFFP by enumerating constrained spanning trees of the ancestry graph. This adaptation relies on the property of *rooted subtree consistency*, defined in the STAR Methods section. Rooted subtree consistency guarantees that we can enumerate the full set of constrained spanning trees by adding one edge at a time. We show that the SC, PEC, and LCC are rooted subtree consistent in STAR Methods.

Algorithm 1: ENUMERATE(F, G, r)

Input: Frequency matrix F, ancestry graph $G = (V_0, E_0)$, root r

Output: All spanning trees of G rooted at r which determine a longitudinally-observed clone tree

$T \leftarrow (\{r\}, \emptyset)$

$H \leftarrow \emptyset$ `// initialize set of frontier edges`

foreach $(r, v) \in E_0$ **do**

`// Add outgoing edges from` r `to the frontier` H `if they meet constraints`

  $t_r^{\max} \leftarrow \min\{t; f_{t,r} = f_{t,v}\}$

  if $\forall t > t_r^{\max}, f_{t,r} = f_{t,v}$ **and** $t_r^{\max} \geq t_v^{\min}$ **then**

    PUSH$(H, (r, v))$

GROW(F, G, T, H)

Algorithm 2: GROW(F, G, T, H)

Input: Frequency matrix F, ancestry graph $G = (V_0, E_0)$, perfect phylogenetic tree $T = (V, E)$, frontier H

Output: All spanning trees of G which contain T and determine a longitudinally-observed clone tree

if $|V| = |V_0|$ **then**
  Output T
else
  $b \leftarrow$ **false**
  while $H \neq \emptyset$ **and** $\neg b$ **do**
    $(p, q) \leftarrow$ POP(H)
    $E \leftarrow E \cup \{(p, q)\}$
    $H' \leftarrow H$
    foreach $(q, v) \in E_0$ **do**
      `// Add all outgoing edges from` q `that could meet longitudinal conditions to the frontier`
      $t_q^{\max} \leftarrow \min\{t; f_{t,q} = f_{t,v}\}$
      if $v \notin V'$ **and** $\forall t > t_q^{\max}, f_{t,q} = f_{t,v}$ **and** $t_q^{\max} \geq t_v^{\min}$ **then**
        PUSH$(H', (q, v))$
    foreach $(v, w) \in H'$ **do**
      `// Remove edges from` H' `which would violate conditions once` (p, q) `is added to the tree`
      if $w = q$ **then**
        Remove (v, w) from H'
      else if $v = p$ **then**
        `// Check if adding` (p, w) `to` T' `would violate conditions`
        $t_p^{\max} \leftarrow \min\{t; f_{t,p} = f_{t,w} + \sum_{s \in \delta(p)} f_{t,s}\}$
        if $\exists t, f_{t,p} < f_{t,w} + \sum_{s \in \delta(p)} f_{t,s}$ `// check the sum condition`
        or $\exists t > t_p^{\max}, f_{t,p} > f_{t,w} + \sum_{s \in \delta(p)} f_{t,s}$ `// check permanent extinction`
        or $\exists (p, s) \in E \cup \{p, q\}, t_p^{\max} < t_s^{\min}$ `// check lineage continuity`
        then
          Remove (v, w) from H'
    $E \leftarrow E \backslash \{(p, q)\}$
    $E_0 \leftarrow E_0 \backslash \{(p, q)\}$
    $b \leftarrow$ BRIDGETEST$((p, q))$ `// returns true if the removal of this edge would cause` G `to become disconnected`

4 MILP formulation for the LVAFFP-U (CALDER), related to STAR Methods

Given interval estimates of mutation frequencies F^- and F^+, we find a maximal longitudinal tree, i.e., a maximal mutation tree that determines a longitudinally-observed clone tree. In addition, we would like the solution to account for as much as possible of the clone mixture proportions - as a proxy to this, we maximize the frequency lower bounds (f^--values) of the mutations included in the tree. Finally, once we have a maximal longitudinal tree the optimizes the previous objectives, we would like the clone proportion matrix with the most zero entries, subject to confidence interval bounds and longitudinal constraints. CALDER optimizes these objectives using the following ILP.

First, we construct the approximate ancestry graph $G_{F^-,F_+} = (V, E)$, which is the directed graph with vertices $V = \{1, \ldots, n\}$ and edges $E = \{(p, q) | f^+_{t,p} \geq f^-_{t,q} \text{ for all } t \in [1, m]\}$. Let r be an artificial root vertex with an outgoing edge to every other vertex in V. Let $E' = E \cup \{(r, w); w \in V\}$ denote this extended edge set. For $v \in V \cup \{r\}$, let $\delta^+(v) = \{w \in V; (v, w) \in E'\}$ be the set of outgoing neighbors of v, and let $\delta^-(v) = \{w \in V; (w, v) \in E'\}$ be the set of incoming neighbors of v. Let $\mathbf{x} \in \{0, 1\}^{|E'|}$ be binary variables indicating the presence or absence of edges in a solution. Let $\hat{F} \in [0, 1]^{m \times n}$ be the inferred frequency matrix. Let $U \in [0, 1]^{m \times n}$ be the inferred matrix of mixture proportions (determined by $\hat{F}$). Let binary variables $\boldsymbol{w} \in \{0, 1\}^n$ indicates the presence or absence of each vertex in the solution (i.e., $w_i = 1$ if and only if vertex i is in the tree). Let binary variables $Y \in \{0, 1\}^{n \times m}$ and $Z \in \{0, 1\}^{n \times m}$ be defined as follows: $y_{t,p} = \mathbb{1}\{t \geq t_p^{\min}\}$, and $z_{t,p} = \mathbb{1}\{t \geq t_p^{\max}\}$. Let $\boldsymbol{t^{\min}} \in \{0, 1, \ldots, m\}^n$ and $\boldsymbol{t^{\max}} \in \{0, 1, \ldots, m + 1\}^n$ be integer variables determined by $\hat{U}$. Finally, h is a hyperparameter corresponding to the minimum allowed clone proportion.

The objective function is constructed to combine the 3 objectives in the context of floating-point values with a fixed precision of 10^{-6}. The first term is the number of edges in the tree, which is equivalent to the number of vertices in the tree. The second term is the frequency lower bounds of included mutations: $f^-_{t,i}$ is provided as input, and $w_i = 1$ if and only if vertex i is included in the tree. The final term is the $L0$-norm of U, i.e., the number of zero entries. The indicator function and product are implemented using standard linearization techniques. Note that not all maximal trees are comparable in terms of $||U||_0$, as some mutations or clusters may enable more 0 values to be inferred than others. We include the second term in the objective to prioritize maximal trees whose vertices account for maximal frequency, as intuitively these trees more completely describe the tumor composition (due to the relationship between F and U).

$$\max_{\hat{F},U,\boldsymbol{x},Y,Z,\boldsymbol{w},\boldsymbol{t^{\min}},\boldsymbol{t^{\max}}} 10^7 \sum_{(i,j)\in E} x_{i,j} + \frac{10^6}{mn}\sum_{t=1}^{m}\sum_{i=1}^{n} w_i f^-_{t,i} + \frac{1}{mn}\sum_{t=1}^{m}\sum_{i=1}^{n} w_i \mathbb{1}\{u_{t,i}=0\}$$

$$\text{s.t.} \sum_{v_i\in\delta^+_r} x_{r,i} = 1 \tag{1}$$

$$\sum_{i\in\delta^-_j} x_{i,j} \geq x_{j,k} \qquad \text{for all } (j,k)\in E \tag{2}$$

$$\sum_{i\in\delta^-_j} x_{i,j} \leq 1 \qquad \text{for all } j\in V \tag{3}$$

$$u_{t,i} = \hat{f}_{t,i} - \sum_{j\in\delta^+(i)} x_{i,j}\hat{f}_{t,j} \qquad \text{for all } i\in V, t\in[1,m] \tag{4}$$

$$0 \geq \hat{f}_{t,i} - y_{t,i} \qquad \text{for all } t\in[1,m], i\in V \tag{5}$$

$$0 \geq u_{t,i} - (1-z_{t,i}) \qquad \text{for all } t\in[1,m], i\in V \tag{6}$$

$$u_{t,i} + (1-y_{t,i}) + z_{t,i} > h \qquad \text{for all } t\in[1,m], i\in V \tag{7}$$

$$x_{i,j} t^{\min}_j \leq t^{\max}_i \qquad \text{for all } (i,j)\in E \tag{8}$$

$$f^-_{t,i} \leq \hat{f}_{t,i} \leq f^+_{t,i} \qquad \text{for all } i\in V, t\in[1,m] \tag{9}$$

$$u_{t,i} \geq 0 \qquad \text{for all } t\in[1,m], i\in V \tag{10}$$

$$\sum_{i=1}^{n} w_i u_{t,i} \leq 1 \qquad \text{for all } t\in[1,m] \tag{11}$$

$$w_j = \sum_{i\in\delta^-_j} x_{i,j} \qquad \text{for all } j\in V \tag{12}$$

Constraint (1) enforces that T has exactly one root vertex.
Constraints (2) and (3) ensures that the resulting graph is a tree.
Constraint (4) is the sum condition relating U and $\hat{F}$.
Constraint (5) enforces the definition of $t^{\min}_i$.
Constraint (6) is the permanent extinction condition.
Constraint (7) enforces the minimum clone proportion threshold h.
Constraint (8) is the lineage continuity condition.
Constraint (9) requires that inferred frequencies respect the given interval bounds.

Constraints (10) and (11) ensure that U is a clone proportion matrix.
Constraint (12) defines $\boldsymbol{w}$.
The constraints defining dummy variables X and Y, as well as the linearization techniques used to represent products of variables, are omitted for the sake of brevity.