Theoretical estimates on the expected number of mutations for reconstructing clonal lineage trees.

Nishat Anjum Bristy[1] [0000−0002−8649−7390] and Russell Schwartz[1,2] [0000−0002−4970−2252]

[1] Ray and Stephanie Lane Computational Biology Department, Carnegie Mellon University, Pittsburgh PA 15213, USA

[2] Department of Biological Sciences, Carnegie Mellon University, Pittsburgh PA 15213, USA

Abstract. Phylogenetics, like many subdisciplines of computational biology, faces a growing challenge of dealing with increasingly large and complicated data sets that have been enabled by ever-improving technologies for sequencing. The issue is particularly acute for studies of somatic evolution, such as of single-cell populations in cancers, where vast single-cell data sets may now identify hundreds of thousands of genetically distinct cells with similar scales of mutations distinguishing them. At the same time, the complexity of the biology of somatic evolution has led to complex phylogeny methods that struggle to scale to even modest data sizes. In this paper, we explore the theoretical and empirical basis for one strategy for managing these large data sets: subsampling mutations to solve the computationally challenging phylogeny problem followed by faster solutions for placement of mutations on a putatively known guide tree. We specifically focus on the fundamental question of determining the number of mutations sufficient to recover the true phylogenetic tree at some level of resolution with high probability. We theoretically analyze variants of several common models that underlie popular tools for building clonal lineage trees. We further evaluate the robustness of these theoretical bounds through simulations of these models, extensions of them, and real biological datasets. The results suggest that modest numbers of mutations suffice to reconstruct clonal trees for typical numbers of clones, supporting the subsampling approach as a general strategy for managing the challenges of ever-growing data sets.

Keywords: Phylogenetics · Somatic mutation · Cancer evolution · Algorithms · Clonal lineage trees

1 Introduction

Reconstructing evolutionary trees, i.e., phylogenetics, is a fundamental problem in biology aimed at understanding how different entities such as species, genes, or cells diversify from a common ancestor. Applications of phylogenetic trees have branched well beyond their classic use in species/organismal evolution [22,14] to cover applications in single-cell somatic evolution [16] and cell lineage tracing [28]. Phylogenetics has proven a particular powerful tool for studies of somatic evolution in cancer progression [25,20] and the theory and tools developed for that purpose have since helped make possible much broader study of somatic evolution in other diseases [10] or even in putatively healthy tissues [1].

A growing problem in phylogenetic studies, at both species and somatic scales, is scaling to ever-larger data sets. At the species level, the field now needs to tackle databases of tens of thousands to even millions of organisms to be reconciled into single trees [29]. For somatic evolution, single-cell datasets can reach into tens of thousands of single cells per sample [26]. Cancer genomes are particularly problematic because they frequently exhibit phenotypes of hypermutability that can lead to thousands of mutations per tumor distributed across a heterogeneous population of cells [18]. While these vast data sets are crucial for shedding new light on somatic mutation processes they are also algorithmically challenging for phylogenetic inference methods. At the same time, somatic evolution applications have the advantage over typical uses in species trees of a somewhat imprecise definition of taxa (typically species for organismal evolution but genetically similar subpopulations of cells called clones in the somatic case). This leaves the option of solving a problem at a lower resolution (fewer defined clones) if the data do not support finer resolution.

Many methods operate either directly on the reads or on a binary mutation matrix (entities $\times$ mutations), where each entry indicates presence or absence of a mutation, which are interpreted in terms of a simplified model of the evolutionary process. Many methods use a "perfect phylogeny" model, which assumes that a mutation is gained once and never lost, equivalent in practice to the "infinite sites" model that there are an infinite number of potentially mutable sites and so each can be mutated only once along the evolutionary history [9]. Under the perfect phylogeny and infinite-sites assumption, the exact phylogeny

can be reconstructed in linear time [9]. More complex models based on finite sites assumptions allow a site to be mutated multiple times along the evolutionary history. Models such as Dollo [6] or k-Dollo [4] offer a compromise in which we assumes a mutation is gained only once and allowed to be lost at most k times along the branches of the phylogeny. Full finite site models may allow arbitrary recurrent mutation or reversion. Particularly as we move to more complicated variant models, such as allowing for complex structural variations (SVs) [3], copy number alterations (CNAs) [21], or combinations of these with more classic single nucleotide variant (SNV) models [7,24], the algorithmic difficulty of the problem may make it infeasible to scale to the sizes of data set now available.

The somatic evolution field has generally dealt with the problem of excessively large data sets heuristically. The most widely used approach is clustering mutations to a small number of equivalence classes that can then be treated as single weighted markers [2]. While valuable, this approach depends on our ability to reliably separate co-occuring mutations by variant allele frequency (VAF), an increasingly difficult proposition for noisy single cell sequencing as mutation numbers grow. Another possibility is the use of consensus tree methods [8], allowing one to apply methods that scale poorly in numbers of markers to subsets of the data and then synthesize the resulting family of trees into a consensus model. In this paper, we focus on a third strategy with precedent in species evolution [5] that we have deployed in our prior work [7]: using a subset of the mutations to solve the harder problem of finding the correct clonal tree topology and then using faster methods to place the remaining mutations onto that putatively known clonal tree. This can be thought of as a variation on a species tree strategy of using a subset of the data to form a reliable backbone tree followed by faster phylogenetic placement methods to place the remainder of the data on the backbone. The mutation sampling version of this problem hinges on the conjecture that if we have a mutation data set that is large compared to the number of clones we need to resolve, then we should be able to find the right tree with a small sample of the data. This paper seeks to put that conjecture on a firmer theoretical basis for somatic evolution models and explore the practical question of how many mutations we need to use in our phylogeny inference to get the right tree topology at a given clonal resolution.

In the remainder of the paper, we assess this conjecture through theoretical and empirical methods. We first determine the bounds for the number of mutations sufficient and necessary for reconstructing evolutionary histories for several variations of phylogeny inference from point mutations (SNVs). For each, we derive theoretical bounds to establish how many mutations one needs to derive the correct tree with confidence. We use experiments on simulated data for these models and more complex variants of them to show that these theoretical bounds yield good results in practice. Our results support the use of the subsampling strategy as a basis for fast and accurate somatic evolution inference in the face of large data sets and costly phylogeny models and provide a basis for further study into extending this strategy of mutation subsampling and placement to more complicated evolutionary and mutation models.

2 Preliminaries

We first define some general terminology we will use throughout this paper; other terminology will be introduced as needed. All considered phylogenetic trees $\mathcal{T}$ are binary trees with node set $\mathcal{V}$, edge set $\mathcal{E}$, and leaf set L. Let $m = |\mathcal{E}|$ denote the number of edges and $n = |L|$ the number of leaves. We represent mutation data by a binary matrix $M \in \{0,1\}^{n\times s}$, whose rows correspond to entities (e.g., species, observed cells, or clones) and columns correspond to mutations, where $M_{i,j} = 1$ indicates that mutation j is present in entity i. Throughout this section, we treat M as generated from an underlying phylogenetic tree T according to one of the evolutionary models considered (perfect, 1-Dollo, or coalescent). Let $\ell_1, \ldots, \ell_m > 0$ be the edge lengths of $\mathcal{T}$, and let $L_{\text{tot}} = \sum_{i=1}^{m} \ell_i$ denote the total tree length. Our goal is to determine the expected number of sampled mutations required for every edge of $\mathcal{T}$ to receive at least one mutation, ensuring that the tree can be reconstructed. Let p_i denote the probability that a mutation lands on edge i; given the branch lengths, these probabilities are proportional to the edge lengths,

$$p_i = \frac{\ell_i}{L_{\text{tot}}}, \qquad 0 < p_i \leq 1, \qquad \sum_{i=1}^{m} p_i = 1. \tag{1}$$

We formalize this as a coverage process over the edges of $\mathcal{T}$. Each mutation independently selects an edge i with probability p_i, and we define the "cover time" random variable T as –

$$T = \min\left\{ t \in \mathbb{N} : \text{every edge } i \in \mathcal{E} \text{ has been hit by at least one of the first } t \text{ sampled mutations} \right\}. \quad (2)$$

The quantity of interest throughout the paper is the expected cover time $\mathbb{E}[T]$, representing the expected number of mutations required for all edges of $\mathcal{T}$ to be observed at least once.

In the following subsections, we divide the analysis into four parts. We begin with the case of a perfect phylogeny with equal branch lengths and derive the exact number of mutations required for complete coverage, showing that it follows the classic $O(m \log m)$ behavior. We then establish general lower and upper bounds for the expected cover time when edge lengths are unequal and sampled from an arbitrary distribution. Finally, we extend the analysis to a loss-supported 1-Dollo phylogeny model and derive the corresponding expected number of sampled mutations.

Theorem 1 (Exact cover time for perfect phylogeny with equal branch lengths). *Let $\mathcal{T}$ be a perfect phylogeny with m edges. Let T be the number of mutations required until every edge has been hit at least once. Then*

$$\mathbb{E}[T] = mH_m, \qquad H_m = \sum_{k=1}^{m} \frac{1}{k} \quad (3)$$

In particular, $\mathbb{E}[T] = m \log m + \gamma + O(1)$ as $m \to \infty$, where γ is the Euler–Mascheroni constant.

Proof. Each sampled mutation falls on one of the m edges with probability $\frac{1}{m}$. With $p_i = 1/m$, the process of assigning mutations to edges is the classical coupon collector with m equiprobable coupons. Let T_i denote the number of additional mutations needed to discover a new edge given that $i-1$ distinct edges have already been seen ($i = 0, 1, \ldots, m$). At that time, the success probability for hitting a previously unseen edge is $q_i = \frac{m-(i-1)}{m}$ Hence T_i is geometric with mean $\mathbb{E}[T_i] = \frac{1}{q_i} = \frac{m}{m-(i-1)}$. Since $\mathbb{E}[T] = \sum_{i=0}^{m} \mathbb{E}[T_i]$ and the T_i's are independent, and thus linearity of expectation suffices.

$$\mathbb{E}[T] = \sum_{i=1}^{m} \mathbb{E}[T_i] = \sum_{i=1}^{m} \frac{1}{q_i} = \sum_{i=1}^{m} \frac{m}{m-(i-1)} = m \sum_{i=1}^{m} \frac{1}{i} = mH_m$$

where H_m is the m th harmonic number. This can be approximated as $\mathbb{E}[T] = m \log m + \gamma + \frac{1}{2m} = m \log m + \gamma + O(1)$, where $\gamma \approx 0.577215$ is the Eular-Mashcheroni's constant. For large m, $\mathbb{E}[T] \approx m \log m + \gamma$. □

We next generalize the problem for unequal edge lengths and thus different probabilities p_i of a sampled mutation getting mapped to an edge. In a classic coupon collectors problem, p_i's are considered fixed parameters. We generalize this to consider p_i's to be random variables following an arbitrary distribution. We will first derive a general upper bound applicable for any such distribution and then consider a special case of the Kingman coalescent with exponentially distributed branch lengths. For the general case, we show in Supplementary Section S1.1 that

$$\mathbb{E}[T] = \sum_{\emptyset \neq S \subseteq [m]} (-1)^{|S|+1} \frac{1}{\sum_{i \in S} p_i} = \int_0^\infty \left[1 - \prod_{i=1}^{m} \left(1 - e^{-p_i u} \right) \right] du. \quad (4)$$

In the equal-probability case $p_i = 1/m$, the integrand simplifies to $1 - (1 - e^{-u/m})^m$, and a direct computation gives $\mathbb{E}[T] = mH_m$ (Supplementary Lemma S1), where $H_m = \sum_{k=1}^{m} 1/k$ is the mth harmonic number. The following theorem generalizes this result for arbitrary edge probabilities.

Theorem 2 (Universal upper bound). *Let $p_1, \ldots, p_m > 0$ with $\sum_i p_i = 1$, and $p_{\min} = \min_i p_i = \frac{\ell_{\min}}{L_{tot}}$. Then*

$$\mathbb{E}[T] \leq \frac{H_m}{p_{\min}}, \qquad H_m = \sum_{k=1}^{m} \frac{1}{k}$$

Equality holds if and only if $p_i = 1/m$ for all edges i.

Proof. Starting from Equation (4), note that $x \mapsto 1 - e^{-xu}$ is increasing in x, so $1 - e^{-p_i u} \geq 1 - e^{-p_{\min} u}$ for all i. All factors lie in $(0, 1]$, hence

$$\prod_{i=1}^{m}(1 - e^{-p_i u}) \;\geq\; (1 - e^{-p_{\min} u})^m \quad \Rightarrow \quad 1 - \prod_{i=1}^{m}(1 - e^{-p_i u}) \;\leq\; 1 - (1 - e^{-p_{\min} u})^m.$$

Integrating both sides over $u \in [0, \infty)$ yeilds and substituting $v = p_{\min} u$ yields

$$\begin{aligned}
\mathbb{E}[T] &\leq \int_0^\infty \left[1 - (1 - e^{-p_{\min} u})^m\right] du \leq \frac{1}{p_{\min}} \int_0^\infty \left[1 - (1 - e^{-v})^m\right] dv \\
&\leq \frac{1}{p_{\min}} \int_1^0 \frac{1 - (1 - x)^m}{-x} dx \qquad \text{(Taking } x = e^{-v}, dv = \frac{dx}{-x}\text{)} \\
&\leq \frac{1}{p_{\min}} \int_0^1 \sum_{k=0}^{m-1} (1 - x)^k dx \leq \frac{1}{p_{\min}} \sum_{k=0}^{m-1} \int_0^1 (1 - x)^k dx \\
&\leq \frac{1}{p_{\min}} \sum_{k=0}^{m-1} \frac{-1}{k+1} [1 - x]_0^1 \leq \frac{1}{p_{\min}} \sum_{k=1}^{m} \frac{1}{k} \leq \frac{H_m}{p_{\min}}
\end{aligned}$$

Equality holds only if $p_i = p_{\min} = \frac{1}{m}$ for all i. □

Consequences. If the edge lengths ℓ_i are known, $p_i = \ell_i / L_{\text{tot}}$ and $\ell_{\min} = \min_i \ell_i$ give the deterministic bound

$$\mathbb{E}[T \mid \mathcal{T}] \;\leq\; H_m \, \frac{L_{\text{tot}}}{\ell_{\min}}.$$

Special case: Kingman coalescent trees. The Kingman coalescent [11] is a stochastic model of evolution in a neutrally evolving population of constant size. Going backward in time, lineages randomly merge pairwise until a single common ancestor is reached. When k lineages are present, the waiting time to the next coalescent event follows an exponential distribution with rate parameter $\binom{k}{2}$. We explore it as an example of an approximation to the edge length ratios we might expect in more realistic clonal trees. We assume $\mathcal{T}$ to be a random rooted Kingman coalescent binary tree on n leaves ($m = 2n - 2$ edges) whose edge lengths ℓ_e are independent exponential random variables determined by the aforementioned rates. In this setting, the probabilities $p_i = \ell_i / L_{\text{tot}}$ are random variables rather than fixed parameters. Taking the expectation over coalescent trees then gives $\mathbb{E}[T] \leq H_m \mathbb{E}_\ell[\frac{1}{p_{\min}}] = H_m \mathbb{E}_\ell[\frac{L_{\text{tot}}}{\ell_{\min}}]$. This is the expectation of the ratio of two dependent random variables, which is hard to simplify to a closed-form and often requires knowing the joint distribution (and covariance) of $\ell_{\min}$ and L_{tot}.

To obtain a tractable analytic estimate, we consider a deterministic approximation to the true coalescent in which each coalescent epoch has length fixed to its expected value. In this mean-epoch model, the time between the coalescent events for k lineages is $1/\binom{k}{2}$. The minimum edge length therefore equals $\ell_{\min} = \frac{1}{\binom{n}{2}}$ and the total expected tree length is $L_{\text{tot}} = 2H_{n-1}$, as we prove in Supplementary Lemma S2. Applying Theorem 2 with $p_{\min} = \ell_{\min} / L_{\text{tot}}$ yields

$$\mathbb{E}[T] \;\leq\; \frac{H_m}{p_{\min}} \;=\; H_m \frac{L_{\text{tot}}}{\ell_{\min}} \;=\; H_m \, (2H_{n-1}) \binom{n}{2} \;=\; n(n-1)\, H_m \, H_{n-1} \tag{5}$$

Since $H_n = \Theta(\log n)$, this leads to an asymptotic upper bound of $O\big(n^2 (\log n)^2\big)$. Thus, under this deterministic mean-epoch approximation, the expected number of mutations required to cover all branches of a coalescent tree scales quadratically with the number of lineages, up to logarithmic factors.

Next, we extend the analysis to evolutionary models that permit mutation losses. In cancer evolution, gained mutations are frequently lost due to independent events or through larger-scale copy number deletions [12]. To account for this, the K-Dollo model assumes that each mutation may be gained once but can be lost up to K times along the phylogeny. The perfect phylogeny corresponds to the special case $K = 0$. We derive the exact expected number of sampled mutations required for full edge coverage under the *1-Dollo model* ($K = 1$), leaving the general K-Dollo case as a direction for future work.

Theorem 3 (Expected cover time under the 1-Dollo model). *Under the 1-Dollo assumption, each mutation is gained once and may be lost at most once. A single mutation can therefore cover either one previously uncovered edge (gain-only) or two uncovered edges (gain–loss). Let $p = P_{\text{loss}} \in [0, 1)$ denote the probability that a gain undergoes loss. Then the expected number of mutations required to cover all m edges is*

$$\mathbb{E}[T] = \frac{m}{1+p}\left(H_m - \sum_{k=1}^{m} \frac{(-p)^{m-k+1}}{k}\right), \qquad H_m = \sum_{k=1}^{m} \frac{1}{k} \tag{6}$$

Proof. Let's consider the process of sequentially assigning mutations to the branches of $\mathcal{T}$. Let i denote the number of edges that are already covered by at least one mutation (equivalently, the number of mutations mapped so far). When a new mutation is sampled, three outcomes are possible: it hits (1) an uncovered gain-only edge, covering one new edge; (2) an uncovered gain–loss pair, covering two new edges; or an already covered edge, covering none. The corresponding state transitions are shown in Figure 1.

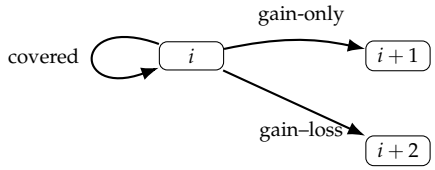

Fig. 1: State transitions under the 1-Dollo model. Each new mutation may cover one (gain-only) or two (gain–loss) new edges, or none (already covered).

Let $u_i = \frac{m-i}{m}$ be the probability that the sampled mutation lands on an uncovered edge. Then, $u_i p$ will be the probability that the mapped mutation will also have a loss associated with it. Let $P_1 = u_i(1-p), P_2 = u_i p, P_0 = 1 - u_i$, corresponding respectively to transitions $i \to i+1$, $i \to i+2$, and $i \to i$. Let $E_i = \mathbb{E}[T_i]$ denote the expected number of remaining mutations given i covered edges. This follows the recurrence

$$E_i = 1 + P_1 E_{i+1} + P_2 E_{i+2} + P_0 E_i \implies E_i = \frac{1 + P_1 E_{i+1} + P_2 E_{i+2}}{P_1 + P_2} = \frac{1}{u_i} + (1-p)E_{i+1} + pE_{i+2}$$

where, $E_m = 0$ is the base case. Let $r = m - i$ be the number of uncovered edges and define $A_r = E_{m-r}$, where A_r is a reparameterization of E_{m-r} denoting the expected number of mutations when r edges are remaining to be mapped. Then, $u_i = r/m$ and $A_r = \frac{m}{r} + (1-p)A_{r-1} + pA_{r-2}$ where the base cases are $A_0 = 0$, $A_1 = m$. Defining $S_r = A_r - A_{r-1}$ gives $S_r - pS_{r-1} = \frac{m}{r}$ with base case, $S_1 = m$. Solving the recurrence, we get:

$$S_r = m(-p)^{r-1} + m\sum_{k=2}^{r} \frac{(-p)^{r-k}}{k}$$

$$\implies A_r = m\sum_{j=1}^{r}(-p)^{j-1} + m\sum_{j=1}^{r}\sum_{k=2}^{j} \frac{(-p)^{j-k}}{k} \qquad \text{(Since } A_r = \sum_{j=1}^{r} S_j\text{)}$$

The first term is geometric, so $\sum_{j=1}^{r}(-p)^{j-1} = \frac{1-(-p)^r}{1+p}$. For the double sum,

$$\sum_{j=1}^{r}\sum_{k=2}^{j} \frac{(-p)^{j-k}}{k} = \sum_{k=2}^{r}\frac{1}{k}\sum_{j=k}^{r}(-p)^{j-k} = \sum_{k=2}^{r}\frac{1}{k}\cdot\frac{1-(-p)^{r-k+1}}{1+p}.$$

Hence

$$\begin{aligned} A_r &= \frac{m}{1+p}\left[\left(1-(-p)^r\right)+\sum_{k=2}^{r}\frac{1}{k}-\sum_{k=2}^{r}\frac{(-p)^{r-k+1}}{k}\right] \\ &= \frac{m}{1+p}\left[\left(1+\sum_{k=2}^{r}\frac{1}{k}\right)-\left((-p)^r-\sum_{k=2}^{r}\frac{(-p)^{r-k+1}}{k}\right)\right] \\ &= \frac{m}{1+p}\left(H_r-\sum_{k=1}^{r}\frac{(-p)^{r-k+1}}{k}\right) \qquad \text{(Since } H_r=\sum_{k=1}^{r}\frac{1}{k}\text{)} \end{aligned}$$

Starting from no covered edges ($i = 0, r = m$),

$$\mathbb{E}[T] = A_m = \frac{m}{1+p}\left(H_m-\sum_{k=1}^{m}\frac{(-p)^{m-k+1}}{k}\right)$$

□

Corollary 1 (No-loss case, i.e. perfect phylogeny). *For $p = 0$ (mutations gained once and never lost), the inner sum vanishes and*

$$\mathbb{E}[T] = mH_m,$$

recovering the classical coupon-collector expectation.

Code Availability

All scripts used for simulation are implemented in Python 3 and available at https://github.com/CMUSchwartzLab/mutation-subsampling.git

3 Experimental Results

We have conducted experiments on simulated and real variation data designed to validate the correctness of the bounds shown in this paper, assess how pessimistic they are in practice, and test how well they extend to more complex but realistic evolutionary scenarios beyond those covered in the proofs. We first validated the robustness of the bounds using simulated data and assessed tree reconstruction accuracies with respect to SPhyR [4] and PAUP [27] for perfect and 1-Dollo phylogenies, and with PAUP for coalescent-derived phylogenies. We further used simulated data to assess the k-Dollo case and the full coalescent model empirically. Additionally, we evaluated the effectiveness of these bounds on real cancer patient samples. For these cases, we limit the comparisons to SNVs as in our theoretical bounds, although the true data contains multiple mutation types, including SNVs, CNAs, and SVs.

Evaluation criteria. To assess the quality of the estimated tree topologies on simulated datasets, we compared them with ground-truth trees using *normalized Robinson-Foulds (RF) distance* [19]. To evaluate branch-length accuracy, especially under mutation subsampling, we further computed the *mean absolute error (MAE)* between the true versus estimated branch lengths after normalizing the branch lengths for each tree by that tree's total branch length, computed over all pairs of leaves in the n-leaf trees. Additionally, to evaluate the accuracy of mutation placement on the inferred trees, we compute the *pairwise ancestral relationship accuracy* as described in Supplementary Section S2.1.

3.1 Evaluation on simulated data

To evaluate how much phylogeny quality is affected by mutation counts across the range from the lower and upper bounds established theoretically, we simulated single-cell WGS data with SNVs based on known ground-truth tumor and coalescent phylogenies. We explain the simulation process for perfect, K-Dollo and coalescent phylogenies in detail in the following sections.

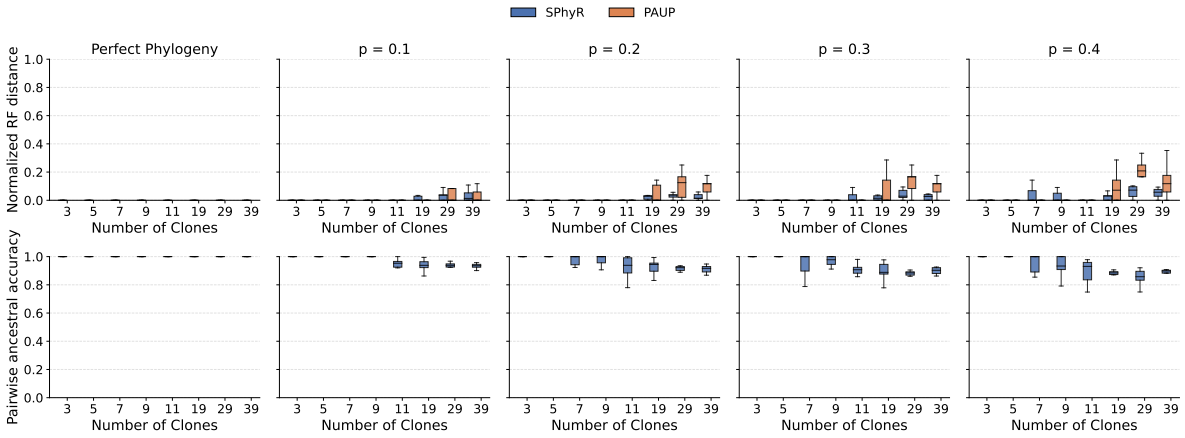

Fig. 2: Results for perfect and 1-Dollo phylogeny reconstruction with SPhyR, where mutations are sampled using Equations 3 and 6 respectively for different number of clones $N \in \{3,5,7,9,11,19,29,39\}$. For the 1-Dollo phylogenies, we consider different loss probabilities, $p \in \{0.1, 0.2, 0.3, 0.4\}$.

Perfect and 1-dollo phylogeny simulations. To evaluate the robustness of our bounds for perfect and 1-Dollo simulations, we first generated random full binary perfect phylogenies with $n \in \{2,3,4,5,6,10,15,20\}$ leaves. This corresponds to $(2n-1) \in \{3,5,7,9,11,19,29,39\}$ total nodes or clones in the tree. Since a full binary tree with n leaves has $m = 2n-2$ edges, we generated mH_m mutations and mapped each of these mH_m mutations to a uniformly chosen edge for each full binary perfect phylogeny, $\mathcal{T}^*$. After assigning all mutations, we counted how many branches of the perfect phylogeny remained without any mutation. We simulated 10 replicates for each value of n. Figure S1 shows the distribution of the number of unmapped branches. The results indicate that using the perfect-phylogeny bound almost always results in every branch receiving at least one mutation. We ran SPhyR [4] on the generated perfect phylogenies with mH_m mutations. Figure 2 shows the results of SPhyR for perfect phylogenies with the aforementioned number of clones. In each case, we reconstruct the ground truth tree with SPhyR, resulting in zero normalized RF distance and pairwise ancestral accuracies = 1.

To generate 1-Dollo phylogenies, we extended each perfect phylogeny $\mathcal{T}^*$ with mH_m mutations by introducing single-loss events. For each mutation, we independently introduced a loss with probability p, removing the mutation and all of its descendants. This produces 1-Dollo instances that contain both gains and single independent losses while preserving the structure of $\mathcal{T}^*$. We then sample a subset of mutations according to Equation (6), run SPhyR [4] and PAUP [27] on the sampled data, and compare the reconstructed trees against the full 1-Dollo phylogeny. Figure 2 shows the 1-Dollo phylogeny reconstruction accuracies. We run 10 replicates for each loss probability $p \in \{0.1, 0.2, 0.3, 0.4\}$ and vary the number of clones as $\{3,5,7,9,11,19,29,39\}$. For each simulated instance, we run SPhyR [4] and compare the reconstructed trees using normalized Robinson–Foulds (RF) distances (top panel) and pairwise ancestral relationship accuracies (bottom panel). As the number of clones increases, the normalized RF distance increases slightly, although the reconstructed trees remain highly accurate. We see similar results when evaluating mutation placements via pairwise ancestral relationship accuracy. With three and five clones, there is little difference between the sampled and full 1-Dollo trees, supporting the accuracy of the reconstructed trees for both cases. However, as the number of clones increases, the gap between the mutations in the full and sampled 1-Dollo trees widens, increasing the complexity for SPhyR and leading to larger reconstruction errors. We see similar results for different values of the mutation loss probability, $p \in \{0.1, 0.2, 0.3, 0.4\}$. According to Equation (6), higher loss probabilities result in fewer sampled mutations, which explains the gradual increase in RF distances and the corresponding decrease in pairwise ancestral accuracies as p increases. Nonetheless, both topology and phylogenetic placement remain highly accurate on the ranges of clones considered, well beyond the 10 clones typically resolved in tumor phylogeny studies.

Empirical evaluation on k-Dollo phylogenies. We generalize beyond the theoretical bounds of the 1-Dollo phylogenies to a more realistic scenario for somatic evolutionary trees, simulating K-Dollo phylogenies follow-

ing SPhyR [4] and the similar method as described in the previous section, by introducing losses in the perfect phylogeny, $\mathcal{T}^*$ for each mutation with loss probability p, allowing at most k losses per mutation. Specifically, we iteratively select a mutation and introduce a loss with probability p if the mutation currently has at most $(k-1)$ losses. Once a loss is introduced for mutation i, all descendants of the corresponding node are set to carry that loss.

Since we do not derive an analytical bound for k-Dollo phylogenies, we instead provide empirical results for $k \in \{1,2,3,4\}$. For each k, we vary the number of sampled mutations as a multiple x of mH_m ($x \times mH_m$), where $x \in \{0.05, 0.1, 0.5, 0.7, 1, 2, 3\}$ and run SPhyR on the simualted instances with its K-Dollo setting. Figure 3 summarizes the empirical estimates. Across all k values, we observe that the normalized RF distances decrease and the pairwise ancestral relationship accuracies increase beginning around $x = 0.7$ - 1.0, indicating that the trees become reliably reconstructable once the number of sampled mutations reaches this range for a loss probability of $p = 0.1$. When we calculated the theoretical ratio for x, we derived the values $\{1.0, 0.89, 0.93, 0.91, 0.9, 0.92, 0.92\}$ respectively for clones $\in \{3, 5, 7, 9, 11, 19, 29\}$, which matches the empirical results from the $k = 1$ case. We additionally simulated k-Dollo phylogenies with higher loss probabilities $p \in \{0.2, 0.3, 0.4\}$. In the supplementary figures (Figures S2, S3, and S4), we observe similar trends to the $p = 0.1$ case. However, as both the loss probability and k increase, SPHYR requires larger number of sampled mutations to infer accurate phylogenies.

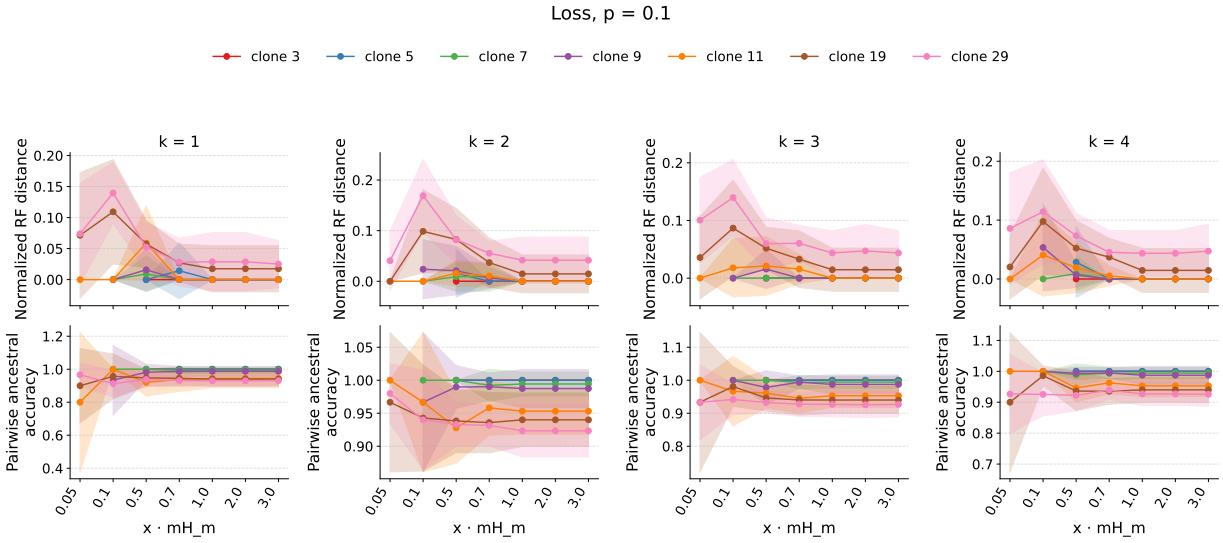

Fig. 3: Results for k-Dollo phylogeny reconstruction with SPhyR [4]. The figure shows normalized phylogeny inference (top) and pairwise ancestral accuracy (bottom) across a range of k, clone numbers, and sampling rates for loss probability, $p = 0.1$.

Coalescent tree simulations. To simulate coalescent trees with n leaves representing n observed clones, we use CellCoal [17], which starts by simulating a genealogy for the sampled cells under the neutral coalescent model with an exponential population growth model [23] for a defined parameterization of the coalescent for cancer cell samples [15]. A healthy somatic cell is considered to define the "outgroup branch" for the cell phylogeny. We simulate 10 replicates for each of $n \in \{3, 5, 7, 9, 11\}$ clone trees, running CellCoal with 0/1 alphabet under the infinite sites model (ISM).

We considered two settings of CellCoal. (1) In the first setting, we explicitly defined the coalescent parameters, setting the effective population size to 10^5, the exponential growth rate to 10^{-5} and 10,000 variant sites, and generated trees whose branch lengths were exponential random variables following the Kingman Coalescent model. (2) In the second setting, we generated trees with branch lengths fixed to the expected mean-epoch lengths under the coalescent model, such that each inter-coalescent interval corresponded ex-

actly to its theoretical mean. For both settings, the simulated healthy outgroup cell is generated from a reference genome and includes representative germline SNP positions, which we exclude from downstream analyses to ensure only somatic variants are considered.

After the simulation, we run PAUP [27] to reconstruct phylogenies from the coalescent-simulated data. To test how phylogeny reconstruction performs with different numbers of sampled mutations, we first run PAUP with all the non-SNP positions from the coalescent simulations. We then repeat the analysis with $n(n-1)H_mH_{n-1}$ subsampled mutations, following the mean-epoch coalescent bound in Equation (5).

Figures 4(a)–(b) show the normalized Robinson–Foulds (RF) distance and the mean absolute error (MAE) of normalized branch lengths for the random coalescent-branch simulations, while Figures 4(c)–(d) show the corresponding results for trees with deterministic mean-epoch branch lengths. In both cases, black lines represent tree reconstructions using the full set of mutations, whereas red lines correspond to subsampling exactly $n(n-1)H_mH_{n-1}$ mutations. We see that across both simulation models, tree topologies were reconstructed with high accuracies even under mutation subsampling, indicating that the predicted number of mutations is sufficient to recover the correct tree structures. However, the branch-length estimation was more challenging for trees with random coalescent branch lengths, PAUP exhibited higher variability and error, whereas trees with deterministic mean-epoch lengths showed substantially more consistent and accurate branch-length estimation.

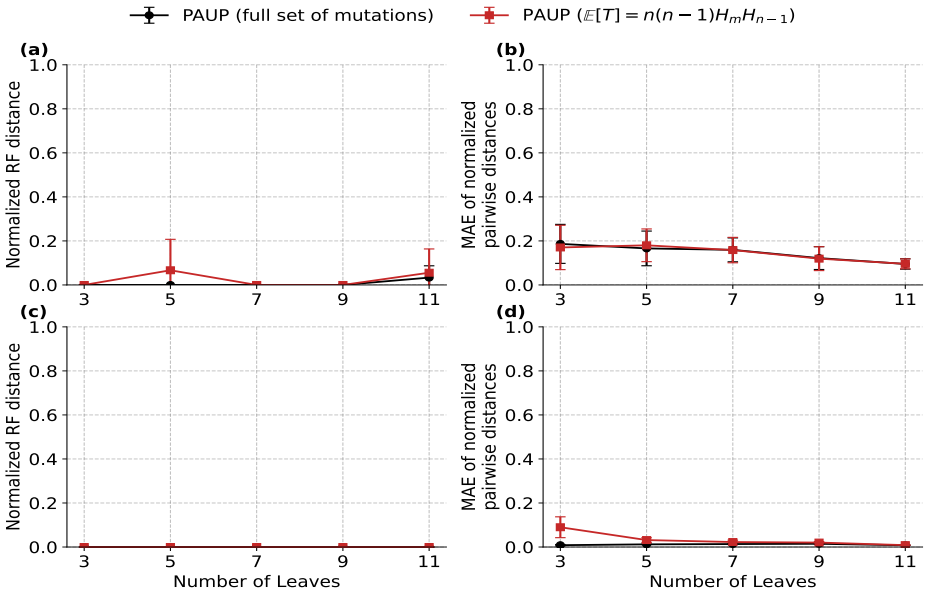

Fig. 4: Results of coalescent phylogeny reconstruction using the full set (black lines) and subsampled set (red lines) of mutations across varying numbers of leaves, $n \in 3, 5, 7, 9, 11$. (a) and (b) show the normalized Robinson–Foulds (RF) distance and the mean absolute error (MAE) in branch-length estimation for random coalescent simulations following the Kingman model. (c) and (d) show the corresponding RF distance and MAE for coalescent simulations with branch lengths fixed to the expected mean-epoch lengths under the coalescent model.

3.2 Evaluation on a metastatic colorectal cancer patient

To evaluate the robustness of our theoretical bounds on a real cancer dataset, we analyzed the metastatic colorectal cancer dataset (CRC1) from Leung et al. [13], which includes 178 single cells and 16 single-nucleotide variants (SNVs) identified through targeted sequencing of 1000 cancer-associated genes. Since the theoretical bounds exceeded the number of SNVs available for reconstructing a perfect phylogeny with more than

four clusters, we tested 10 replicates of three and four taxa trees, comparing reconstructions from the sampled set of mutations versus the full set of mutations. We additionally restricted the number of mutations to the maximum available in the dataset and reconstructed five and six clusters trees using PAUP [27]. In each case, PAUP produced identical trees from the sampled versus full mutation sets, resulting in a normalized Robinson–Foulds (RF) distance of 0. To further assess similarity, we compared the L1 distance of the pairwise branch lengths normalized by total tree length for the full and sampled clonal trees. As shown in Figure 5, the number of sampled mutations increases with the number of taxa, while the L1 distance between the branch-length ratios of the full and sampled trees decreases, indicating increasing structural similarity. The results confirm that following the theoretical bounds reliably yields equivalent topologies to the full data set, although with some error in inferred edge lengths.

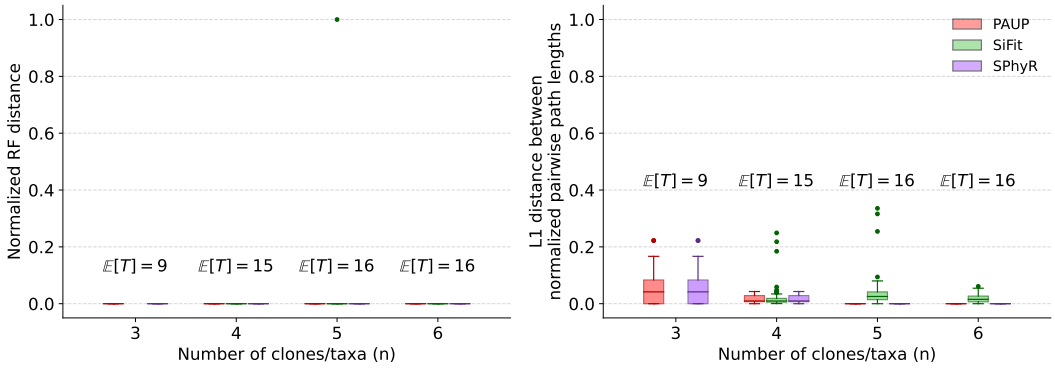

Fig. 5: Results on the CRC1 patient dataset [13] for $n \in 3, 4, 5, 6$ taxa (clusters). We ran PAUP [27] using the neighbor-joining method for 10 replicates of the sampled mutations with $\mathbb{E}[T] \in 9, 15, 16, 16$ for $n \in 3, 4, 5, 6$, respectively.

4 Discussion

This paper examined the problem of how many mutations one needs to reliably reconstruct clonal lineage trees from SNV data under various evolutionary models. The question is motivated by a practical problem of how to solve for complex and computationally intensive evolutionary models needed in somatic clone tree applications in the face of increasingly large variation data sets. We established theoretical bounds for some simplified evolutionary models commonly used in clonal lineage studies, showing that modest numbers of variants are sufficient to derive the correct tree topologies for realistic numbers of clones. Empirical studies with simulated and real tumor data confirm that these bounds provide good guidance in practice despite the simplified evolutionary models from which they derive.

This work provides a firmer basis for a practical strategy of dealing with large variant sets and algorithms that scale poorly in their size by subsampling to smaller numbers of variants to derive the correct tree followed by using faster algorithms to place the remaining variants on the tree. Considerable room remains, though, to extend the work further. The theory so far covers only some relatively simplified evolutionary models typical of those assumed in somatic phylogeny methods, but might be extended to more involved models that better capture the true biology. Results on simulated and real data suggest that the bounds we establish hold at least approximately for more realistic data. The present work has only looked at SNV data, and both the theoretical and empirical studies ought to be extended to cover the harder cases of evolution by CNA and SV events important in somatic evolution processes. There is likely also room for improvement in algorithms for both inference of topologies from small variant sets and on fast placement of unsampled variants on that topology, as well as other extensions to more effectively use unsampled variants to validate topologies or assess uncertainty within them. Furthermore, there may be other uses of the derived bounds to explore. For example, while we focused on the scenario of how to subsample mutations when they are available in excess, there is a parallel problem of how to define the maximum level of resolution of clonal structure our data can support when mutations are sparse.

References

1. De, S.: Somatic mosaicism in healthy human tissues. Trends in Genetics **27**(6), 217–223 (2011)
2. Deshwar, A.G., Vembu, S., Yung, C.K., Jang, G.H., Stein, L., Morris, Q.: Phylowgs: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome biology **16**(1), 35 (2015)
3. Eaton, J., Wang, J., Schwartz, R.: Deconvolution and phylogeny inference of structural variations in tumor genomic samples. Bioinformatics **34**(13), i357–i365 (2018)
4. El-Kebir, M.: Sphyr: tumor phylogeny estimation from single-cell sequencing data under loss and error. Bioinformatics **34**(17), i671–i679 (2018)
5. Erdos, P.L., Steel, M.A., Székely, L.A., Warnow, T.J.: A few logs suffice to build (almost) all trees (i). Random Structures and Algorithms **14**(2), 153–184 (1999)
6. Farris, J.S.: Phylogenetic analysis under dollo's law. Systematic Biology **26**(1), 77–88 (1977)
7. Fu, X., Lei, H., Tao, Y., Schwartz, R.: Reconstructing tumor clonal lineage trees incorporating single-nucleotide variants, copy number alterations and structural variations. Bioinformatics **38**(Supplement_1), i125–i133 (2022)
8. Govek, K., Sikes, C., Oesper, L.: A consensus approach to infer tumor evolutionary histories. In: Proceedings of the 2018 Acm international conference on bioinformatics, computational biology, and health informatics. pp. 63–72 (2018)
9. Gusfield, D.: Efficient algorithms for inferring evolutionary trees. Networks **21**(1), 19–28 (1991)
10. Kennedy, S.R., Loeb, L.A., Herr, A.J.: Somatic mutations in aging, cancer and neurodegeneration. Mechanisms of ageing and development **133**(4), 118–126 (2012)
11. Kingman, J.F.C.: The coalescent. Stochastic processes and their applications **13**(3), 235–248 (1982)
12. Kuipers, J., Jahn, K., Raphael, B.J., Beerenwinkel, N.: Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. Genome research **27**(11), 1885–1894 (2017)
13. Leung, M.L., Davis, A., Gao, R., Casasent, A., Wang, Y., Sei, E., Vilar, E., Maru, D., Kopetz, S., Navin, N.E.: Single-cell dna sequencing reveals a late-dissemination model in metastatic colorectal cancer. Genome research **27**(8), 1287–1299 (2017)
14. Nei, M., Kumar, S.: Molecular evolution and phylogenetics. Oxford university press (2000)
15. Ohtsuki, H., Innan, H.: Forward and backward evolutionary processes and allele frequency spectrum in a cancer cell population. Theoretical Population Biology **117**, 43–50 (2017)
16. Pennington, G., Smith, C.A., Shackney, S., Schwartz, R.: Reconstructing tumor phylogenies from heterogeneous single-cell data. Journal of Bioinformatics and Computational Biology **5**(02a), 407–427 (2007)
17. Posada, D.: Cellcoal: coalescent simulation of single-cell sequencing samples. Molecular biology and evolution **37**(5), 1535–1542 (2020)
18. Roberts, S.A., Gordenin, D.A.: Hypermutation in human cancer genomes: footprints and mechanisms. Nature Reviews Cancer **14**(12), 786–800 (2014)
19. Robinson, D.F., Foulds, L.R.: Comparison of phylogenetic trees. Mathematical biosciences **53**(1-2), 131–147 (1981)
20. Schwartz, R., Schäffer, A.A.: The evolution of tumour phylogenetics: principles and practice. Nature Reviews Genetics **18**(4), 213–229 (2017)
21. Schwarz, R.F., Trinh, A., Sipos, B., Brenton, J.D., Goldman, N., Markowetz, F.: Phylogenetic quantification of intratumour heterogeneity. PLoS computational biology **10**(4), e1003535 (2014)
22. Semple, C., Steel, M., et al.: Phylogenetics, vol. 24. Oxford University Press on Demand (2003)
23. Slatkin, M., Hudson, R.R.: Pairwise comparisons of mitochondrial dna sequences in stable and exponentially growing populations. Genetics **129**(2), 555–562 (1991)
24. Sollier, E., Kuipers, J., Takahashi, K., Beerenwinkel, N., Jahn, K.: Compass: joint copy number and mutation phylogeny reconstruction from amplicon single-cell sequencing data. Nature communications **14**(1), 4921 (2023)
25. Somarelli, J.A., Ware, K.E., Kostadinov, R., Robinson, J.M., Amri, H., Abu-Asab, M., Fourie, N., Diogo, R., Swofford, D., Townsend, J.P.: Phylooncology: understanding cancer through phylogenetic analysis. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer **1867**(2), 101–108 (2017)
26. Svensson, V., Vento-Tormo, R., Teichmann, S.A.: Exponential scaling of single-cell rna-seq in the past decade. Nature protocols **13**(4), 599–604 (2018)
27. Wilgenbusch, J.C., Swofford, D.: Inferring evolutionary trees with paup. Current protocols in bioinformatics (1), 6–4 (2003)
28. Woodworth, M.B., Girskis, K.M., Walsh, C.A.: Building a lineage from single cells: genetic techniques for cell lineage tracking. Nature Reviews Genetics **18**(4), 230–244 (2017)
29. Zaharias, P., Warnow, T.: Recent progress on methods for estimating and updating large phylogenies. Philosophical Transactions of the Royal Society B **377**(1861), 20210244 (2022)