PhyDOSE: Design of Follow-up Single-cell Sequencing Experiments of Tumors

Leah Weber[1◕], Nuraini Aguse[1◕], Nicholas Chia[2,3], Mohammed El-Kebir[1*]

1 Dept. of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America
2 Microbiome Program, Center for Individualized Medicine, Mayo Clinic, Rochester, Minnesota, United States of America
3 Division of Surgical Research, Department of Surgery, Mayo Clinic, Rochester, Minnesota, United States of America

◕ These authors contributed equally to this work.
* melkebir@illinois.edu

Abstract

The combination of bulk and single-cell DNA sequencing data of the same tumor enables the inference of high-fidelity phylogenies that form the input to many important downstream analyses in cancer genomics. While many studies simultaneously perform bulk and single-cell sequencing, some studies have analyzed initial bulk data to identify which mutations to target in a follow-up single-cell sequencing experiment, thereby decreasing cost. Bulk data provide an additional untapped source of valuable information, composed of candidate phylogenies and associated clonal prevalence. Here, we introduce PhyDOSE, a method that uses this information to strategically optimize the design of follow-up single cell experiments. Underpinning our method is the observation that only a small number of clones uniquely distinguish one candidate tree from all other trees. We incorporate distinguishing features into a probabilistic model that infers the number of cells to sequence so as to confidently reconstruct the phylogeny of the tumor. We validate PhyDOSE using simulations and a retrospective analysis of a leukemia patient, concluding that PhyDOSE's computed number of cells resolves tree ambiguity even in the presence of typical single-cell sequencing errors. We also conduct a retrospective analysis on an acute myeloid leukemia cohort, demonstrating the potential to achieve similar results with a significant reduction in the number of cells sequenced. In a prospective analysis, we demonstrate that only a small number of cells suffice to disambiguate the solution space of trees in a recent lung cancer cohort. In summary, PhyDOSE proposes cost-efficient single-cell sequencing experiments that yield high-fidelity phylogenies, which will improve downstream analyses aimed at deepening our understanding of cancer biology.

Author summary

Cancer development in a patient can be explained using a phylogeny — a tree that describes the evolutionary history of a tumor and has therapeutic implications. A tumor phylogeny is constructed from sequencing data, commonly obtained using either bulk or single-cell DNA sequencing technology. The accuracy of tumor phylogeny inference increases when both types of data are used, but single-cell sequencing may

become prohibitively costly with increasing number of cells. Here, we propose a method that uses bulk sequencing data to guide the design of a follow-up single-cell sequencing experiment. Our results suggest that PhyDOSE provides a significant decrease in the number of cells to sequence compared to the number of cells sequenced in existing studies. The ability to make informed decisions based on prior data can help reduce the cost of follow-up single cell sequencing experiments of tumors, improving accuracy of tumor phylogeny inference and ultimately getting us closer to understanding and treating cancer.

Introduction

Tumorigenesis follows an evolutionary process during which cells gain and accumulate somatic mutations that lead to cancer [1]. The most natural expression of an evolutionary process is a *phylogeny* — a tree that describes the order and branching points of events in the history of a cellular population. Tumor phylogenies are critical to understanding and ultimately treating cancer, with recent studies using tumor phylogenies to identify mutations that drive cancer progression [2, 3], assess the interplay between the immune system and the clonal architecture of a tumor [4, 5], and identify common evolutionary patterns in tumorigenesis and metastasis [6, 7]. These downstream analyses critically rely on accurate phylogenies that are inferred from sequencing data of a tumor.

The majority of current cancer genomics data consist of pairs of matched normal and tumor samples that have undergone bulk DNA sequencing. Bulk data is composed of sequences from cells with distinct genomes. More specifically, we observe frequencies $\mathbf{f} = [f_i]$ for the set of somatic mutations in the tumor (Fig 1a). Many deconvolution methods have been proposed for tumor phylogeny inference from such data [8–13], typically inferring a set $\mathcal{T}$ of equally plausible trees (Fig 1b). These approaches are unsatisfactory, as candidate trees with different topologies may alter conclusions in downstream analyses. Single-cell sequencing (SCS), as opposed to bulk sequencing, enables us to observe specific clones present within the tumor. These clones correspond to the leaves of the true phylogeny, allowing phylogeny inference methods to reconstruct the tree itself once we observe all clones in the tumor [14–17]. However, the elevated error rates of SCS, as well as its high cost [18], make it prohibitive as a standalone method for phylogeny inference. As such, hybrid methods have been recently proposed to infer high-fidelity phylogenies from combined bulk and SCS data obtained from the same tumor [19, 20].

Several hybrid datasets have been obtained by performing bulk and single-cell DNA sequencing simultaneously [21, 22]. However, there is merit in first performing bulk sequencing to guide follow-up SCS experiments. For instance, several studies first identified a subset of single-nucleotide variants from the bulk data to target in subsequent SCS experiments, thereby reducing costs compared to conventional whole-genome SCS approaches [23–25]. A recently introduced method, SCOPIT, computes how many cells are needed to observe all clones of a tumor, given estimates on the smallest prevalence of a clone as well as the number of clones to detect [26]. The authors provide no guidance on how to obtain these two quantities. Here, we build upon this work by directly incorporating knowledge encoded by the trees $\mathcal{T}$ inferred from the initial bulk sequencing data. Indeed, by using data from a SCS experiment we may eliminate trees from $\mathcal{T}$ that do not align with the observed clones (Fig 1). In other words, if we observe all clones in a tumor, it is possible to determine the phylogeny of the tumor. However, is it possible to achieve the same goal by observing fewer clones? If so, how many cells are necessary for us to observe the required clones?

We introduce Phylogenetic Design Of Single-cell sequencing Experiments

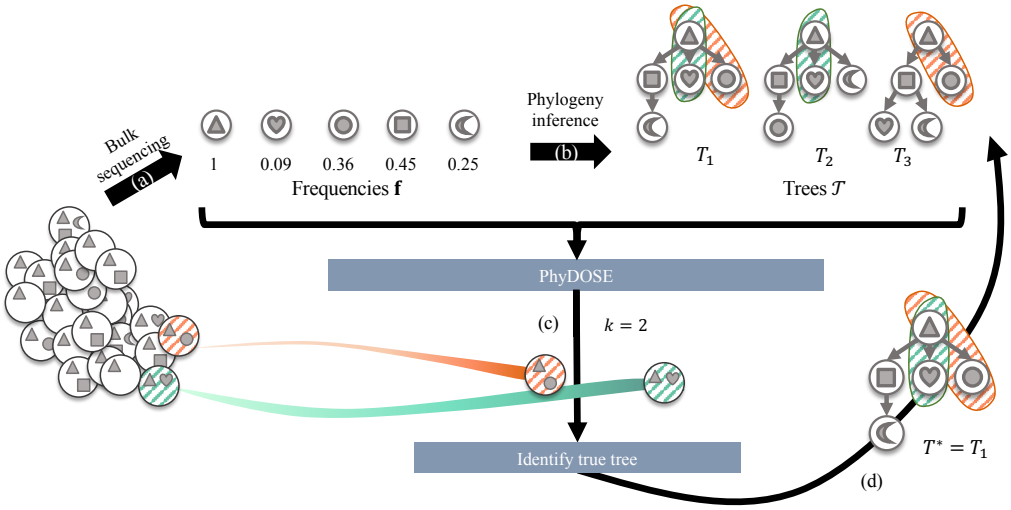


Fig 1. PhyDOSE computes the number of single cells to sequence to identify the true phylogeny. (a) Mutation frequencies **f** obtained from bulk DNA sequencing data. (b) The solution space $\mathcal{T}$ of trees inferred from **f**. We show a distinguishing feature of T_1 (orange and green). (c) For tree T_1, PhyDOSE suggests that $k = 2$ single cells suffice to observe clones that are unique to T_1. (d) In a follow-up SCS experiment we observe $k = 2$ cells, one from the orange clone and one from the green clone. As such, we eliminate trees T_2 and T_3, concluding that phylogeny T_1 is the true phylogeny T^*.

(PhyDOSE), a method to strategically design a follow-up SCS experiment aimed at inferring the true phylogeny (Fig 1). Given a set $\mathcal{T}$ of candidate trees inferred from initial bulk data, we describe how to distinguish a single tree T among the rest using features unique to T. In particular, if our SCS experiment results in observing cells corresponding to a distinguishing feature of T, we may conclude that T is in fact the true tree. This means that we can typically identify T using only a subset of the clones. To determine the number of cells to sequence, we introduce a probabilistic model that incorporates SCS errors and models successful SCS experiments as a tail probability of a multinomial distribution (Fig 1c). Finally, we reconcile the sampled cells utilizing these distinguishing features to infer the true phylogeny (Fig 1d). We validate PhyDOSE using both simulated data and a retrospective analysis of a leukemia patient that has undergone both bulk and SCS sequencing. We also demonstrate the utility of PhyDOSE by prospectively computing how many cells are needed to resolve the uncertainty in phylogenies of a recent lung cancer cohort. The cost-efficient SCS experiments enabled by PhyDOSE will yield high-fidelity phylogenies, improving downstream analyses aimed at understanding tumorigenesis and developing treatment plans.

Materials and methods

We introduce Phylogenetic Design Of Single-cell sequencing Experiments (PhyDOSE), a method to determine the number of single cells to sequence to identify the true phylogeny given initial bulk sequencing data. PhyDOSE is implemented in C++/R and is available at `https://github.com/elkebir-group/PhyDOSE`. This section describes the various methodological components of PhyDOSE.

Problem Statement

Let n be the number of single-nucleotide variants, or simply *mutations*, identified from initial bulk sequencing data of a matched normal and tumor biopsy sample. For each mutation i, we observe the *variant allele frequency* (VAF), i.e. the fraction of aligned reads that harbor the tumor allele at the locus of mutation i. Specialized methods exist that combine copy number information and VAFs to infer a *cancer cell fraction* f_i for each mutation i, which is the proportion of cells in the tumor biopsy that contain at least one copy of the mutation [3, 27–29]. Here, we refer to cancer cell fractions as *frequencies*. Typically, phylogenies $\mathcal{T}$ inferred by current methods from frequencies $\mathbf{f} = [f_i]$ adhere to the infinite sites assumption. That is, each mutation i is introduced exactly once at vertex v_i and never subsequently lost.

When we sequence a single cell from the same tumor biopsy, assuming no errors, we identify a clone of the tumor. In other words, we observe a set of mutations that must form a connected path in the unknown true phylogeny T^*. By repeatedly sequencing single cells until we observe all clones in the tumor, we will have observed all root-to-vertex paths of T^*, thus identifying tree T^* itself. We assume that (i) the true unknown phylogeny T^* is among the trees in $\mathcal{T}$ and that (ii) mutations among single cells that we sample from the tumor biopsy follow the same distribution as $\mathbf{f}$. These assumptions are important for the mathematical derivation of PhyDOSE but it is typical for violations to occur in practice. Through simulations, we explore the impact of violating these assumptions and show that our approach is robust to many realistic scenarios.

This leads to the following question and problem statement with respect to these two assumptions. How many single cells do we need to identify T^* with confidence level γ?

Problem 1 (SCS Power Calculation (SCS-PC))**.** Given a set $\mathcal{T}$ of candidate phylogenies, frequencies $\mathbf{f}$ and confidence level γ, find the minimum number k^* of single cells needed to determine the true phylogeny T^* among $\mathcal{T}$ with probability at least γ.

Clearly, we do not know which phylogeny in $\mathcal{T}$ is the true underlying phylogeny T^* of the tumor. Thus, we consider a slightly different problem: In the T-SCS-PC problem (defined formally at the end of the section), we are given an arbitrary phylogeny $T \in \mathcal{T}$ and want to perform a similar power calculation when conditioning on T being the true phylogeny. By solving the T-SCS-PC problem for all trees $T_1, \ldots, T_{|\mathcal{T}|}$, we obtain the numbers $k(T_1), \ldots, k(T_{|\mathcal{T}|})$ of single cells needed for each tree. As T^* is in $\mathcal{T}$, the maximum number among $k(T_1), \ldots, k(T_{|\mathcal{T}|})$ is an upper bound on the number of required SCS experiments to identify T^* with probability at least γ. To solve the T-SCS-PC problem, we need to reason for which SCS experiments we can conclude that T is the true phylogeny.

Observe that each tree T in $\mathcal{T}$ describes a unique set of clones, corresponding to the sets of mutations encountered in all root-to-vertex paths of T (Fig 1). Thus, if we observe all clones of a phylogeny T in our SCS experiments, we may conclude that T is the true phylogeny. What is the probability of doing so? To answer this question, we must compute the prevalence of each clone in the tumor biopsy.

For phylogenies that adhere to the infinite sites assumption, the *prevalence* $\mathbf{u}(T, \mathbf{f}) = [u_i]$ of the clones in the tumor biopsy are uniquely determined by the phylogeny T and frequencies $\mathbf{f}$ as

$$u_i = f_i - \sum_{j \in \delta_T(i)} f_j \qquad \forall i \in [n]. \tag{1}$$

where $\delta_T(i)$ is the set of children of the node where mutation i was introduced [9].

Tumor phylogeny inference methods guarantee that the inferred phylogenies $\mathcal{T}$ from frequencies $\mathbf{f}$ have clonal prevalence $\mathbf{u}(T, \mathbf{f}) = [u_i]$ that are nonnegative and that

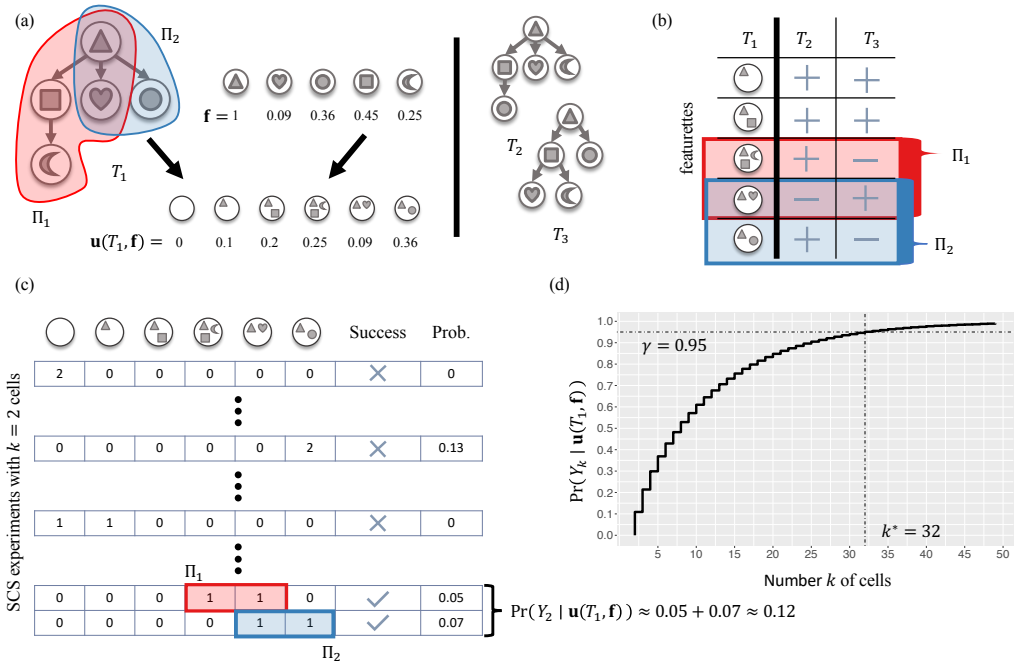


Fig 2. The SCS Power Calculation for Phylogeny T (T-SCS-PC) problem. (a) We are given frequencies $\mathbf{f}$ and a tree T_1 that we want to distinguish from the other trees $\{T_2, T_3\}$. The pair $(T_1, \mathbf{f})$ uniquely determine clonal prevalence $\mathbf{u}(T_1, \mathbf{f})$. (b) Featurettes of T_1 correspond to root-to-vertex paths, yielding distinguishing features Π_1 and Π_2, each with one featurette absent in T_2 and another absent in T_3. (c) With $k = 2$ cells, we must observe clones from either Π_1 or Π_2 for a successful outcome, resulting in probability $\Pr(Y_2 \mid \mathbf{u}(T_1, \mathbf{f})) \approx 0.12$. (d) To increase this probability to $\gamma = 0.95$, we need $k^* = 32$ cells.

$\sum_{i=1}^{n} u_i \leq 1$, where the remainder $u_0 = 1 - \sum_{i=1}^{n} u_i$ is the prevalence of the normal clone. Thus, conditioning on a phylogeny T and frequencies $\mathbf{f}$, sequencing one cell from the tumor will lead us to observe one of the $n+1$ clones of T with probabilities $(u_0, \ldots, u_n)$. In other words, the outcome of this SCS experiment with one cell is a draw from the categorical distribution $\mathrm{Cat}(u_0, \ldots, u_n)$. The possible outcomes of a SCS experiment composed of k cells thus follow a multinomial distribution $\mathrm{Mult}(u_0, \ldots, u_n)$. Thus, the probability of observing all tumor clones of T in such a SCS experiment with k cells corresponds to the tail probability of the multinomial where each of the n tumor clones is observed at least once.

The corresponding *power calculation* is to determine the smallest number for k where the tail probability is greater or equal to the confidence level γ. Note that this power calculation for observing all clones has been previously introduced [26].

Importantly, in many cases we need not observe all clones of T to distinguish T from the remaining phylogenies $\mathcal{T} \setminus \{T\}$ (Fig 2). This means that we may conclude that T is the true phylogeny with a SCS experiment with fewer cells. To formalize this notion, we start by defining a featurette.

Definition 1. A *featurette* τ is a subset of mutations.

We say that a featurette τ is *present* in a phylogeny T if the nodes/mutations of τ form a connected path of T starting at the root node, otherwise we say that τ is *absent* in T. The same featurette, however, may be present in more that one phylogeny. Thus, multiple featurettes may be required to distinguish a phylogeny T from the remaining

phylogenies $\mathcal{T} \setminus \{T\}$.

Definition 2. A set Π of featurettes is a *distinguishing feature of* T if (i) for all featurettes $\tau \in \Pi$ it holds that τ is present in T, and (ii) for each remaining phylogeny $T' \in \mathcal{T} \setminus \{T\}$ there exists a featurette $\tau' \in \Pi$ where τ' is absent in T'.

Thus, a SCS experiment where we observe one cell from each clone of a distinguishing feature Π of T enables us to conclude that phylogeny T is the true phylogeny. As discussed, every phylogeny T has a *trivial distinguishing feature*, which is composed of all featurettes present in T. Moreover, T may have multiple distinguishing features. Therefore, we must consider the complete set of all distinguishing features, which we call the distinguishing feature family.

Definition 3. The set $\Phi(T, \mathcal{T} \setminus \{T\})$ composed of all distinguishing features of T with respect to $\mathcal{T} \setminus \{T\}$ is a *distinguishing feature family of* T.

Let $(c_0, \ldots, c_n)$ be the outcome of a SCS experiment of k cells, where $c_i \geq 0$ is the number of cells observed of clone i and $\sum_{i=0}^{n} c_i = k$. This experiment is *successful* if, among the k sequenced cells, we observe the clones of at least one distinguishing feature $\Pi \in \Phi(T, \mathcal{T} \setminus \{T\})$ — i.e. $c_i > 0$ for all clones i in some distinguishing feature $\Pi \in \Phi(T, \mathcal{T} \setminus \{T\})$. As discussed, conditioning on frequencies $\mathbf{f}$ and T being the true phylogeny, outcomes $(c_0, \ldots, c_n)$ of SCS experiments of k cells follow a multinomial distribution $\mathrm{Mult}(k, u_0, \ldots, u_n)$ where $\mathbf{u}(T, \mathbf{f}) = [u_i]$ is defined as in (1). Let Y_k denote the event of a successful outcome. We are interested in computing the probability $\Pr(Y_k \mid \mathbf{u}(T, \mathbf{f}))$, which equals the sum of the probabilities of all successful outcomes. More specifically, we want to determine the smallest number k^* of single cells to sequence such that $\Pr(Y_{k^*} \mid \mathbf{u}(T, \mathbf{f}))$ is at least the prescribed confidence level γ (Fig 2).

Problem 2 (SCS Power Calculation for Phylogeny T (T-SCS-PC))**.** Given a set $\mathcal{T}$ of candidate phylogenies and a phylogeny $T \in \mathcal{T}$, frequencies $\mathbf{f}$ and confidence level γ, find the minimum number k^* of single cells needed such that $\Pr(Y_k \mid \mathbf{u}(T, \mathbf{f})) \geq \gamma$.

. In Section A.1 in S1 Text, we prove that the above problem is NP-hard.

Theorem 1. T-SCS-PC is NP-hard.

Multinomial Power Calculation

To solve the T-SCS-PC problem, it suffices to have an algorithm that computes $\Pr(Y_k \mid \mathbf{u}(T, \mathbf{f}))$, which is the probability of concluding that T is the true phylogeny. Using this algorithm we identify k^* by starting from $k = 0$ and simply incrementing k until the corresponding probability $\Pr(Y_k \mid \mathbf{u}(T, \mathbf{f}))$ exceeds the prescribed confidence level γ. In the following, we describe how to efficiently compute $\Pr(Y_k \mid \mathbf{u}(T, \mathbf{f}))$.

Recall that the outcome of a SCS experiment composed of k cells corresponds to a vector $\mathbf{c} = [c_i]$, where $c_i \geq 0$ is the number of cells that we observe from clone i and $\sum_{i=0}^{n} c_i = k$. In a successful outcome $\mathbf{c}$ we observe at least one cell for each featurette in at least one distinguishing feature $\Pi \in \Phi(T, \mathcal{T} \setminus \{T\})$, where $\Phi(T, \mathcal{T} \setminus \{T\})$ is the distinguishing feature family. For brevity, we will write Φ rather than $\Phi(T, \mathcal{T} \setminus \{T\})$.

Let $\mathbf{c}(\Pi, k)$ denote the set of all outcomes where we observe at least one cell for each featurette in a distinguishing feature Π — i.e. $\sum_{i=0}^{n} c_i = k$, and for all $i \in \{0, \ldots, n\}$ it holds that $c_i > 0$ if clone i is a featurette in Π and $c_i \geq 0$ otherwise. The set $\mathbf{c}(\Phi, k)$ of successful outcomes is defined as the union $\bigcup_{\Pi \in \Phi} \mathbf{c}(\Pi, k)$. The probability of any SCS outcome $\mathbf{c} = (c_0, \ldots, c_n)$ is distributed according to $\mathrm{Mult}(k, \mathbf{u}(T, \mathbf{f}))$. Since successful

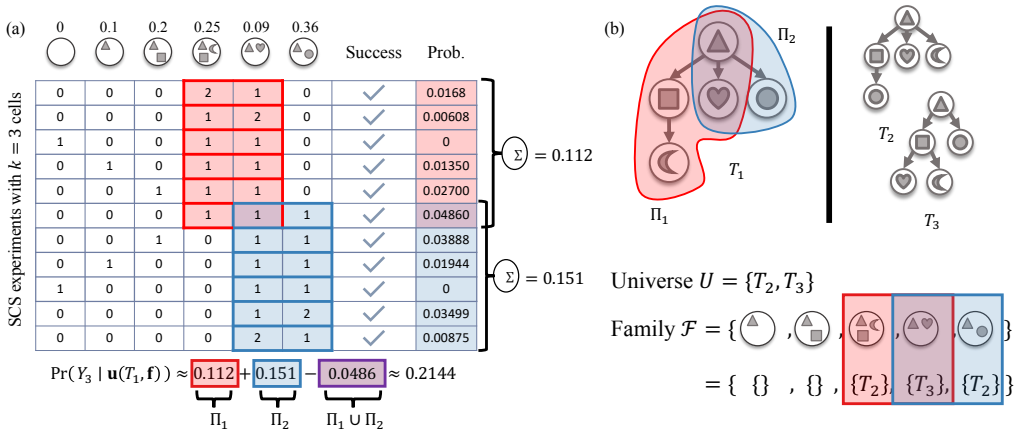


Fig 3. PhyDOSE implementation details. (a) To account for minimal distinguishing features that share featurettes, we use the inclusion-exclusion principle to compute $\Pr(Y_k \mid \mathbf{u}(T, \mathbf{f}))$. Here, Π_1 (red) and Π_2 (blue) share a featurette (with 'triangle' and 'heart' mutations). (b) To enumerate the set Φ^* of minimal distinguishing features of T_1, we reduce the problem to Set Cover and repeatedly identify minimum covers. Here, the universe U is composed of trees $\{T_2, T_3\}$ and there is a subset in $\mathcal{F}$ for each featurette τ of T_1 composed of the trees where τ is absent.

outcomes enable us to conclude that T is the true phylogeny, we have

$$\Pr(Y_k \mid \mathbf{u}(T, \mathbf{f})) = \sum_{\boldsymbol{\ell} \in \mathbf{c}(\Phi, k)} \mathrm{Mult}(\boldsymbol{\ell} \mid k, \mathbf{u}(T, \mathbf{f})) = \sum_{\boldsymbol{\ell} \in \mathbf{c}(\Phi, k)} \frac{k!}{\prod_{i=0}^{n} \ell_i!} \prod_{i=0}^{n} u_i^{\ell_i}. \tag{2}$$

If there is only one distinguishing feature Π, i.e. $\Phi = \{\Pi\}$, then the desired probability is a standard tail probability of the multinomial where we sum up the probabilities of outcomes $\mathbf{c}(\Pi, k) = [c_i]$ such that $\sum_{i=0}^{n} c_i = k$, $c_i > 0$ if clone i is a featurette of Π and $c_i \geq 0$ otherwise. A fast calculation of this tail probability was developed using a connection to the conditional probability of independent Poisson random variables [26, 30]. If there are multiple distinguishing features but they are pairwise disjoint — i.e. no two distinct distinguishing features share the same featurette — then we simply have

$$\Pr(Y_k \mid \mathbf{u}(T, \mathbf{f})) = \sum_{\Pi \in \Phi} \sum_{\boldsymbol{\ell} \in \mathbf{c}(\Pi, k)} \mathrm{Mult}(\boldsymbol{\ell} \mid k, \mathbf{u}(T, \mathbf{f})), \tag{3}$$

and we can apply the fast computation [26] to obtain each independent tail probability. However, the equality in the above equation does not hold if the family Φ is composed of distinguishing features with overlapping featurettes. Incorrectly applying this equation will lead us to overestimate the value of k^*. Since single-cell sequencing is expensive, overestimating the number of cells to sequence in a SCS experiment can be costly and unnecessary. One naive way would be to simply brute force all $(n+1)^k$ SCS outcomes, but this will not scale. Instead, to calculate $\Pr(Y_k \mid \mathbf{u}(T, \mathbf{f}))$ exactly, we propose to use the inclusion-exclusion principle as follows.

$$\Pr(Y_k \mid \mathbf{u}(T, \mathbf{f})) = \sum_{\emptyset \subsetneq \Phi' \subseteq \Phi} (-1)^{|\Phi'|+1} \sum_{\boldsymbol{\ell} \in \mathbf{c}(I(\Phi'), k)} \mathrm{Mult}(\boldsymbol{\ell} \mid k, \mathbf{u}(T, \mathbf{f})), \tag{4}$$

where $I(\Phi')$ is the set of all featurettes in Φ', i.e. $I(\Phi') = \bigcup_{\Pi \in \Phi'} \Pi$ (Fig 3a).

Thus, we need to compute $2^{|\Phi|} - 1$ tail probabilities, which each can be done using the fast calculation in SCOPIT [26].

In the worst case, Φ has $O(2^n)$ distinguishing features resulting in $O(2^n)$ tail probabilities. We now describe one final optimization that will significantly reduce the number of required computations. This is based on the following observation.

Observation 1. If Π is a distinguishing feature of T then for all featurettes τ present in T it holds that $\Pi \cup \{\tau\}$ is a distinguishing feature of T.

This means that distinguishing features in Φ form a partially ordered set under the set inclusion relation. We call a distinguishing feature Π *minimal* if there does not exist another distinguishing feature $\Pi' \in \Phi$ that is a proper subset of Π, i.e. $\Pi' \subsetneq \Pi$.

A direct consequence of Observation 1 is that the outcome of an SCS experiment is successful when we observe all featurettes of a distinguishing feature Π, and remains so even if we observe additional featurettes $\tau' \notin \Pi$.

As such, successful outcomes w.r.t. Φ equal those w.r.t. the set Φ^* of all minimal distinguishing features of T.

Observation 2. It holds that $\mathbf{c}(\Phi^*, k) = \mathbf{c}(\Phi, k)$.

Therefore, it suffices to restrict our attention to only Φ^* rather than the complete family Φ when computing $\Pr(Y_k \mid \mathbf{u}(T, \mathbf{f}))$ using (4). Section B.1 in S1 Text describes how to find Φ^* by reducing the problem to that of finding all minimal set covers, which we solve in an iterative fashion using integer linear programming.

Consideration of SCS Error Rates

One current challenge with SCS is that the false negative rate per site is quite high with typical rates up to 0.4 for the commonly used multiple displacement amplification (MDA) method [31]. On the other hand, current false positive rates are low and are typically less than 0.0005 for MDA-based whole-genome amplification [31]. A *false negative* is defined as not observing a mutation that is present in the cell. A *false positive* occurs when we observe the presence of a mutation that did not occur in that cell.

With PhyDOSE, we propose one possible method for incorporating the false negative rate β when it is known. Specifically, sampled cells follow a categorical distribution $\mathbf{u} = [u_0, \ldots, u_n]$ when conditioned on tree T. Hence, the probability of sampling a cell from clone i equals u_i. True positives, i.e. correctly observing a mutation in a clone, follow a Bernoulli distribution with parameter $1 - \beta$. To observe a featurette/clone i that has n_i mutations and a prevalence of u_i, we thus need to have n_i true positives. In other words, assuming independence among mutations, we require n_i successful draws from a Bernoulli distribution parameterized by $1 - \beta$. As such, we derive new clonal prevalence $\mathbf{u}'(T, \mathbf{f}, \beta) = [u'_i]$ from $\mathbf{u}(T, \mathbf{f}) = [u_i]$. Additionally assuming independence between the events of a cell being sampled from clone i and the absence of false negatives, we set $u'_i = u_i(1 - \beta)^{n_i}$ where n_i is the number of mutations in featurette/clone i. We set u'_0 to be equal to $1 - \sum_{i=1}^{n} u'_i$. This adjustment results in a reduction of the clonal prevalence and ultimately increases the value of k^*. The issue of false positives is less serious as error rates are low enough to be negligible.

Prioritizing Candidate Trees Post SCS Experiment

The final step is to prioritize candidate trees after performing a SCS experiment with the number k^* of cells computed by PhyDOSE. To this end, we compute the *support* of each tree $T \in \mathcal{T}$. Intuitively, support(T) is the number of cells that support the conclusion that T is the actual phylogeny. Formally, we say that a distinguishing feature Π of a tree T is *observed* if each featurette of Π is observed in at least one cell.

ID	% of Trees	Clones	Mutations	Prevalence Noise	FNR β	Doublet δ
sim1a	100%	7	7	0%	0	0
sim1b	10%	7	7	0%	0	0
sim2a	100%	7	7	5%	0	0
sim2b	10%	7	7	5%	0	0
sim2c	100%	7	7	20%	0	0
sim3a	100%	7	7	5%	0.2	0.1
sim3b	10%	7	7	5%	0.2	0.1
sim4a	100%	10	100	5%	0.2	0.1

Table 1. Simulation conditions. We generate simulated data under eight conditions with 100 instances each. These conditions have varying subsets of candidate trees, number of clones, number of mutations per clone, clonal prevalence distortions, false negative and doublet rates.

Using this, we define support(T) as the number of cells that correspond to featurettes of an observed distinguishing feature Π of T. Per Observation 1, it suffices to restrict our attention to the set Π^* of minimal distinguishing features.

There are two outcomes of a SCS experiment with k^* cells. Either there is no tree $T \in \mathcal{T}$ with non-zero support or there are one or more trees with non-zero support. In the former case, the SCS experiment has failed, which is expected to occur with probability $1 - \gamma$. In the latter case, which may occur in the presence of false negatives and false positives, we return the set of trees with maximum support.

Alternatively, we may use existing methods that infer tumor phylogeny from SCS data [14, 16, 32] or a combination of SCS and bulk data [20, 33].

Results

In this section, we demonstrate the application of PhyDOSE to simulated and real data. We begin by validating our method using simulated data. Next, we provide retrospective results for a leukemia patient [23] and an acute myeloid leukemia cohort [34] where both bulk and single-cell DNA sequencing have been performed [23]. Finally, we use PhyDOSE to perform a prospective analysis to determine the required number of single cells to identify the true phylogeny in a non-small cell lung cancer patient cohort [3].

Simulations

Design. To assess the performance of PhyDOSE, we generated simulated data where the ground truth tree T^* is known. Given a fixed number c of clones and n mutations, we first generated a ground truth tree T^* with c vertices uniformly at random using Prüfer sequences [35] and randomly distributed the n mutations to the c clones while ensuring that every clone had at least one mutation. Next, we generated clonal prevalences $\mathbf{u} = [u_i]$ by drawing from a symmetric $n+1$-dimensional Dirichlet distribution with concentration parameter 0.2. We used rejection sampling to ensure that each clonal prevalence u_i was at least 0.05. Let $\sigma(i)$ be the set of clones that contain mutation i. We generated frequencies $\mathbf{f} = [f_i]$ by setting $f_i = \sum_{j \in \sigma(i)} u_j$ for each mutation $i \in \{1, \ldots, n\}$.

To validate our method using *in silico* SCS experiments, we introduced errors related to single-cell sequencing by varying false negative rates $\beta \in \{0, 0.2\}$ and doublet rates $\delta \in \{0, 0.1\}$. We generated, for each simulation instance, $10,000$ single cells sampled either under the specified false negative rates β and doublet rates δ according to the bulk clonal prevalence $\mathbf{u}$. Additionally, we varied single-cell clonal prevalences

from the bulk clonal prevalences by resampling $\hat{\mathbf{u}} \sim \textsc{Dir}(\lambda \mathbf{u})$. We tuned the parameter λ so that the clonal prevalence varied by an absolute average of 5% from the clones of the ground truth tree T^*, which resulted in $\lambda = 2000$. We performed further sensitivity analysis (S2 Fig) on the robustness to clonal prevalence distortion by investigating the impacts of setting $\lambda = 50$, equating to an absolute average difference of 20% between the bulk and single cell clonal prevalence.

In total, we generated 100 simulation instances under eight varying conditions as specified in Table 1. We used the SPRUCE algorithm to enumerate the set $\mathcal{T}$ of trees given frequencies $\mathbf{f}$ [9]. In order to consider violations of the PhyDOSE assumption that the true tree is among the candidate set, for simulation conditions 'b' (sim1b, sim2b and sim3b), we randomly sampled 10% of the trees outputted by SPRUCE [9]. For sim2c, we introduce a clonal prevalence distortion of 20%.

For sim4a, we first used PyClone [36] to cluster the 100 mutations before enumerating the set.

Results. We now compare the number k^* of single cells computed by PhyDOSE to the naive method of requiring all clones to be observed. In addition, we perform *in silico* SCS experiments using PhyDOSE's computed number k^* to assess its accuracy in light of varying sequencing and sampling errors.

The number $|\mathcal{T}|$ of trees in our simulations ranged from 3 to 506 for the full enumerated set of trees (simulation conditions 'a') with a median of 59 trees (Fig 4a). For the set of downsampled trees (simulation conditions 'b'), the number of trees ranged from 2 to 50 with a median of 7 trees (Fig 4a).

We ran PhyDOSE to identify the minimal distinguishing feature family Φ^* for each tree in each simulation instance. This yielded a single minimum distinguishing feature in each case for the simulations instances with fully enumerated tree sets, whereas for the downsampled tree set the number of distinguishing features was between 1 and 6 with a median of 2. Fig 4b shows the number of featurettes in each minimal distinguishing feature identified by PhyDOSE , ranging from 1 to 5 with a median of 3 for conditions 'a' and 2 for conditions 'b'. Importantly, this number is smaller than the total number of 5 featurettes. As such, running PhyDOSE resulted in a median reduction of $\sim 88\%$ over all simulations in the follow-up experimental design compared to the naive approach of requiring all featurettes/clones to be observed (Fig 4c,d).

In particular, with a confidence level of $\gamma = 0.95$, PhyDOSE computed a median number of $k^* = 41$ cells compared to $k^* = 544$ cells computed by the naive method (Fig 4d). With a false negative error rate $\beta = 0.2$, PhyDOSE computed $k^* = 137$ cells compared to $k^* = 720$ cells by the naive method (Fig 4d). . Thus, with increasing false negative rates $\beta \in \{0, 0.2\}$ we observed that (i) PhyDOSE continued to outperform the naive method and (ii) more cells were needed to identify T^*.

Next, we assess the accuracy of PhyDOSE's k^* value for varying single-cell sequencing error conditions. To do so, we ran our approach for prioritizing candidate trees as well as SPhyR [16] on sampled single cells. For the former, we performed 100 experiments for each simulation instance, reporting the number of experiments that successfully recovered the ground truth tree T^*. To test the precision of k^* we also considered the performance of SPhyR when randomly sampling half and double the k^* cells determined by PhyDOSE. S3 Fig shows that the prioritization approach worked well in the case the candidate tree set contains the ground truth tree (median of 89% success rate for conditions sim1a, sim2a and sim3a), while performance dropped when downsampling candidate trees (median success rate of 0% for sim1b, sim2b and sim3b).

Nevertheless, for both sets of simulation conditions, SPhyR [16] was able to identify the true tree in the majority of cases after sampling PhyDOSE computed number k^* of cells. In the cases where the estimated tree T and true tree T^* were different, SPhyR

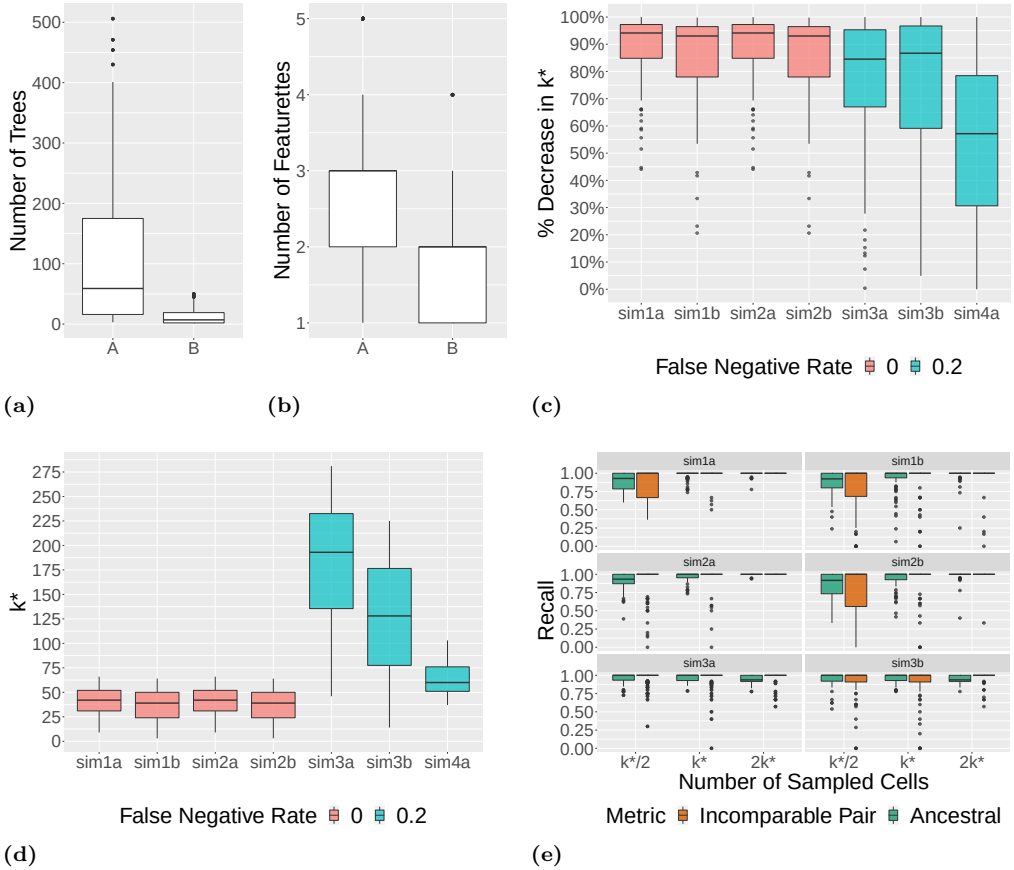


Fig 4. Simulations demonstrate that PhyDOSE's calculated number of single cells resolves tree ambiguity in bulk sequencing data. We used confidence level $\gamma = 0.95$ to determine the number k^* of single cells to sequence. (a) Number $|\mathcal{T}|$ of trees output by SPRUCE [9]. (b) Number $|\Pi|$ of featurettes among minimal distinguishing features Φ^*. (c) Percent decrease in k^* when utilizing PhyDOSE instead of the naive method. (d) Number k^* of cells identified by PhyDOSE. (e) Recall metrics of the tree inferred by SPhyR [16] by randomly sampling $k^*/2$, k^* and $2k^*$ simulated single cells.

still performed well in accordance with two commonly-used tree distance metrics, ancestral and incomparable pair recall. *Ancestral pair recall* is defined as $|\mathcal{A}(T) \cap \mathcal{A}(T^*)|/|\mathcal{A}(T^*)|$ where $\mathcal{A}(T)$ ($\mathcal{A}(T^*)$) is the set of ordered pairs of mutations that occur on distinct edges of the same branch of T (T^*). *Incomparable pair recall* is defined as $|\mathcal{I}(T) \cap \mathcal{I}(T^*)|/|\mathcal{I}(T^*)|$ where $\mathcal{I}(T)$ ($\mathcal{I}(T^*)$)is the set of unordered pairs of mutation that occur on edges in distinct branches in T (T^*). The median of both metrics is 1 when sampling k^* cells (Fig 4e). Additionally, we found greater gains in performance between sampling k^* cells versus $k^*/2$ cells than sampling $2k^*$ cells versus k^* cells, with this trend less pronounced in sim3a and sim3b.

Next, we assessed the impact of clonal prevalence distortion between bulk and single cell data. We found PhyDOSE to be robust to random clonal prevalence noise between bulk and single-cell sequencing as evidenced by a drop of only 1% in the median percentage of successful *in silico* SCS experiment when there is no downsampling of trees (S2 Fig). Additionally, the recall performance metrics (Fig 4e, S2 Fig) are also not substantially different between sim1a versus sim2a and sim2c. We attribute this to the

fact that PhyDOSE's use of distinguishing features critically relies on the clonal prevalence of a few key clones. Furthermore, only when $u_i > \hat{u}_i$ for such a key clone i will PhyDOSE will underestimate $k(T)$, yielding conservative estimates otherwise.

Moreover, we note a drop in the median ancestral pair recall between k^* and $2k^*$ when introducing false negatives and doublets in sim3a and sim3b (Fig. 4e). We attribute this to utilizing the 100th percentile for k^*, setting a confidence level of 0.95 and adjusting for $\beta = 0.2$ leading to highly conservative estimates on the number of cells to sequence. Therefore we do not anticipate a significant change in performance to be evident within these three particular sampling scenarios ($k^*/2$, k^* and $2k^*$). We note that there is similar performance between sim2a/sim2b and sim3a/sim3b at k^*, which likely demonstrates the benefit of adjusting PhyDOSE for the false negative rate. Although PhyDOSE nor SPhyR [16] explicitly account for doublets, tree recall metrics remain high.

Even when the candidate set is significantly downsampled prior to running PhyDOSE, SPhyR still returned trees with high recall performance. There was a small increase in variation between simulations with the full set of candidate trees and the reduced set (Fig 4e). For sim1 and sim2, the distribution of k^* overlaps substantially with a median per replication difference in k^* of 1 (IQR: 0-3). Therefore, similar performance is expected in sim1a vs sim1b and sim2a vs sim2b. The median per replication difference in k^* between sim3a and sim3b was 55 (IQR: 35.5-82). Thus, elevated variation in the recall metrics is expected in those replications where k^* is significantly underestimated.

S4 Fig shows PhyDOSE performance with mutation clusters inferred by PyClone [36]. We additionally included the clustered pair recall in our analysis where clustered pair recall is $|\mathcal{C}(T) \cap \mathcal{C}(T^*)|/|\mathcal{C}(T^*)|$, where $\mathcal{C}(T)$ ($\mathcal{C}(T^*)$) is defined as the set of unordered pairs of mutations that are introduced on the same edge in T (T^*). At k^* cells, the median ancestral pair recall was 0.96, the incomparable pair recall was 0.86 and the clustered pair recall was 0.94, showing a reduction in performance from the first three simulations due to the additional errors introduce by PyClone [36].

In summary, our simulations demonstrate that PhyDOSE's distinguishing feature analysis results in significantly fewer cells to sequence than the naive approach without a subsequent loss in power to identify the true phylogeny. Moreover, we find that PhyDOSE is robust to increasing values of false negatives, doublets and clonal prevalence noise that are typical to real data as well to the case when $T^* \notin \mathcal{T}$.

Running time. Performing the calculation of k^* is fast [26] when $|\Phi| = 1$ and the median of $|\Phi|$ in our simulations was 1. Therefore, the main bottleneck in PhyDOSE is the determination of Φ for each tree in the candidate set as this involves iteratively solving an ILP for each tree until all minimal distinguishing features are found. Note that this step is embarrassingly parallelized because Φ is computed independently for each $T \in \mathcal{T}$. To explore a practical upper bound on $|\mathcal{T}|$, we calculated the runtime in seconds for each simulation instance when Φ is solved sequentially for each $T \in \mathcal{T}$. To additionally explore how PhyDOSE scales with $|\mathcal{T}|$, we generated an additional simulation set of with 10 mutations and no mutations clusters and set a two-hour time limit. The largest sized set was $40,089$ trees with a median of 112 trees. The results are displayed in S5 Fig. Simulations where $|\mathcal{T}| \leq 5000$ completed within the two hour time limit when solved sequentially, suggesting an upper bound on $|\mathcal{T}|$ to use in practice without parallelization. The mean runtime per tree grows linearly with $|\mathcal{T}|$, thus providing a means of estimating total runtime with parallelization and/or large sized inputs.

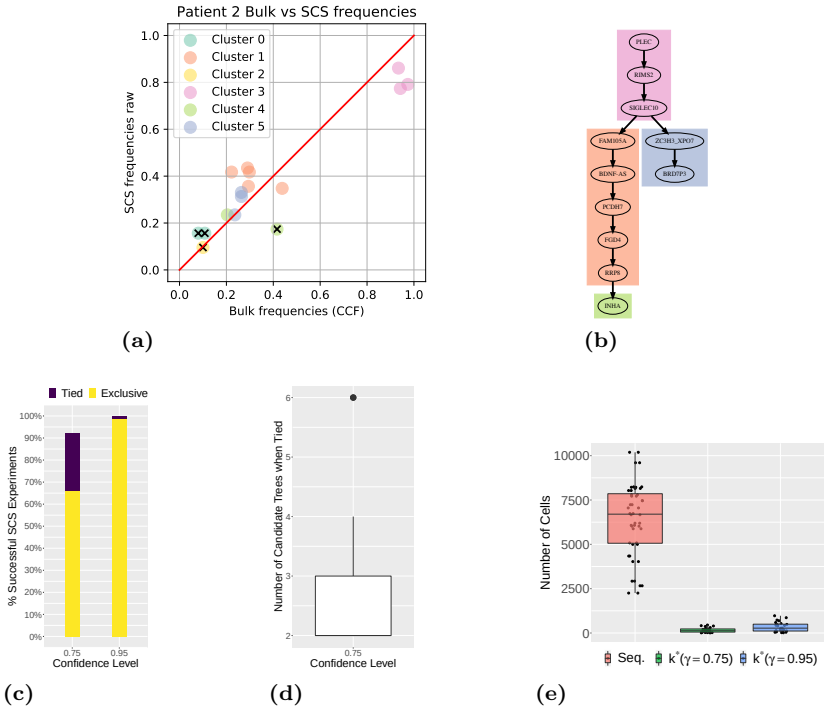


Fig 5. Retrospective analysis of ALL patient 2 [23] and AML cohort [34] demonstrates that fewer cells suffice for replication. Panels (a)-(d) consider ALL patient 2 [23] and panel (e) considers the AML cohort [34]. (a) There is a strong correlation between bulk and single-cell mutation frequencies. Colors indicate mutation clusters from SCS data and excluded mutations are indicated by 'x'. (b) Phylogeny T^* that is consistent with the SCS and bulk data. (c) Percent of successful outcomes in 100 *in silico* SCS experiments, obtained by sampling from the 115 sequenced cells without replacement following PhyDOSE's calculated number $k(T^*)$ of cells (103 for $\gamma = 0.95$ and 50 for $\gamma = 0.75$). Exclusive outcomes (yellow) uniquely identified T^* whereas tied outcomes (purple) yielded a small set of candidate phylogenies that include T^*. (d) Number of candidate phylogenies in the case of ties. (e) The distribution of PhyDOSE's k^* for $\gamma \in \{0.75, 0.95\}$ of all patients in the AML cohort with $|\mathcal{T}| > 2$ as well as the number of cells that were originally sequenced.

Retrospective Analysis of an Acute Lymphoblastic Leukemia Patient

We considered a cohort of six childhood acute lymphoblastic leukemia (ALL) patients whose blood was sequenced using bulk and targeted single-cell DNA sequencing [23]. The number of sequenced single cells per patient varied between 96 and 150. To validate our approach, we used PhyDOSE to calculate the number $k(T^*)$ of cells needed to identify the true phylogeny T^* that is consistent with both data types, thereby retrospectively determining whether fewer single cells suffice to determine T^*, decreasing the cost of replicate experiments. In addition, we assessed whether the calculated number $k(T^*)$ yielded T^* using *in silico* SCS experiments.

Due to the absence of published copy-number aberration information for this dataset, we focused our attention on patient 2 whose single-cell phylogeny adhered to the infinite sites assumption and the variant allele frequencies suggested the absence of copy-number aberrations (as detailed in Section C in S1 Text). For this patient, 16 autosomal mutations in 115 cells were sequenced [23]. We note that the authors had no

knowledge of the number of cells that would suffice to infer the tumor phylogeny of the patient. Using the infinite sites assumption and assuming the absence of copy-number aberrations, we define the cancer cell fraction, or frequency f_i of each mutation i in the bulk data as $2 \cdot \text{VAF}(i)$. We define the *SCS mutation frequency* as the fraction of single cells that harbor the mutation. Strikingly, there is a clear correlation between the bulk and SCS mutation frequencies, supporting PhyDOSE's first assumption (Fig 5a). We excluded mutation *CMTM8* because of a notable discrepancy in frequencies (0.4 in bulk vs. 0.2 in SCS). Using SPRUCE [9], we enumerated the set $\mathcal{T}$ of trees from the bulk data, yielding over 2.5 million trees. This number is mainly driven by 3 mutations (*ATRNL1*, *LINC00052* and *TRRAP*) with a VAF less than 0.05. Excluding these 3 mutations resulted in a more tractable number of 2576 trees. We note that in practice we may similarly exclude mutations because of very low VAFs or less importance in downstream analyses. Fig 5b shows the single tree $T^* \in \mathcal{T}$ that was consistent with the cleaned single-cell data, supporting PhyDOSE's second assumption.

We ran PhyDOSE using varying confidence levels $\gamma \in \{0.75, 0.95\}$ and an estimated false negative rate of $\beta = 0.2$ reported by the authors [23]. PhyDOSE calculated that $k(T^*) = 103$ cells suffice to identify T^* with confidence level $\gamma = 0.95$. Indeed, performing 100 *in silico* SCS experiments, by sampling $k(T^*)$ cells among the 115 sequenced cells without replacement, yielded a success rate of 99% (Fig 5c).

To reduce costs, we explored what would have happened retrospectively with a lower confidence level γ of 0.75. PhyDOSE calculated that $k(T^*) = 50$ cells are needed for $\gamma = 0.75$, which is a significant cost savings over $\gamma = 0.95$. Performing 100 *in silico* SCS experiments yielded a success rate of uniquely identifying T^* of 66%, which was lower than the expected rate of 75%. Furthermore, we noted that in an additional 26% of experiments the correct phylogeny T^* was among the trees with the highest overall support (Fig 5c). The number of trees in the tied set of successes varied from 2 to 6 (S7d Fig), showing that although PhyDOSE did not uniquely identify the tree, it was able to significantly reduce the original set of 2576 trees (S7 Fig and S8 Fig).

In summary, this retrospective analysis shows that the true tree for patient 2 could have been identified confidently with fewer cells than the 115 cells initially sequenced [23]. With a lower confidence level γ, PhyDOSE computes that far fewer cells are required, significantly reducing costs but at the expense of a lower success rate of uniquely identifying the true phylogeny. Nevertheless, the resulting SCS experiment will eliminate a large fraction of the original set of candidate phylogenies due to the incorporation of distinguishing features in the PhyDOSE power calculation.

Retrospective Analysis of an Acute Myeloid Leukemia Cohort

In [34], high-throughput targeted microfluidic single-sequencing was performed on a cohort of 77 patients with acute myeloid leukemia (AML). The authors additionally performed bulk sequencing in order to confirm the presence of a mutation in the single-cell data. We note that the authors restricted their analysis to somatic mutations (SNVs and indels) that did not occur in regions affected by additional copy number aberrations.

Here, we utilized the published bulk sequencing VAFs of the SNVs in each patient, eliminating any mutations not detected via bulk sequencing, to enumerate a set of candidate trees using SPRUCE [9]. The median number of mutations per patient was 4 (IQR: 3-5). We retrospectively used PhyDOSE at confidence levels $\gamma \in \{0.75, 0.95\}$ to estimate the cells needed to perform an equivalent single-cell experiment. The mean of the per patient published false negative rate ($\beta = 4.9\%$) was used to estimate the system error *a priori*. In the original study, a median of 7,584 cells per patient (IQR: 6,194-8,361) were sequenced. Fig 5e shows the distribution of PhyDOSE k^* for all patients with greater than one tree in the candidate set (median is 2, IQR: 2-6, max is

316) at $\gamma \in \{0.75, 0.95\}$ versus the total number of cells sequenced in [34]. For $\gamma = 0.95$, the median value of k^* was 274 cells (IQR: 230-497). This is a significant reduction from the number of cells sequenced per patient in [34] with a median percent reduction at confidence level $\gamma = 0.95$ of 95.4% (IQR: 92.2%-98.0%) necessary to replicate the experiment with similar results (S1 Table).

Prospective Analysis of a Non-small Cell Lung Cancer Cohort

Using PhyDOSE, we prospectively determined the number of cells needed to uniquely identify the true phylogeny for the 25 out of 100 patients in the TRACERx non-small-cell lung cancer cohort that have multiple candidate trees [3]. The authors previously identified the set $\mathcal{T}$ of candidate trees for each patient using CITUP [11] after clustering mutations with PyClone [36]. The authors also reported the cancer cell fraction of each mutation cluster in each bulk sample. The number of trees in the candidate set for each patient ranged from 2 to 17, with each containing mutation clusters with between 5 and 882 mutations (S2 Table).

Unlike in the simulations and ALL patient 2, multiple bulk samples per patient were available for analysis. Therefore, we calculated k^* for each sample independently for all 25 patients at varying confidence level $\gamma \in \{0.75, 0.95\}$. Mutation clusters alleviate the issue of false negatives, i.e. it suffices to only observe a single mutation to impute the presence of the other mutations in the same cluster. Here, with a typical SCS false negative rate of 0.2, the probability of all mutations in the smallest cluster (with size 5) dropping out thus equals $0.2^5 = 0.00032$, a probability that can be neglected. As such, we set $\beta = 0$. The reported k^* value is the minimum k^* over the set of available samples, subsequently identifying which of the samples is the best to utilize for the SCS experiment. PhyDOSE was able to return a finite value of k^* for 23 out of the 25 patients.PhyDOSE will return ∞ when for each sample of the patient there is a featurette in every distinguishing feature where the clonal prevalance is 0. For two of the 23 patients the calculated k^* was over 400 due to featurettes in the distinguishing features with low clonal prevalence. For the remaining 21 patients, the median value of k^* was 29 for $\gamma = 0.95$ and 14 for $\gamma = 0.75$ (S9 Fig). These strikingly low values of k^* for the majority of the 25 patients with multiple candidate trees demonstrate the benefit of using PhyDOSE to strategically optimize the design of follow-up single cell experiments.

Discussion

In this work, we showed that the mutation frequencies $\mathbf{f}$ and the set $\mathcal{T}$ of tumor phylogenies inferred from initial bulk data contain valuable information to provide guidance for follow-up SCS experiments. We introduced PhyDOSE, a method to calculate the number k^* of single cells needed to infer the true phylogeny T^* given $\mathbf{f}$, $\mathcal{T}$ and a user-specified confidence level γ. Underpinning our method is the observation that often only a subset of clones suffices to distinguish one tree $T \in \mathcal{T}$ from the remaining trees $\mathcal{T} \setminus \{T\}$.

We validated PhyDOSE using simulations and a retrospective analysis of leukemia patients [23, 34], concluding that PhyDOSE's computed number k^* of cells resolves tree ambiguity, even in the presence of SCS errors. Our simulations showed that PhyDOSE remains robust in the presence of SCS errors such as elevated false negative and doublet rates as well as deviations between the distribution of clones in bulk and single-cell samples. Moreover, even in the case of an incomplete candidate tree set, PhyDOSE's computed number k^* of cells sufficed to recover the true tree in a follow-up SCS experiment. In a prospective analysis, we demonstrated that only a small number of

cells suffice to disambiguate the solution space of trees in a recent non-small cell lung cancer cohort [3]. Finally, we introduced the notion of support as a way to prioritize candidate trees given SCS data, often correctly selecting the ground truth tree among the candidate trees. Existing methods that infer tumor phylogeny from SCS data [14, 16, 32] or a combination of SCS and bulk data [20, 33] may be used as an alternative. In summary, PhyDOSE proposes cost-efficient SCS experiments that will yield high-fidelity phylogenies, which may consequently improve downstream analyses in cancer genomics aimed at deepening our understanding of cancer biology.

There are several future research directions. First, in the case of multiple bulk samples, rather than selecting cells from a single sample, a better strategy would be to select cells across samples. To model this accurately, we must consider a multinomial mixture model. Second, to further reduce SCS costs, we might want to include a mutation selection step as part of our approach to perform targeted rather than whole-genome sequencing. Third, similar ideas can be used to design follow-up sequencing experiments using alternative sequencing technologies such as long read sequencing. Alternatively, performing additional bulk sequencing rather than single-cell sequencing might be more cost-effective, especially when obtaining a bulk sample with distinct clonal prevalences [10, 37]. Fourth, we plan to replace the integer linear program used for identifying minimal distinguishing features with a combinatorial algorithm. This will enable us to develop an easy-to-use and install R package with a Shiny user interface. Fifth, to improve robustness in the presence of SCS errors, we plan to explore alternative definitions of successful SCS experiment outcomes, requiring that more than one cells is observed of each featurette of a distinguishing feature. This will enable us to address errors such as doublets and false positives in a SCS experiment. Sixth, the concept of distinguishing features may be useful to summarize diverse solution spaces in cancer phylognetics [38]. Finally, we plan to explore evolutionary models beyond the infinite sites model, such as the Dollo parsimony model where mutations might be lost [16], requiring a more careful definition of the probability of observing a distinguishing feature.

Supporting information

S1 Fig. Reduction from Set Cover to T-SCS-PC. Given a family $\mathcal{F}$ of subsets $\{S_1, \ldots, S_{|\mathcal{F}|}\}$ on a universe $U = \{1, \ldots, n\}$, we construct $n+1$ trees $\mathcal{T} = \{T_0, \ldots, T_n\}$ with mutations $\{\top, \bot, S_1, \ldots, S_{|\mathcal{F}|}\}$. We seek to distinguish T_0 from the remaining trees $\{T_1, \ldots, T_n\}$. The key concept captured by the reduction is that there is a cover of size k if and only if $\Pr(Y_k \mid \mathbf{u}(T_0, \mathbf{f}))$ is greater than 0. Here, S_1 and S_4 form a cover of size $k = 2$ of the universe U and the corresponding probability $\Pr(Y_k \mid \mathbf{u}(T_0, \mathbf{f}))$ is greater than 0.

S2 Fig. PhyDOSE sensitivity analysis of clonal prevalence distortion (a) The mean absolute percentage difference of the single cell clonal prevalence per replication from simulated bulk clonal prevalence at values of $\lambda \in \{50, 2000\}$. These values of λ resulted in mean absolute percentage difference of 5% and 20% respectively. (b) $\lambda = 2000$ was selected for sim2a/b but further sensitivity analysis was performed with $\lambda = 50$. The recall metrics are compared between sim2a at $\lambda = 2000$ and sim2c at $\lambda = 50$. (b) Recall metrics when inferring T^* with SPhyR [16] by randomly sampling $k^*/2, k^*, 2k^*$ simulated single cells.

S3 Fig. Percentage of *in silico* single-cell sequencing experiments correctly identifying T^* as the true tree utilizing PhyDOSE's k^* at γ =0.95 and $\beta \in \{0.0, 0.2\}$. PhyDOSE has a high success rate in alignment with $\gamma = 0.95$

(dashed horizontal line) when utilizing the entire set $\mathcal{T}$ in condition a although the introduction of false negatives results in a reduction in success rate and greater variance. In the cases where $T^* \in \mathcal{T}$, the support metric was still able to successfully identify T^* as the true tree.

S4 Fig. Empirical runtime analysis on the bottleneck step of PhyDOSE (a) Total runtime in seconds for each input set $\mathcal{T}$ that completed in the two hour limit when solved sequentially. (b) Fraction of the input set $\mathcal{T}$ completed within the two hour limit when solved sequentially. (c) Mean runtime in seconds per tree $T \in \mathcal{T}$.

S6 Fig. Retrospective analysis of ALL patient 2 where [23] sequenced 115 single cells, targeting 16 mutations. (a) Overview of single-cell data, focusing on a subset of 14 mutations. Each row is a single cell and columns correspond to mutations. Black cells indicate presence of the SNV, whereas white cells indicate absence. Text in each cell indicates a true positive ('1'), false negative ('1*'), true negative ('0') or false positive ('0*') entry. (b) Overview of the set $\mathcal{T}$ of trees, where each edge is labeled by its frequency in $\mathcal{T}$. Red edges encode the tree T^* that is consistent with SCS data in (a). In particular, the SCS data encodes a tree composed of four mutation clusters, each indicated by a different shading.

S7 Fig. Distribution of number of supporting cells for each candidate tree across 100 *in silico* SCS experiments at varying success probabilities for ALL patient 2 [23]. For $\gamma = 0.75$, we sampled $k(T^*) = 50$ cells from the SCS data. For $\gamma = 0.95$, we sampled $k(T^*) = 103$ cells. S8 Fig shows the candidate trees.

S8 Fig. Candidate trees that had a non-zero support across 100 *in silico* SCS experiments for ALL patient 2 [23]. For $\gamma = 0.75$, we sampled $k(T^*) = 50$ cells from the SCS data. For $\gamma = 0.95$, we sampled $k(T^*) = 103$ cells. Labels match tree indices in S7 Fig.

S1 Table Prospective analysis of TRACERx non-small-cell lung cancer cohort. Table shows the patient identifier, the number of mutation clusters, the number of bulk samples, the minimum number of mutations per cluster, the maximum number of mutations per cluster, the size of the candidate set $\mathcal{T}$ of trees as determined by [3], PhyDOSE's k^* calculated at confidence levels of $\gamma \in \{0.75, 0.95\}$ and the recommended sample label from which the single cells should be drawn.

S2 Table Retrospective analysis of an acute myeloid leukemia (AML) cohort. Table shows the patient identifier, the number of mutations, the total cells sequenced [34], the size of the candidate set $\mathcal{T}$ of trees as determined by SPRUCE [9], PhyDOSE's k^* calculated at confidence levels of $\gamma \in \{0.75, 0.95\}$ (% reduction from total sequenced).

S9 Fig. PhyDOSE calculated k^* for the lung cancer cohort at varying confidence levels. Patients CRUK0013, CRUK0037, CRUK0068, CRUK0076 were excluded from the plot, but are shown in S1 Table.

S1 Text Supplementary materials.

Acknowledgments

This work was supported by UIUC Center for Computational Biotechnology and Genomic Medicine (grant: CSN 1624790) and the National Science Foundation (grant: CCF 1850502).

References

1. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976;194(4260):23–8.
2. McGranahan N, Favero F, de Bruin EC, Birkbak NJ, Szallasi Z, Swanton C. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. Science Translational Medicine. 2015;7(283):283ra54–283ra54.
3. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TB, Veeriah S, et al. Tracking the evolution of non–small-cell lung cancer. New England Journal of Medicine. 2017;376(22):2109–2121.
4. Zhang AW, McPherson A, Milne K, Kroeger DR, Hamilton PT, Miranda A, et al. Interfaces of Malignant and Immunologic Clonal Dynamics in Ovarian Cancer. Cell. 2018;173(7):1755–1769.e22.
5. Łuksza M, Riaz N, Makarov V, Balachandran VP, Hellmann MD, Solovyov A, et al. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. Nature. 2017;551(7681):517.
6. Turajlic S, Xu H, Litchfield K, Rowan A, Horswell S, Chambers T, et al. Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal. Cell. 2018;.
7. Turajlic S, Xu H, Litchfield K, Rowan A, Chambers T, Lopez JI, et al. Tracking Cancer Evolution Reveals Constrained Routes to Metastases: TRACERx Renal. Cell. 2018;0(0).
8. Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biology. 2015;16(1):35. doi:10.1186/s13059-015-0602-8.
9. El-Kebir M, Oesper L, Acheson-Field H, Raphael BJ. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. Bioinformatics. 2015;31(12):i62–i70.
10. El-Kebir M, Satas G, Oesper L, Raphael BJ. Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures. Cell Systems. 2016;3(1):43–53.
11. Malikic S, McPherson AW, Donmez N, Sahinalp CS. Clonality Inference in Multiple Tumor Samples using Phylogeny. Bioinformatics. 2015;doi:10.1093/bioinformatics/btv003.
12. Yuan K, Sakoparnig T, Markowetz F, Beerenwinkel N. BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. Genome Biology. 2015;16(1):36.

13. Popic V, Salari R, Hajirasouliha I, Kashef-Haghighi D, West RB, Batzoglou S. Fast and scalable inference of multi-sample cancer lineages. Genome biology. 2015;16(1):91.

14. Jahn K, Kuipers J, Beerenwinkel N. Tree inference for single-cell data. Genome Biology. 2016;17(1):86.

15. Ross EM, Markowetz F. OncoNEM: inferring tumor evolution from single-cell sequencing data. Genome Biology. 2016;17(1):69.

16. El-Kebir M. SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. Bioinformatics. 2018;34(17):i671–i679.

17. Zafar H, Tzen A, Navin N, Chen K, Nakhleh L. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. Genome biology. 2017;18(1):178.

18. Navin NE. Cancer genomics: one cell at a time. Genome Biology. 2014;15(8):452.

19. Malikic S, Jahn K, Kuipers J, Sahinalp SC, Beerenwinkel N. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. Nature communications. 2019;10(1):1–12.

20. Malikic S, Mehrabadi FR, Ciccolella S, Rahman MK, Ricketts C, Haghshenas E, et al. PhISCS: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. Genome Research. 2019;29(11):1860–1877.

21. Kuboki Y, Fischer CG, Guthrie VB, Huang W, Yu J, Chianchiano P, et al. Single-cell sequencing defines genetic heterogeneity in pancreatic cancer precursor lesions. The Journal of Pathology. 2019;247(3):347–356.

22. Leung ML, Davis A, Gao R, Casasent A, Wang Y, Sei E, et al. Single cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. Genome Research. 2017; p. gr.209973.116.

23. Gawad C, Koh W, Quake SR. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. Proceedings of the National Academy of Sciences. 2014;111(50):17947–17952. doi:10.1073/pnas.1420822111.

24. McPherson A, Roth A, Laks E, Masud T, Bashashati A, Zhang AW, et al. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. Nature Genetics. 2016;.

25. Kim C, Gao R, Sei E, Brandt R, Hartman J, Hatschek T, et al. Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. Cell. 2018;173(4):879–893.

26. Davis A, Gao R, Navin NE. SCOPIT: sample size calculations for single-cell sequencing experiments. BMC bioinformatics. 2019;20(1):566.

27. Bolli N, Avet-Loiseau H, Wedge DC, Van Loo P, Alexandrov LB, Martincorena I, et al. Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. Nature communications. 2014;5.

28. Dentro SC, Wedge DC, Van Loo P. Principles of Reconstructing the Subclonal Architecture of Cancers. Cold Spring Harbor Perspectives in Medicine. 2017;7(8):a026625.

29. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, et al. The landscape of cancer genes and mutational processes in breast cancer. Nature. 2012;486(7403):400–404.

30. Levin B. A Representation for Multinomial Cumulative Distribution Functions. The Annals of Statistics. 1981;9(5):1123–1126.

31. Fu Y, Li C, Lu S, Zhou W, Tang F, Xie XS, et al. Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(38):11923–11928.

32. Zafar H, Navin N, Chen K, Nakhleh L. SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. Genome research. 2019;29(11):1847–1859.

33. Malikic S, Jahn K, Kuipers J, Sahinalp SC, Beerenwinkel N. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. bioRxiv. 2017;doi:10.1101/234914.

34. Morita K, Wang F, Jahn K, Kuipers J, Yan Y, Matthews J, et al. Clonal Evolution of Acute Myeloid Leukemia Revealed by High-Throughput Single-Cell Genomics. bioRxiv. 2020;doi:10.1101/2020.02.07.925743.

35. Prüfer H. Neuer beweis eines satzes uber permutationen. Arch Math Phys. 1918;27:742–4.

36. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, et al. PyClone: statistical inference of clonal population structure in cancer. Nature methods. 2014;11(4):396–398.

37. Qi Y, Pradhan D, El-Kebir M. Implications of non-uniqueness in phylogenetic deconvolution of bulk DNA samples of tumors. Algorithms for Molecular Biology. 2019;14(1):23–14.

38. Aguse N, Qi Y, El-Kebir M. Summarizing the solution space in tumor phylogeny inference by multiple consensus trees. Bioinformatics. 2019;35(14):i408–i416.

Supplementary Text — PhyDOSE: Design of Follow-up Single-cell Sequencing Experiments of Tumors

Leah Weber [1,*] Nuraini Aguse [1,*] Nicholas Chia [2,3] Mohammed El-Kebir [1,†]

[1]Dept. of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801
[2]Microbiome Program, Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905
[3]Division of Surgical Research, Department of Surgery, Mayo Clinic, Rochester, MN 55905

[*]Shared first authorship; [†]Corresponding author: melkebir@illinois.edu

Contents

A Supplementary Text **2**
A.1 Complexity 2

B Supplementary Methods **4**
B.1 Finding the Minimal Distinguishing Feature Family Φ^* 4

C Supplementary Results **5**
C.1 Retrospective Analysis of an Acute Lymphoblastic Leukemia Patient 9

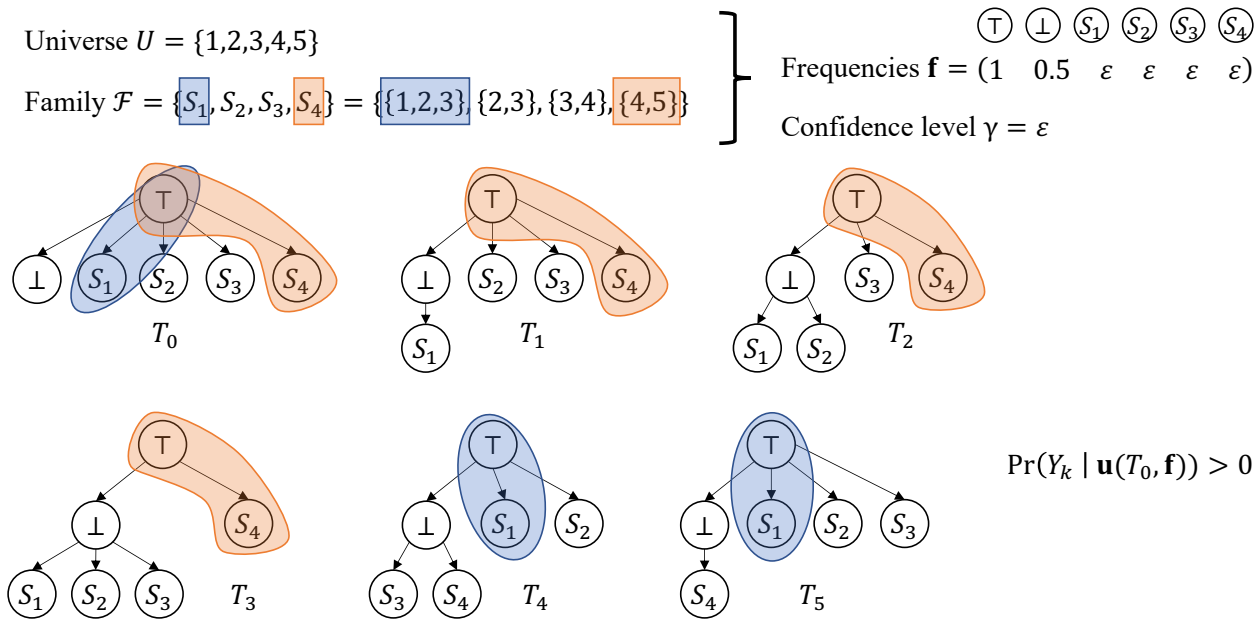


Fig S1: **Reduction from SET COVER to T-SCS-PC**. Given a family $\mathcal{F}$ of subsets $\{S_1, \ldots, S_{|\mathcal{F}|}\}$ on a universe $U = \{1, \ldots, n\}$, we construct $n+1$ trees $\mathcal{T} = \{T_0, \ldots, T_n\}$ with mutations $\{\top, \bot, S_1, \ldots, S_{|\mathcal{F}|}\}$. We seek to distinguish T_0 from the remaining trees $\{T_1, \ldots, T_n\}$. The key concept captured by the reduction is that there is a cover of size k if and only if $\Pr(Y_k \mid \mathbf{u}(T_0, \mathbf{f}))$ is greater than 0. Here, S_1 and S_4 form a cover of size $k = 2$ of the universe U and the corresponding probability $\Pr(Y_k \mid \mathbf{u}(T_0, \mathbf{f}))$ is greater than 0.

A Supplementary Text

A.1 Complexity

Theorem 1. T-SCS-PC is NP-hard.

We prove the theorem using a polynomial-time reduction from the SET COVER problem, a known NP-hard problem [1].

Problem 1 (SET COVER)**.** Given a family $\mathcal{F}$ of subsets $\{S_1, \ldots, S_{|\mathcal{F}|}\}$ over a universe $U = \{1, \ldots, n\}$, find a cover $C \subseteq \mathcal{F}$ such that $\bigcup_{S \in C} S = U$ and C has minimum cardinality.

Specifically, we reduce a SET COVER instance $(\mathcal{F}, U)$ to an T-SCS-PC instance $(\mathcal{T}, T, \mathbf{f}, \gamma)$ as follows. The set $\mathcal{T} = \{T_0, \ldots, T_n\}$ includes one tree T_i for each element i in the universe U and an additional tree T_0. All trees in $\mathcal{T}$ have $|\mathcal{F}| + 2$ vertices, corresponding to subsets $\{S_1, \ldots, S_{|\mathcal{F}|}\}$ and two additional mutations $\{\top, \bot\}$. Each tree in $\mathcal{T}$ includes the edge $(\top, \bot)$. Additionally, if element $i \in U$ is absent from subset S_j then there is an edge $(\top, S_j)$ in tree T_i, otherwise T_i includes an edge $(\bot, S_j)$. Tree T_0 includes edges $(\top, S_j)$ for all subsets S_j. As for the frequencies $\mathbf{f}$, we set $f_\top = 1$, $f_\bot = 0.5$ and the remaining frequencies $f_{S_j} = \varepsilon$ for all subsets $S_j \in \mathcal{F}$. Moreover, we set the confidence level γ to ε as well. In the corresponding T-SCS-PC instance $(\mathcal{T}, T_0, \mathbf{f}, \varepsilon)$, the tree of interest is T_0. Fig. S1 shows an example.

The key idea is that as $\gamma = \varepsilon > 0$ is a small positive infinitesimal constant, this T-SCS-PC instance seeks the smallest number k^* of cells such that $\Pr(Y_{k*} \mid \mathbf{u}(T_0, \mathbf{f}))$ is non-zero. In particular, this number k^* of cells will only be achieved if there is a distinguishing feature Π of the same size k^*. By our reduction, there is a 1-1 correspondence between set covers of U and distinguishing features Π of T_0 with respect to $\{T_1, \ldots, T_n\}$. Specifically, a set cover C of size k corresponds to a distinguishing feature $\Pi(C)$ of the same size k, and vice versa. As such, we have the following lemma whose proof is in the supplement.

Lemma 1. Let $(\mathcal{T}, T_0, \mathbf{f}, \gamma = \varepsilon)$ be the T-SCS-PC instance corresponding to SET COVER instance $(U, \mathcal{F})$. A minimum cover has size k^* if and only if k^* is the smallest integer such that $\Pr(Y_{k^*} \mid \mathbf{u}(T_0, \mathbf{f})) \geq \gamma$.

Proof. ($\Rightarrow$) Let C be a minimum cover of the SET COVER instance $(U, \mathcal{F})$. By the premise, we have that $|C| = k^*$. We start by showing that $\Pr(Y_{k^*} \mid \mathbf{u}(T_0, \mathbf{f})) \geq \gamma$ by constructing a distinguishing feature $\Pi(C)$ of T_0 where $|\Pi| = k^*$. Observe that for each subset S_j in C we have that $\{\top, S_j\}$ is a featurette of T_0. We define $\Pi(C)$ to be composed of featurettes $\{\top, S_j\}$ for all subsets $S_j \in C$. Thus, $|\Pi(C)| = k^*$. To show that $\Pi(C)$ is a distinguishing feature of T_0, it remains to show that at least one featurette $\tau \in \Pi(C)$ is absent in each tree in $\mathcal{T} \setminus T_0 = \{T_1, \ldots, T_n\}$. Consider any tree $T_i \neq T_0$. Since C is a cover, the element i of the universe U corresponding to tree T_i must be covered by some subset $S_j \in C$. This means that tree T_i contains the edge $(\perp, S_j)$, which means that the featurette $\{\top, S_j\}$ in $\Pi(C)$ is absent from T_i. Hence, $\Pi(C)$ is a distinguishing feature of T_0.

We now must show that $\Pr(Y_{k^*} \mid \mathbf{u}(T_0, \mathbf{f})) \geq \gamma$. We do so by focusing on distinguishing feature $\Pi(C)$. By construction of T_0 and $\mathbf{f}$, it follows from (1) that each featurette $\{\top, S_j\}$ in $\Pi(C)$ has a clonal prevalence $u_j = \varepsilon$. This means that a SCS experiment of k^* cells where we only observe the k^* featurettes/clones has a probability that is strictly greater than 0. Therefore, $\Pr(Y_{k^*} \mid \mathbf{u}(T_0, \mathbf{f})) > 0$. Since ε is a small positive infinitesimal constant, we have that $\Pr(Y_{k^*} \mid \mathbf{u}(T_0, \mathbf{f})) \geq \gamma = \varepsilon$.

It remains to show that k^* is the smallest integer where $\Pr(Y_{k^*} \mid \mathbf{u}(T_0, \mathbf{f})) \geq \varepsilon$. Assume for a contradiction that the smallest integer k' where $\Pr(Y_{k'} \mid \mathbf{u}(T_0, \mathbf{f})) \geq \varepsilon$ is strictly smaller than k^*. This means that there exists a minimal distinguishing feature Π' of size at most k'. By definition Π' is composed of featurettes corresponding to root-to-vertex paths in T_0. Since Π' is minimal, it will not contain the featurette $\{\top, \perp\}$ as this featurette is present in all remaining trees $\{T_1, \ldots, T_n\}$. Thus, Π' is composed of featurettes of the form $\{\top, S_j\}$ where $S_j \in \mathcal{F}$. Since Π' is a distinguishing feature, no tree $T_i \in \{T_1, \ldots, T_n\}$ contains all featurettes of Π'. By construction of $\{T_1, \ldots, T_n\}$, this means that the subsets encoded in Π' form a cover of the universe U. Thus, there exists a cover with size strictly smaller than k^*, contradicting the premise. Therefore, k^* is indeed the smallest integer where $\Pr(Y_{k^*} \mid \mathbf{u}(T_0, \mathbf{f})) \geq \gamma = \varepsilon$.

($\Leftarrow$) Let k^* be the smallest integer such that $\Pr(Y_{k^*} \mid \mathbf{u}(T_0, \mathbf{f})) \geq \gamma = \varepsilon$. We start by showing that the size of a minimum distinguishing feature Π of T_0 has to be exactly k^*. Clearly, if $|\Pi| > k^*$ then $\Pr(Y_{k^*} \mid \mathbf{u}(T_0, \mathbf{f})) = 0$ as there exists no successful SCS experiment with k^* cells. On the other hand, if $|\Pi| < k^*$ then there exists a successful SCS experiment with $|\Pi|$ cells. In other words, $\Pr(Y_{|\Pi|} \mid \mathbf{u}(T_0, \mathbf{f})) \geq \varepsilon$. This contradicts that k^* is the smallest integer where $\Pr(Y_{|\Pi|} \mid \mathbf{u}(T_0, \mathbf{f})) \geq \varepsilon$. Hence, $|\Pi| = k^*$.

Consider a minimum distinguishing feature Π of T_0. By the previous argument, we know that $|\Pi| = k^*$. We will show that Π encodes a cover $C(\Pi)$ of U of size k^*. Since Π is minimal, it will not contain the featurette $\{\top, \perp\}$ of T_0 as this featurette is present in all remaining trees $\{T_1, \ldots, T_n\}$. Thus, Π is composed of k featurettes of the form $\{\top, S_j\}$ where $S_j \in \mathcal{F}$. Let $C(\Pi)$ be defined as the collection of subsets $S_j \in \mathcal{F}$ where $\{\top, S_j\}$ in Π. Since Π is a distinguishing feature, no tree $T_i \in \{T_1, \ldots, T_n\}$ contains all featurettes of Π. By construction of $\{T_1, \ldots, T_n\}$, this means that $C(\Pi)$ is a cover of size k of the universe U.

Finally, we must show that there exists no cover C' of U with size $|C'|$ strictly smaller than k^*. Suppose for a contradiction that such a cover C' exists. By construction, C' encodes a distinguishing feature $\Pi(C')$ composed of featurettes $\{\top, S_j\}$ for all subsets $S_j \in C'$. Thus, $|\Pi(C')| = |C'|$. To show that $\Pi(C')$ is a distinguishing feature of T_0, we must show that (i) all features $\tau \in \Pi(C')$ are present in T_0, and (ii) at least one featurette $\tau \in \Pi(C)$ is absent in each tree in $\mathcal{T} \setminus T_0 = \{T_1, \ldots, T_n\}$. Condition (i) holds by construction of $\Pi(C')$ and T_0, i.e. for each subset S_j in C' we have that $\{\top, S_j\}$ is a featurette of T_0. As for condition (ii), consider any tree $T_i \neq T_0$. Since C' is a cover, the element i of the universe U corresponding to tree T_i must be covered by some subset $S_j \in C'$. This means that tree T_i contains the edge $(\perp, S_j)$, which means that the featurette $\{\top, S_j\}$ in $\Pi(C')$ is absent from T_i. Hence, $\Pi(C')$ is a distinguishing feature of T_0. This in turn means that $\Pr(Y_{|\Pi(C')|} \mid \mathbf{u}(T_0, \mathbf{f})) > 0$. In other words, $\Pr(Y_{|\Pi(C')|} \mid \mathbf{u}(T_0, \mathbf{f})) \geq \gamma = \varepsilon$, thus contradicting the premise. Hence, minimum set covers of $(U, \mathcal{F})$ have cardinality k^*. $\square$

The theorem follows from the above lemma, as the reduction to obtain $(\mathcal{T}, T_0, \mathbf{f}, \gamma = \varepsilon)$ from $(U, \mathcal{F})$ takes only polynomial time.

B Supplementary Methods

B.1 Finding the Minimal Distinguishing Feature Family Φ^*

To perform the calculation in (4), it is necessary to first find the minimal distinguishing feature family Φ^*. Using similar ideas as in our hardness proof (Section A.1), we consider the reverse reduction from the problem of finding a minimal distinguishing feature to that of finding a minimum size set cover (Problem 1).

We define the universe $U = \{1, \ldots, m\}$ to be the set $\mathcal{T} \setminus \{T\} = \{T_1, \ldots, T_m\}$ of trees excluding the tree T for which we want to solve the T-SCS-PC problem. We define the family $\mathcal{F} = \{S_1, \ldots, S_n\}$ of subsets to correspond to the n featurettes present in T. Specifically, the subset S_j corresponding to featurette τ_j that is present in T is composed of elements $i \in U$ corresponding to trees T_i where τ_j is absent. The key idea is that S_j is indicating in which input trees featurette τ_j of T is absent. We note that $\mathcal{F}$ is a multi-set as distinct featurettes τ_j and $\tau_{j'}$ may be absent in the same set of trees, thus leading to $S_j = S_{j'}$ (Fig. 3(b)). There is a bijection between set covers of $(U, \mathcal{F})$ and distinguishing features of T. That is, each distinguishing feature $\Pi = \{\tau_1, \ldots, \tau_{|\Pi|}\}$ corresponds to the same-sized cover $\Pi(C)$ composed of subsets $\{S_1, \ldots, S_{|\Pi(C)|}\}$, and vice versa. In particular, a minimal distinguishing feature corresponds to a minimal set cover. Thus, we may use the following integer linear program (ILP) to find a minimum set cover C and thus a corresponding *minimum* distinguishing feature $\Pi(C)$ — a minimal distinguishing feature of minimum size.

$$\min \sum_{j=1}^{n} x_j \text{ s.t. } \sum_{j \in [n]: i \in S_j} x_j \geq 1, \qquad \forall i \in [m]. \tag{1}$$

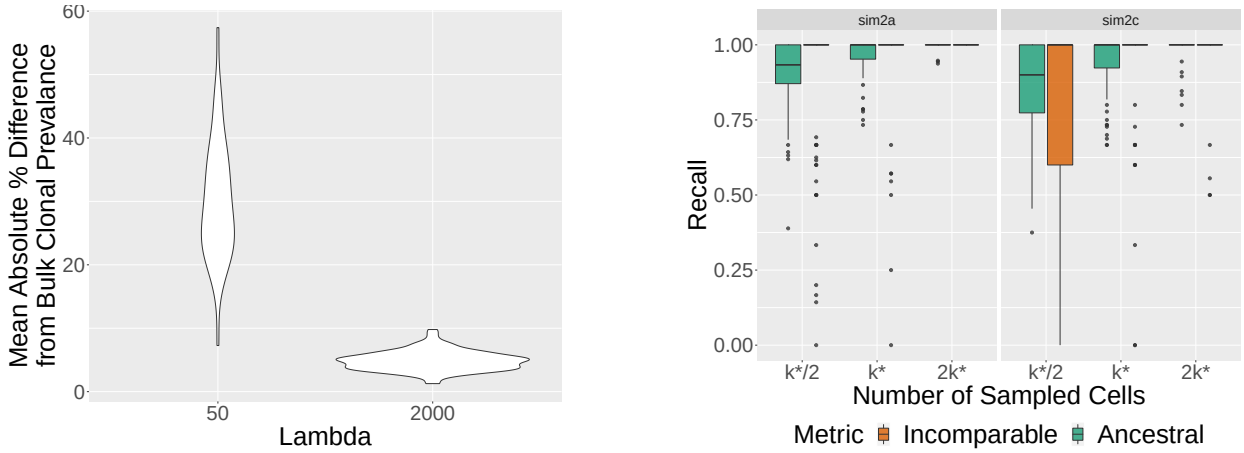
In order to find the *next* minimum distinguishing feature $\Pi(C')$ that is not contained within $\Pi(C)$, we add the following constraint to the ILP.

$$\sum_{j \in C} x_j \leq |C| - 1, \qquad \forall C \in \mathcal{C}^*, \tag{2}$$

where $\mathcal{C}^* = \{C\}$. By repeatedly adding identified minimum set covers to $\mathcal{C}^*$ until the ILP becomes infeasible, we identify all minimal set covers $\mathcal{C}^*$ and thus all minimal distinguishing features Φ^*. Fig. 3(b) shows an example. We use IBM ILOG CPLEX v12.9 to solve the ILP[1].

[1] https://www.ibm.com/analytics/cplex-optimizer

C Supplementary Results



(a) (b)

Fig S2: **PhyDOSE sensitivity analysis of clonal prevalence distortion** (a) The mean absolute percentage difference of the single cell clonal prevalence per replication from simulated bulk clonal prevalence at values of $\lambda \in \{50, 2000\}$. These values of λ resulted in mean absolute percentage difference of 5% and 20% respectively. (b) $\lambda = 2000$ was selected for sim2a/b but further sensitivity analysis was performed with $\lambda = 50$. The recall metrics are compared between sim2a at $\lambda = 2000$ and sim2c at $\lambda = 50$. (b) Recall metrics when inferring T^* with SPhyR [2] by randomly sampling $k^*/2, k^*, 2k^*$ simulated single cells.

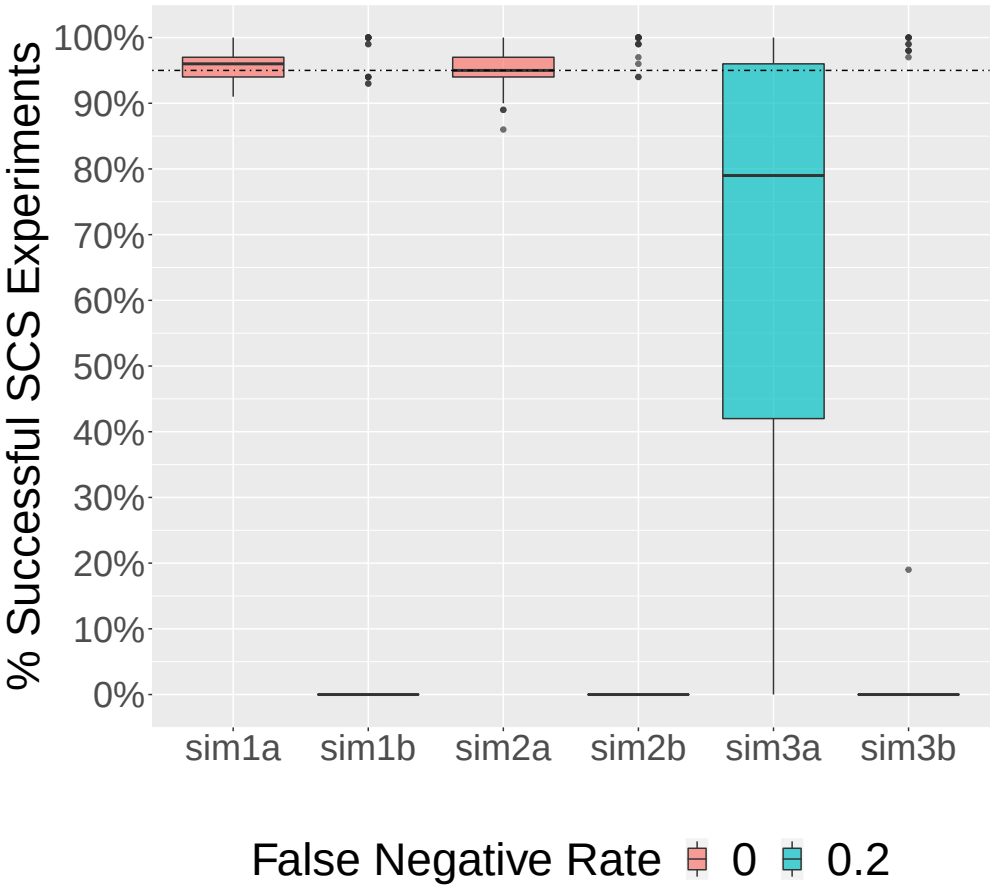


Fig S3: **Percentage of *in silico* single-cell sequencing experiments correctly identifying T^* as the true tree utilizing PhyDOSE's k^* at γ =0.95 and** $\beta \in \{0.0, 0.2\}$. PhyDOSE has a high success rate in alignment with $\gamma = 0.95$ (dashed horizontal line) when utilizing the entire set $\mathcal{T}$ in condition a. However, the introduction of false negatives results in a reduction in success rate and greater variance. In the cases where $T^* \in \mathcal{T}$, the support metric was still able to successfully identify T^* as the true tree.

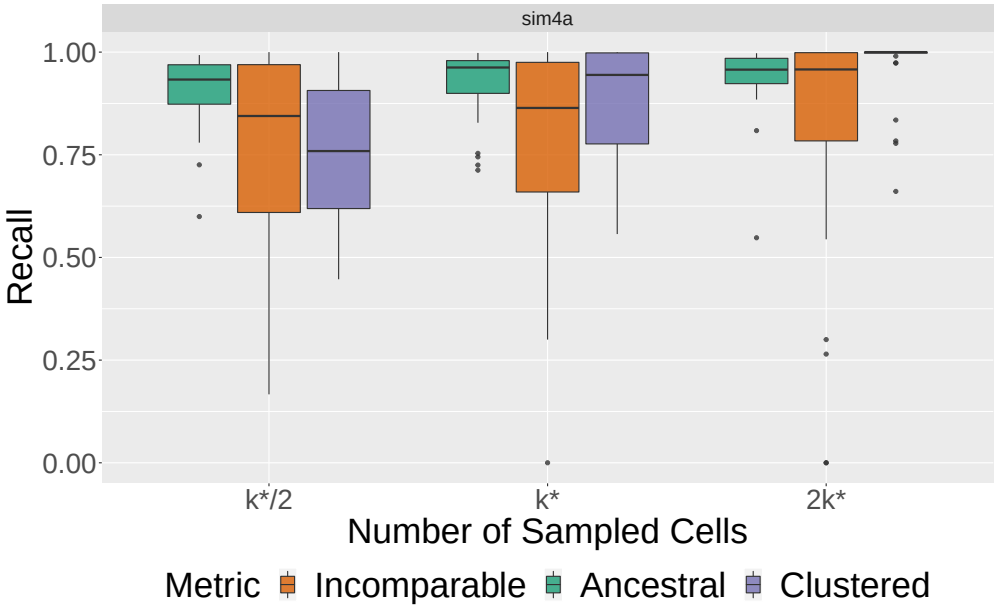
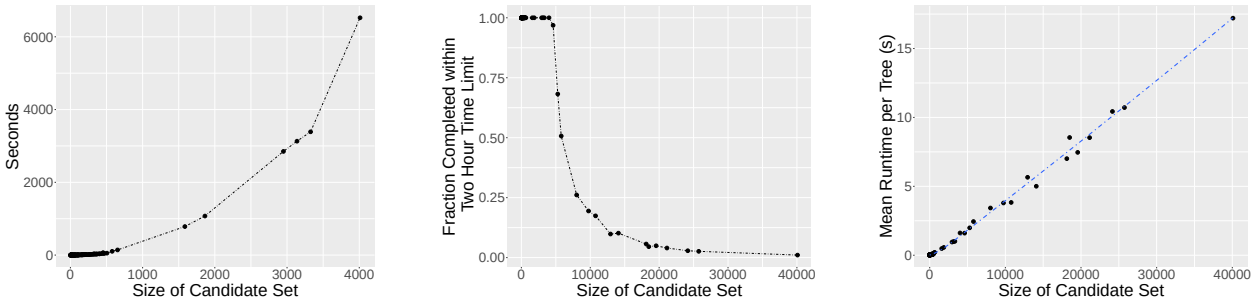


Fig S4: **PhyDOSE performance with mutation clustering** Recall metrics when inferring T^* with SPhyR [2] by randomly sampling $k^*/2, k^*, 2k^*$ simulated single cells.



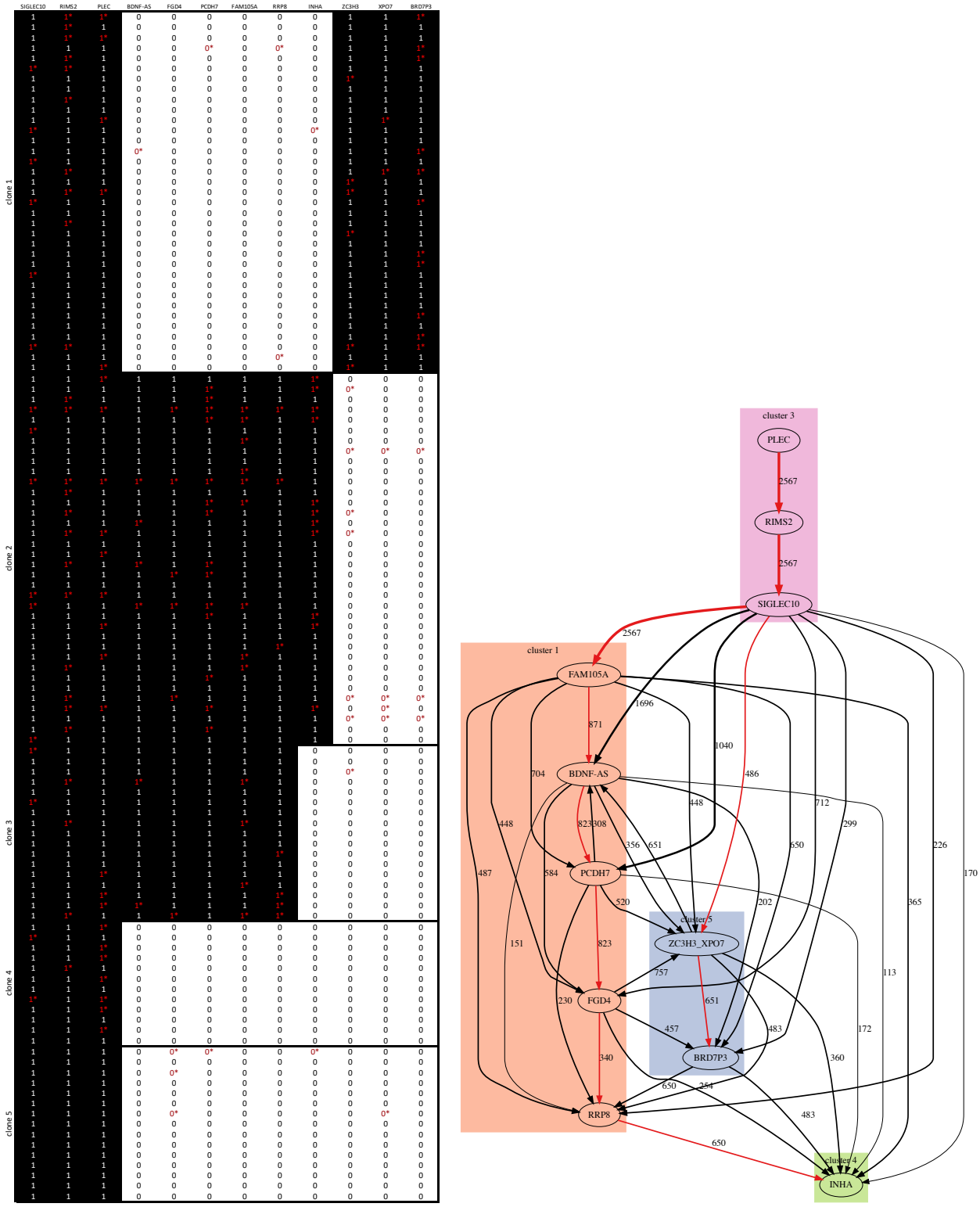
(a) (b) (c)

Fig S5: **Empirical runtime analysis on the bottleneck step of PhyDOSE** (a) Total runtime in seconds for each input set $\mathcal{T}$ that completed in the two hour limit when solved sequentially. (b) Fraction of the input set $\mathcal{T}$ completed within the two hour limit when solved sequentially. (c) Mean runtime in seconds per tree $T \in \mathcal{T}$.

C.1 Retrospective Analysis of an Acute Lymphoblastic Leukemia Patient

For each patient, the sequenced cells are clustered into 2 to 7 clones using an EM-based approach [3]. Based on the fact that false negatives occur more frequently than false positives, we designated an SNV as present if at least 30% of cells in the clone had the mutation. We then checked if the resulting binary, clone-by-SNV matrices adhered to the infinite sites assumption, which was the case for only patients 2 and 3. While the VAFs of all 16 SNVs in patient 2 are less than 0.5, patient 3 had 6 out 49 SNVs with a VAF larger than 0.5, which is indicative of copy number aberrations. Since no copy number information was available to infer cancer cell fractions, we excluded patient 3 from our analysis, thus restricting our attention to patient 2.

For patient 2, 115 cells were clustered into 5 clones [3]. This patient has 16 SNVs, from which we excluded mutations *CMTM8*, *ATRNL1*, *LINC00052* and *TRRAP* for reasons that we described in the main text. The majority voting rule described above yielded a binary clone-by-SNV matrix with 4 mutations clusters that each correspond to SNVs that co-occur in every clone (Fig. S6a), corresponding to a two-state perfect phylogeny T_{SCS} on the mutation clusters (Fig. 5(b)). To obtain the set $\mathcal{T}$ of candidate phylogenies, we considered the bulk data. Specifically, we merged mutations ZC3H3 and XPO7 as they had the same VAF in the bulk data and occurred in the same mutation cluster in the cleaned SCS data (Fig. 5(a)). Using SPRUCE [4], we enumerated $|\mathcal{T}| = 2576$ trees (Fig. S6b). Only one tree $T^* \in \mathcal{T}$ was consistent with T_{SCS}, i.e. each mutation cluster of T_{SCS} formed a connected path in T^* and subsequently collapsing these paths in T^* yields T_{SCS}. Comparing the cleaned single-cell data to the raw values, we computed a false negative rate β of 0.2 for the 14 mutations (Fig. S6a), which was in line with the value reported by [3].



(a) (b)

Fig S6: **Retrospective analysis of ALL patient 2 where [3] sequenced 115 single cells, targeting 16 mutations.** (a) Overview of single-cell data, focusing on a subset of 14 mutations. Each row is a single cell and columns correspond to mutations. Black cells indicate presence of the SNV, whereas white cells indicate absence. Text in each cell indicates a true positive ('1'), false negative ('1*'), true negative ('0') or false positive ('0*') entry. (b) Overview of the set $\mathcal{T}$ of trees, where each edge is labeled by its frequency in $\mathcal{T}$. Red edges encode the tree T^* that is consistent with SCS data in (a). In particular, the SCS data encodes a tree composed of four mutation clusters, each indicated by a different shading.

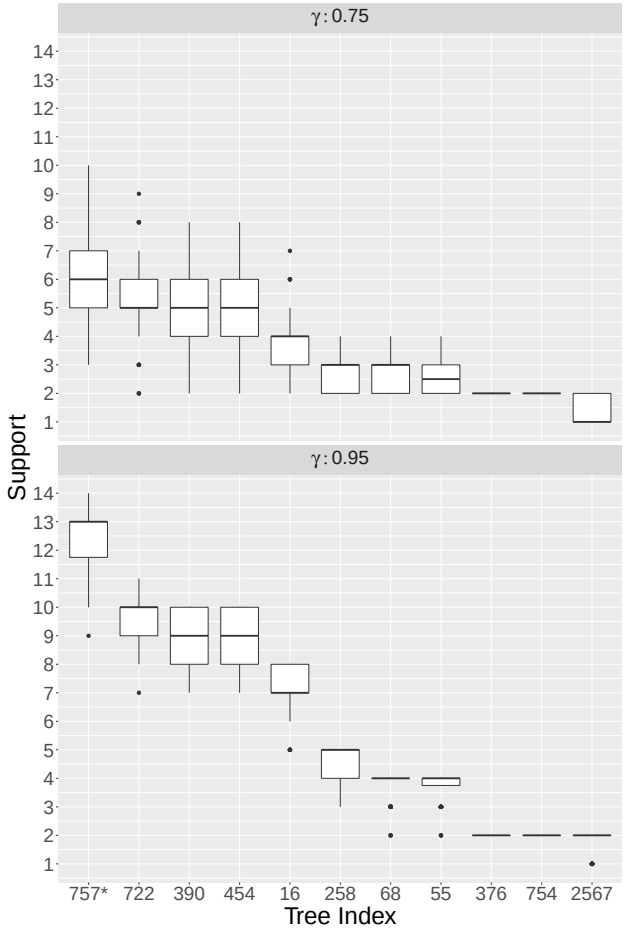
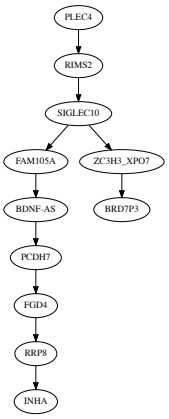
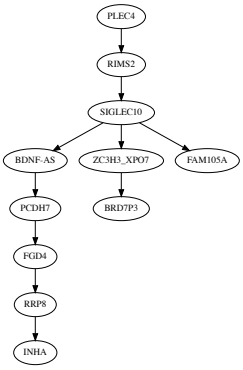


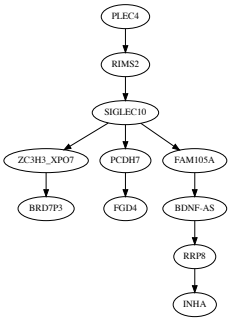
Fig S7: **Distribution of number of supporting cells for each candidate tree across 100 *in silico* SCS experiments at varying success probabilities for ALL patient 2 [3].** For $\gamma = 0.75$, we sampled $k(T^*) = 50$ cells from the SCS data. For $\gamma = 0.95$, we sampled $k(T^*) = 103$ cells. Fig. S8 shows the candidate trees.



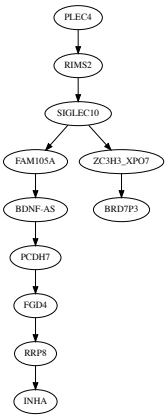
(a) 757 (T^*)



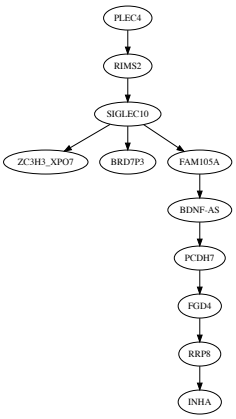
(b) 722



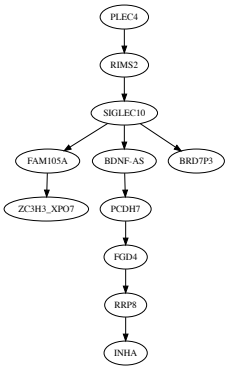
(c) 390



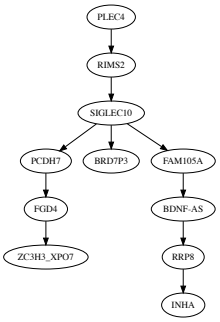
(d) 454



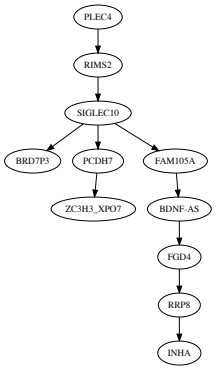
(e) 16



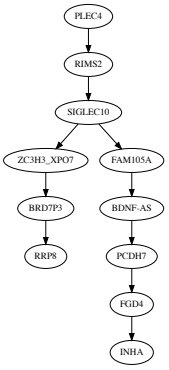
(f) 258



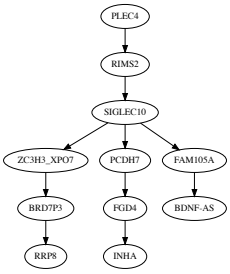
(g) 68



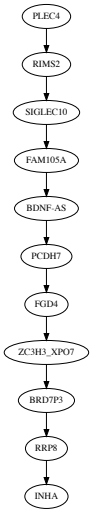
(h) 55



(i) 754



(j) 376



(k) 2567

Fig S8: **Candidate trees that had a non-zero support across 100 *in silico* SCS experiments for ALL patient 2 [3].** For $\gamma = 0.75$, we sampled $k(T^*) = 50$ cells from the SCS data. For $\gamma = 0.95$, we sampled $k(T^*) = 103$ cells. Labels match tree indices in Fig. S7.

patient	mutations	total sequenced [5]	$\lvert\mathcal{T}\rvert$	$k^*_{\gamma=0.75}$ (% red.)	$k^*_{\gamma=0.95}$ (% red.)
AML-4-001	5	5878	8	451 (92.3%)	973 (83.4%)
AML-5-001	4	10191	8	396 (96.1%)	730 (92.8%)
AML-14-001	3	7053	4	129 (98.2%)	278 (96.1%)
AML-22-001	6	2925	57	209 (92.9%)	402 (86.3%)
AML-26-001	5	9601	24	306 (96.8%)	590 (93.9%)
AML-29-001	3	7795	2	78 (99.0%)	169 (97.8%)
AML-30-001	3	6668	2	83 (98.6%)	178 (97.3%)
AML-32-001	4	8241	2	7 (99.9%)	14 (99.8%)
AML-36-001	3	8153	2	246 (97.0%)	531 (93.5%)
AML-38-001	8	7235	316	366 (95.0%)	692 (90.4%)
AML-39-001	4	6735	6	56 (99.2%)	120 (98.2%)
AML-40-001	3	6194	2	60 (99.0%)	128 (98.0%)
AML-41-001	7	5082	89	413 (91.9%)	1155 (83.0%)
AML-42-001	3	2252	2	125 (94.4%)	270 (88.0%)
AML-45-001	3	7716	2	29 (99.7%)	62 (99.2%)
AML-51-001	3	8219	2	216 (97.3%)	467 (94.3%)
AML-59-001	3	2662	2	41 (98.5%)	88 (96.7%)
AML-61-001	4	4320	2	64 (98.5%)	138 (96.8%)
AML-62-001	3	4207	2	9 (98.5%)	20 (99.5%)
AML-64-001	4	4999	3	225 (95.5%)	471 (95.5%)
AML-67-001	5	6024	2	21 (99.7%)	44 (99.2%)
AML-69-001	3	7462	2	202 (97.3%)	436 (94.2%)
AML-75-001	3	6073	2	20 (99.7%)	42 (99.3%)
AML-76-001	4	8033	3	196(97.5%)	423 (94.7%)

Table S1: **Prospective analysis of an acute myeloid leukemia (AML) cohort.** Table shows the patient identifier, the number of mutations,the total cells sequenced [5], the size of the candidate set $\mathcal{T}$ of trees as determined by SPRUCE [4], PhyDOSE's k^* calculated at confidence levels of $\gamma \in \{0.75, 0.95\}$ (% reduction from total sequenced).

patient	clusters	samples	min muts	max muts	$\lvert\mathcal{T}\rvert$	$k^*(\gamma = 0.75)$	$k^*(\gamma = 0.95)$	sel. sample
CRUK0004	7	4	10	78	2	15	32	R3
CRUK0005	6	4	24	536	2	12	26	R3
CRUK0011	8	3	12	335	3	16	34	R2
CRUK0012	5	2	11	84	2	7	15	R1
CRUK0013	9	5	5	114	8	487	1051	LN1
CRUK0022	5	2	5	158	2	11	23	R1
CRUK0023	10	4	5	226	2	10	20	R1
CRUK0025	7	3	10	350	2	12	25	R2
CRUK0028	5	2	5	72	2	15	33	R1
CRUK0031	7	3	15	675	2	14	30	R1
CRUK0037	10	5	6	397	17	∞	∞	N/A
CRUK0038	4	2	6	107	2	20	43	R1
CRUK0046	5	4	5	186	2	14	29	R1
CRUK0049	6	2	7	882	4	17	37	R2
CRUK0063	8	5	5	167	2	16	33	R4
CRUK0067	5	2	24	263	2	14	29	R1
CRUK0068	10	4	11	532	3	∞	∞	N/A
CRUK0070	10	5	6	254	2	11	24	R6
CRUK0076	9	4	8	846	4	21971	47470	R2
CRUK0077	7	4	9	586	2	10	22	R4
CRUK0084	6	4	5	332	2	8	17	R2
CRUK0094	6	4	6	50	2	10	22	R4
CRUK0095	4	3	6	216	2	20	42	R2
CRUK0099	5	4	5	438	2	21	46	R6
CRUK0100	8	3	5	777	3	12	24	R2

Table S2: **Prospective analysis of TRACERx non-small-cell lung cancer cohort.** Table shows the patient identifier, the number of mutation clusters, the number of bulk samples, the minimum number of mutations per cluster, the maximum number of mutations per cluster, the size of the candidate set $\mathcal{T}$ of trees as determined by [6], PhyDOSE's k^* calculated at confidence levels of $\gamma \in \{0.75, 0.95\}$ and the recommended sample label from which the single cells should be drawn.

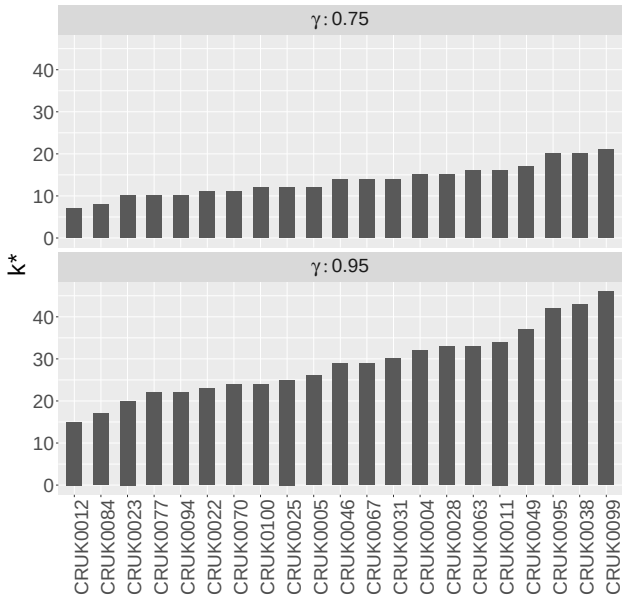


Fig S9: **PhyDOSE calculated k^* for the lung cancer cohort at varying confidence levels.** Patients CRUK0013, CRUK0037, CRUK0068, CRUK0076 were excluded from the plot, but are shown in Table S2.

References

[1] Karp RM. In: Miller RE, Thatcher JW, Bohlinger JD, editors. Reducibility among Combinatorial Problems. Springer; 1972. p. 85–103.

[2] El-Kebir M. SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. Bioinformatics. 2018;34(17):i671–i679.

[3] Gawad C, Koh W, Quake SR. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. Proceedings of the National Academy of Sciences. 2014;111(50):17947–17952. doi:10.1073/pnas.1420822111.

[4] El-Kebir M, Oesper L, Acheson-Field H, Raphael BJ. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. Bioinformatics. 2015;31(12):i62–i70.

[5] Morita K, Wang F, Jahn K, Kuipers J, Yan Y, Matthews J, et al. Clonal Evolution of Acute Myeloid Leukemia Revealed by High-Throughput Single-Cell Genomics. bioRxiv. 2020;doi:10.1101/2020.02.07.925743.

[6] Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TB, Veeriah S, et al. Tracking the evolution of non–small-cell lung cancer. New England Journal of Medicine. 2017;376(22):2109–2121.