Examining Tumor Phylogeny Inference in Noisy Sequencing Data

Kiran Tomlinson
Department of Computer Science
Carleton College
Northfield, MN
tomlinsonk@carleton.edu

Layla Oesper
Department of Computer Science
Carleton College
Northfield, MN
loesper@carleton.edu

***Abstract*—A number of methods have recently been proposed to reconstruct the evolutionary history of a tumor from noisy DNA sequencing data. We investigate when and how well these histories can be reconstructed from multi-sample bulk sequencing data when considering only single nucleotide variants (SNVs). We formalize this as the Enumeration Variant Allele Frequency Factorization Problem and provide a novel proof for an upper bound on the number of possible phylogenies consistent with a given dataset. In addition, we propose and assess two methods for increasing the robustness and performance of an existing graph based phylogenetic inference method. We apply our approaches to noisy simulated data and find that low coverage and high noise make it more difficult to identify phylogenies. We also apply our methods to both chronic lymphocytic leukemia and clear cell renal cell carcinoma datasets.**

***Index Terms*—Cancer genomics, tumor phylogeny, evolution.**

I. Introduction

Cancer is a disease originating from somatic mutations in a single founder cell and characterized by the runaway proliferation of that cell's aberrant descendants. The clonal theory of cancer [1] posits that new somatic mutations will continue to arise in descendants of the founder cell, driving the progression of the disease. As a result, a tumor may be a heterogeneous mix of different tumor cell populations, or clones, each with their own set of somatic mutations. A number of recent studies have highlighted the prevalence of such *intra-tumor heterogeneity* [2], [3]. Characterization of the particular clones appearing in a specific tumor and how they evolved has important implications for understanding and, ultimately, treating the disease [4], [5].

In recent years there has been increased interest in computational methods that use noisy DNA sequencing data to reconstruct the evolutionary history of a tumor in terms of ancestral relationships between somatic mutations. Some recent approaches have focused on using single-cell sequencing data to reconstruct tumor phylogenies [6], [7], [8]. However, despite significant technological advances, single-cell sequencing remains error-prone and, in many cases, prohibitively expensive. Therefore, we focus here on the data from more economical bulk sequencing, in which a collection of potentially heterogeneous cells are sequenced together, obfuscating the relationships between mutations. Furthermore, there are many other sources of error in the data, including the sequencing process, read alignment, and variant calling algorithms. Thus, specialized methods are required to robustly analyze such noisy bulk sequencing data.

A number of recent computational methods have been developed to infer tumor phylogenetic trees using multi-sample bulk sequencing data. Many methods restrict attention to single nucleotide variants (SNVs) [9], [10], [11], [12], [13] and use observations about the observed frequencies of each mutation to identify possible ancestral relationships. Specifically, these methods make this problem more tractable by using the infinite sites assumption (ISA) which states that any locus in the genome mutates at most once during the history of tumor. For example, AncesTree [9] creates a graph called the ancestry graph from mutation frequencies and then identifies spanning trees of that graph adhering to the ISA. Reports that the ISA is often violated in tumors [14] have led to the investigation of the removal of the ISA in limited contexts [8], [15]. A few other methods also consider structural variants or copy number aberrations [16], [17], [18], [19] in addition to SNVs, but this has proven challenging. Finally, a few methods make note that multiple tumor evolutionary trees may be consistent with a given sequencing dataset and attempt to enumerate these trees [10], [18], [19]. In this vein, a recent paper [20] observed that multiple such trees typically exist in noise free simulations. However, it is unclear how the conclusions from that work are affected by the multiple sources of noise present in bulk sequencing data and to what extent they transfer to real sequencing data.

In this paper, we investigate when and how well clonal evolutionary trees can be reconstructed from multi-sample bulk sequencing data using the ancestry graph approach of [9], which relies on the ISA. In particular, we focus on the performance of this method when applied to noisy data. We describe a relaxation of the ancestry graph approach that makes it more robust to noise and introduce a method of simplifying the ancestry graph to reduce computational cost. We show the effects of coverage, noise, and other parameters in reconstructing clonal trees in noisy simulated data and apply our methods to cancer sequencing datasets from two studies [21], [22].

This project is supported by NSF CRII award IIS-1657380 and by Elledge, Eugster, and Class of '49 Fellowships from Carleton College.

978-1-5386-5488-0/18/$31.00 ©2018 IEEE

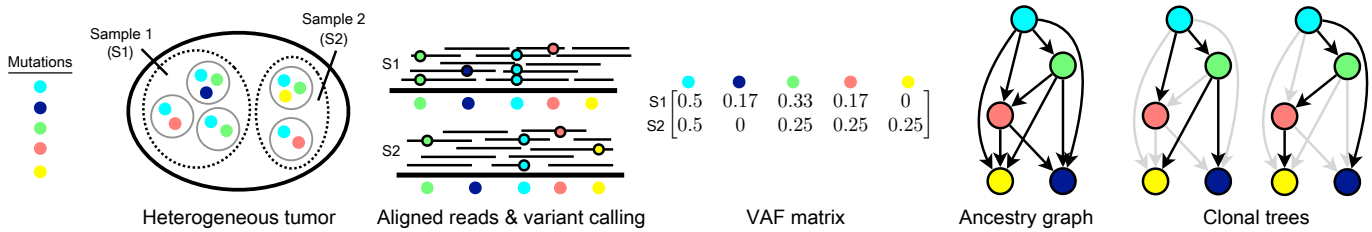


Fig. 1. Overview of the clonal tree inference process. From left to right: several samples are taken from a heterogeneous tumor, either from different parts of the tumor or from different time points; the samples are sequenced, the resulting reads are aligned to a reference genome, and variants are identified in the aligned reads of each sample; the variant and reference read counts are used to build the VAF matrix; we construct the ancestry graph of the VAF matrix; each spanning tree of the ancestry graph adhering to the sum condition is a possible clonal tree, two of which are displayed. Note that the second tree shown would be eliminated if we could determine the coincidence of mutations, since the dark blue mutation always appears with the green mutation in the tumor.

II. Methods

We first describe the ancestry graph method [9] and then formalize the problem of using this approach to enumerate all tumor phylogenies consistent with a given dataset. We also provide a novel proof of an upper bound on the number of possible trees consistent with this approach, present a relaxation that allows the method to be more robust to noise, and introduce a simplification that improves computational efficiency.

A. Problem Formalization

1) Definitions: Let s be the number of samples sequenced from a tumor and let n be the total number of observed mutations across all samples. Mutations are labeled $1, \ldots, n$. The $s \times n$ *variant allele frequency (VAF) matrix* F stores in F_{ij} the fraction of reads from sample i that contain mutation j. A *clonal tree* T (or tumor phylogeny) is a rooted tree on n nodes with each node labeled by a distinct mutation. More generally, nodes may be labeled with disjoint sets of mutations, with a corresponding decrease in the number of nodes. Each node represents a cell population containing all mutations along its root-node path. The infinite sites assumption guarantees that a clonal tree is a perfect phylogeny, so we can also store the tree as an $n \times n$ *clonal matrix* B, where $B_{\ell j} = 1$ if cell population ℓ contains mutation j and 0 otherwise. Finally, samples from the tumor contain mixtures of cell populations: the $s \times n$ *usage matrix* U stores in $U_{i\ell}$ the frequency of cells from population ℓ in sample i.

2) The VAFFP and the Ancestry Graph: The authors of [9] formalized the Variant Allele Frequency Factorization Problem (VAFFP), also called the Perfect Phylogeny Mixture Problem in [20], as follows:

Given: A VAF matrix F.

Find: A usage matrix U and a clonal matrix B such that:

$$F = \frac{1}{2} UB \tag{1}$$

The factor of $1/2$ arises because all mutations are assumed to be heterozygous SNVs. The VAFFP is known to be NP-complete [9], but in practice, many datasets are small enough that finding solutions is feasible.

In [9], the authors present an approach for solving the VAFFP using the *ancestry graph* of F (see Fig. 1 for a visual overview of this method). When necessary to avoid confusion, we will refer to the ancestry graph as the *strict* ancestry graph. The ancestry graph G_F contains n nodes, one labeled by each mutation. Additionally, G_F includes a directed edge from node j to node k if $F_{ij} \geq F_{ik}$ $\forall i \in \{1, \ldots, s\}$. These edges encode the *ancestry condition*: under the infinite sites assumption, an ancestral mutation must be more frequent than a descendant mutation. The possible clonal trees are exactly the set of directed spanning trees of G_F that adhere to the *sum condition* (2). Using $C(j)$ to denote the children of mutation j in a clonal tree T, the sum condition requires that:

$$\sum_{k \in C(j)} F_{ik} \leq F_{ij} \qquad \forall i \in \{1, \ldots, s\} \tag{2}$$

That is, the sum of observed frequencies of children mutations in a clonal tree cannot exceed the frequency of their parent mutation in any sample.

Each spanning tree T of G_F adhering to the sum condition yields a solution to VAFFP (see the rightmost part of Fig. 1 for examples). We can construct the clonal matrix B from T by tracing node labels along each root-node path. We can then efficiently compute U without back-substitution [9]:

$$U_{ij} = 2\Big(F_{ij} - \sum_{k \in C(j)} F_{ik}\Big) \tag{3}$$

3) The Enumeration Variant Allele Frequency Factorization Problem (E-VAFFP): Here, we define the focus of our work, the enumeration version of the VAFFP.

Given: A VAF matrix F.

Find: The set $\mathcal{T}(G_F)$ of *all* trees that span the ancestry graph G_F and adhere to the sum condition.

When $\mathcal{T}(G_F) \neq \emptyset$, we say that an E-VAFFP solution exists or that F admits an E-VAFFP solution. In this paper, we explore the relationship between $\mathcal{T}(G_F)$ and the underlying tumor evolutionary tree and investigate relaxations and extensions to the E-VAFFP.

B. Finding and Counting E-VAFFP Solutions

To solve the E-VAFFP, we use the same method employed in [11], [18], [20], a modified version of the Gabow-Myers algorithm [23]. This algorithm recursively constructs all spanning trees of a graph through a structured depth-first search.

It is straightforward to modify the Gabow-Myers algorithm to avoid execution branches violating the sum condition.

We can place an upper bound on $|\mathcal{T}(G_F)|$ by counting the number of spanning trees of G_F. Tutte's Matrix-Tree Theorem [24] provides a polynomial-time method of counting the spanning trees of a directed graph from the graph's Laplacian matrix. The Laplacian matrix of a graph is obtained by subtracting its adjacency matrix from its in-degree matrix.

Theorem 1 (Tutte's Matrix-Tree Theorem): The number of spanning trees of a directed graph G rooted at r is $\det(\hat{L}_r)$, where $\hat{L}_r$ is obtained by removing the rth row and column from the Laplacian matrix of G.

In some cases, it is possible to avoid calculating the determinant: Pradhan and El-Kebir [20] proved that if G is acyclic with a unique root r, then the number of spanning trees of G rooted at r is the product of in-degrees of all nodes $v \neq r$ in G. Their proof uses a bijection between spanning trees and a set of functions of known size. We rephrase Pradhan and El-Kebir's theorem in a way that makes more direct use of Theorem 1 and invites a novel proof.

Theorem 2: If G is a directed acyclic graph, then the number of spanning trees of G rooted at r is the product of the diagonal elements of $\hat{L}_r$.

Proof: Using Theorem 1, it suffices to show that $\det(\hat{L}_r)$ is the product of its diagonal when G is acyclic. We proceed by induction on the size of $\hat{L}_r$. If $\hat{L}_r$ is 1×1, then its determinant is indeed the lone diagonal element. Now suppose $\hat{L}_r$ is $n\times n$. Since G is acyclic, there must exist a node in $G-\{r\}$ with no incoming edges. Equivalently, $\hat{L}_r$ must contain a column i with all off-diagonal elements zero. By performing a cofactor expansion along column i, we see that $\det(\hat{L}_r)$ is the product of its ith diagonal element with the determinant of the $(n-1)\times(n-1)$ minor resulting from the removal of column i and row i. Note that this minor still has the same vital property as $\hat{L}_r$: it stores the adjacency of an acyclic forest in its off-diagonal elements. We could thus perform the same cofactor expansion on the minor, allowing us to conclude that the determinant of the minor is the product of its diagonal by the inductive hypothesis. Therefore the determinant of $\hat{L}_r$ is also the product of its diagonal. ■

Since the diagonal product of $\hat{L}_r$ is precisely the product of in-degrees of all nodes $v \neq r$ in G, Theorem 2 is equivalent to Pradhan and El-Kebir's.

We can guarantee that G_F is acyclic by merging any mutations with equal frequencies. Then, all spanning trees of G_F must share the same root r. This allows us to place an upper bound on $|\mathcal{T}(G_F)|$ using Theorem 2: it cannot exceed the diagonal product of $\hat{L}_r$. While this bound is computable in polynomial time, determining the exact size of this set is NP-hard and is conjectured to be #P-complete [20].

C. Pruning Transitive Edges

Theorem 2 allows us to see that the number of spanning trees of an n-node DAG grows exponentially with n when the average in-degree is held constant. Even with only 20 mutations, the number of spanning trees of G_F can exceed 10^{17}, dramatically slowing down the identification of clonal trees. In order to reduce the number of spanning trees while preserving core ancestral relationships, we explore removing transitive edges from the ancestry graph (see Fig. 2). This approach may be especially useful if we expect that the tumor has a branching, rather than a linear topology.

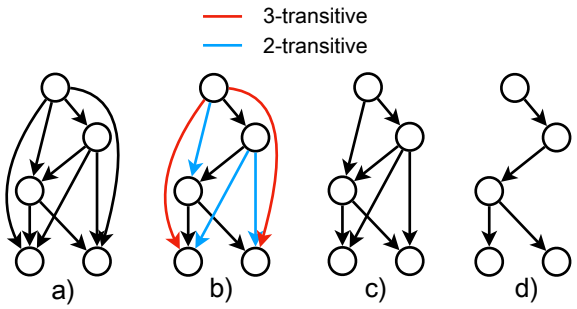


Fig. 2. Example of partial transitive reduction. a) An ancestry graph G_F. b) The transitive edges in G_F. The red edges are 3-transitive and the blue edges are 2-transitive. c) The 3-PTR of G_F. d) The transitive reduction of G_F; equivalently, the 2-PTR of G_F.

For a directed acyclic graph G, we define an edge $(u,v) \in G$ to be *k-transitive* if there exists a path from u to v of length k (see Fig. 2b). We say that an edge is $\geq k$-transitive if it is i-transitive for some $i \geq k$. By pruning G of $\geq k$-transitive edges for a chosen k, we can reduce the number of spanning trees while maintaining the general structure of G. We call the graph resulting from removing all $\geq k$-transitive edges the *k partial transitive reduction (k-PTR)* of G. Note that the 2-PTR is the standard transitive reduction [25] of a graph (see Fig. 2d). To construct the k-PTR of G, we first find the transitive reduction R of G using Hsu's algorithm [26]. Then, we can easily identify if (u,v) is $\geq k$-transitive by checking the path length from u to v in R. This may be done efficiently by pre-computing the all-pairs shortest path matrix of R with n breadth-first searches.

D. Relaxations of the E-VAFFP

1) Approximate Ancestry Graph: Real sequencing data is rife with noise, but the E-VAFFP assumes F is measured precisely. In real datasets, there are often no spanning trees of G_F. To handle less idealized data, we use a method based on the probabilistic approach from [9]. This approach defines the *approximate ancestry graph* of F: a complete n-node directed graph with nodes labeled by mutations and edges (j,k) weighted by the probability that mutation j is ancestral to mutation k given their observed frequencies. To calculate this probability, we model the VAF of mutation j in sample i with the beta-distributed random variable X_{ij}, as in [9]. If $X_{ij} \geq X_{ik}$, then this provides evidence that mutation j is ancestral to mutation k. The overall probability that j is ancestral to k is defined based on the sample with the weakest evidence:

$$\Pr[j \text{ ancestral to } k] := \min_i \Pr[X_{ij} \geq X_{ik}] \qquad (4)$$

The probabilities on the right hand side of (4) can be directly calculated from the read counts that generate F [27].

Just as with the strict ancestry graph, we can use the Gabow-Myers algorithm [23] to enumerate all spanning trees of the approximate ancestry graph whose observed frequencies satisfy the sum condition. Once these are found, the most probable (i.e. max weight) tree can be selected. Alternatively, if there are too many spanning trees to enumerate exhaustively, we can use an algorithm of Camerini, Fratta, and Maffioli [28] to enumerate weighted spanning trees in descending weight order until one satisfying the sum condition is found. Unlike Gabow-Myers, this algorithm does not offer a simple sum condition modification. Using this method, we can potentially find the most probable clonal tree without the need to enumerate every tree. The disadvantage of the algorithm from [28] is that it is much slower when no solutions exist, since it is forced to explore the entire space of spanning trees rather than just those satisfying the sum condition.

Note that the approximate ancestry graph does not admit more E-VAFFP solutions than the strict ancestry graph. This is because any tree violating the sum condition in the strict graph will necessarily violate it in the approximate graph as the sum condition only relies on F. Additionally, any spanning tree of the approximate graph that does not exist in the strict graph must violate the ancestry condition (and therefore violates the sum condition), since it includes an edge not present in the strict graph.

The key benefits of the approximate ancestry graph are that it provides an ordering on solutions and that it allows the exploration of novel tree topologies not present in the strict graph. To make use of these topologies, however, we need to weaken the sum condition.

2) Relaxed Sum Condition: Adding leniency to the sum condition allows the identification of possible clonal trees obscured by noise. For a small threshold ε, we can relax the sum condition to require that:

$$\sum_{k \in C(j)} F_{ik} \leq F_{ij} + \varepsilon \qquad \forall i \in \{1, \ldots, s\} \tag{5}$$

We then can identify the smallest ε that allows one valid spanning tree to be found in the approximate ancestry graph. This is equivalent to finding the spanning tree whose maximal violation of the sum condition is minimal.

III. Results

We investigated E-VAFFP solutions in simulated noisy data and comparatively assessed the strict and approximate ancestry graph approaches on two real datasets of 3 chronic lymphocytic leukemia (CLL) patients from [21] and 8 clear cell renal cell carcinoma (ccRCC) patients from [22]. In particular, we examined the effect of noise on the existence of E-VAFFP solutions and on the degree to which trees in $\mathcal{T}(G_F)$ reflect the underlying evolutionary tree. In addition to the findings presented here, we also simulated error-free data and reproduced the relationships between n, s, and the size of $\mathcal{T}(G_F)$ reported in [20].

A. Simulated Data

On simulated data, we present findings on the existence and quality of strict and approximate E-VAFFP solutions in noisy DNA sequencing data. We also separately evaluate the usefulness of pruning transitive edges from the ancestry graph. We first describe our data simulation procedure.

1) Simulating Noisy VAF Data: Our data simulation process consists of four steps: (1) randomly generating an evolutionary tree topology, (2) choosing the cellular frequencies, (3) determining the mutation frequencies, and (4) drawing variant reads from a binomial distribution, allowing direct computation of F. We also describe our method of varying noise levels in the simulated VAF matrix.

Given a number of mutations n, a number of samples s, and an average sequencing coverage c, we generate a random tumor phylogenetic tree T and an $s \times n$ VAF matrix consistent with T. For simplicity, we say that each clone contains a single new mutation not shared by its parent, so we interchangeably refer to n as the number of clones. Making no assumptions about the topology of tumor phylogenies, T is constructed iteratively by adding each mutation as the child of a random node already in T. From T, we can construct the clonal matrix B as in Section II-A2. We then choose the frequency of each of the n clones in the simulated tumor. Clone i is assigned frequency u_i such that $\sum_i u_i = 1$. To choose $u_1, \ldots, u_n$, we sample uniformly from the *standard simplex*, the set of points in $\mathbb{R}^n$ whose coordinates are non-negative and have sum 1, using a method described in [29].

We then calculate the frequencies of the n mutations in the tumor. Storing the mutation and cellular frequencies in the row vectors $\vec{f}$ and $\vec{u}$, respectively, we find $\vec{f}$ using (1):

$$\vec{f} = \frac{1}{2}\vec{u}B \tag{6}$$

The last step is to simulate reads in each of the s samples. For simplicity, we model a thoroughly mixed tumor, in which the expected cellular composition of every sample matches that of the tumor as a whole. For each sample i and for each mutation j, we simulate $r_{ij} \sim Poisson(c)$ reads, where c is the mean coverage. The number of variant reads v_{ij} of mutation j in sample i is drawn from a binomial distribution: $v_{ij} \sim Binom(r_{ij}, f_j)$. The $s \times n$ VAF matrix F then contains entries $F_{ij} = v_{ij}/r_{ij}$.

We simulate additional noise in the sampling and sequencing process by adding overdispersion to the binomial distribution. We replace f_j with a beta-distributed random variable with mean f_j. The parameters α and β of the beta distribution are chosen to be:

$$\alpha = \frac{(1-\rho)}{\rho} f_j \qquad \beta = \frac{(1-\rho)}{\rho}(1 - f_j)$$

where $\rho \in (0, 1)$ is the overdispersion parameter. This results in a beta distribution with mean f_j and with variance proportional to ρ. By varying ρ, we can simulate sequencing data with more or less noise. When we do not add overdispersion, we denote it with the shorthand $\rho = 0$.

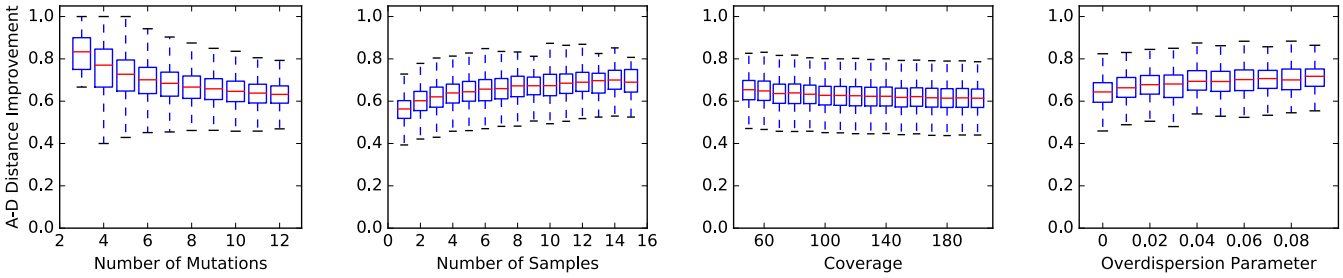


Fig. 3. Parameter effects on E-VAFFP solution quality. An A-D improvement of 0 signifies that trees in $\mathcal{T}(G_F)$ are no better than random, while an improvement close to 1 signifies that $\mathcal{T}(G_F)$ are nearly identical to the underlying evolutionary tree. Note that solution quality is measured only when solutions exist, which may be rare.

2) Existence of E-VAFFP Solutions: Under this data simulation process, we find that in the majority of cases, no spanning trees of G_F satisfy the sum condition. The rarity of E-VAFFP solutions is exacerbated by having many clones, many samples, low coverage, and high noise. We ran 10000 trials of the data simulation and ancestry graph procedure for each parameter setting (n from 3 to 12, s from 1 to 15, coverage from $50\times$ to $200\times$, and ρ from 0 to 0.09) and measured the fraction of trials with $|\mathcal{T}(G_F)| \geq 1$. We refer to these as *solvable* trials. Each parameter was tested independently, using the default values $n = 10$, $s = 5$, $60\times$ coverage, and $\rho = 0$ for parameters held constant.

At the default parameter settings, only 14% of trials had any valid clonal trees. Higher coverage significantly increased the fraction of solvable trials, up to 47% at $200\times$ coverage. Meanwhile, more mutations and samples both significantly decreased the proportion of solvable trials, as did adding overdispersion. At $\rho = 0.09$, E-VAFFP solutions existed in only 89 of the 10000 trials. A high number of samples exhibited a similarly strong effect, with just 103 solvable trials at $s = 15$. We note that E-VAFFP solutions were also rare in real datasets (see section III-B).

3) Existence of Approximate Solutions: Relaxing the sum condition (5) dramatically increases the fraction of solvable trials, even with small ε. There is a linear increase in the proportion of solvable trials from 14% at $\varepsilon = 0$ to 64% at $\varepsilon = 0.05$. This is accompanied by a dramatic increase in the mean size of $\mathcal{T}(G_F)$, from 2000 to 69000. There is therefore a trade off between lower computational effort and an increased likelihood of finding a possible clonal tree.

4) E-VAFFP Solution Quality: To measure the quality of clonal trees generated by the ancestry graph approach, we calculate the mean ancestor-descendant (A-D) distance [30] between each tree in $\mathcal{T}(G_F)$ and the underlying evolutionary tree. Note that standard phylogenetic distance measures, such as Robinson-Foulds [31], do not apply to clonal trees as their internal nodes are labeled in addition to their leaves. To quantify the useful information gained from E-VAFFP solutions, we measure how much more similar trees in $\mathcal{T}(G_F)$ are to the underlying tree than an equinumerous set of randomly generated trees. Formally, with $\overline{AD}(S)$ denoting mean A-D distance between trees in the set S and the underlying tree, we define the *A-D improvement* to be:

$$\frac{\overline{AD}(\text{random}) - \overline{AD}(\mathcal{T}(G_F))}{\overline{AD}(\text{random})} \tag{7}$$

This measure quantifies the decrease in incorrectly identified ancestral relationships relative to the random baseline. For example, an A-D improvement of 0 would indicate that trees in $\mathcal{T}(G_F)$ are no better than random, while an A-D improvement of 1 would mean that $\mathcal{T}(G_F)$ contains only the correct tree.

With default parameters, trees in $\mathcal{T}(G_F)$ exhibited a mean A-D improvement of 0.64, showing that they accurately capture 64% of ancestral patterns in the data missed by the random baseline. Increasing the number of mutations not only makes solutions rarer, but also decreases the quality of solutions when they are present. More samples, on the other hand, is a marked benefit to the similarity of trees in $\mathcal{T}(G_F)$ to the underlying tree. (See Fig. 3.) Our results regarding n and s agree with those presented in [20] on error-free simulated data.

Conditioned on the existence of solutions, we find that higher noise makes trees in $\mathcal{T}(G_F)$ more closely capture ancestral relationships in the underlying tree (see Fig. 3). Higher coverage has a slight negative impact on solution quality as measured by A-D distance. At $50\times$ coverage, the A-D improvement was 0.65 and it decreased to 0.61 at $200\times$ coverage. Meanwhile, higher values of the overdispersion parameter ρ also led to higher-quality trees, in the rare case that any could be found at all. With no overdispersion, the A-D improvement was 0.64 and it reached 0.72 at $\rho = 0.09$. This could indicate that good E-VAFFP solutions are more robust to noise than solutions dissimilar to the underlying tree. Thus, when more noise is present, poor trees are preferentially excluded from $\mathcal{T}(G_F)$, causing the mean A-D improvement to increase. However, these parameters have such strong negative impacts on the existence of solutions that the presence of noise is still deleterious to our ability to infer phylogeny. For instance, the total number of correctly reported ancestral relationships across all 10000 trials does decrease as ρ increases, since there are so few trials with $|\mathcal{T}(G_F)| \geq 1$ when ρ is high.

5) Approximate Solution Quality: Solution quality responds in the same way to changes in s, coverage, and overdispersion

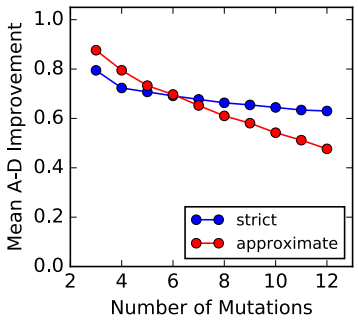
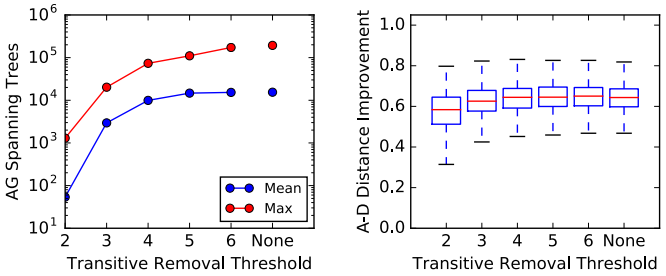


Fig. 4. Difference in the relationship between n and A-D improvement with the strict and approximate ancestry graph methods (with $\varepsilon = 0$). As the number of mutations increases, both methods worsen, but the approximate ancestry graph does so more rapidly.



(a) (b)

Fig. 5. Effect of partial transitive reduction on (a) the number of ancestry graph spanning trees and (b) the quality of clonal trees derived from PTRs of the ancestry graph. 'None' represents the unpruned ancestry graph.

in the approximate ancestry graph as in the strict ancestry graph. However, we found an intriguing difference in the response to number of mutations n. Choosing the max-weight sum-constrained spanning tree of the approximate graph provides noticeably better solutions than the strict approach for small n. However, the approximate method drops off more sharply in quality as n grows, with the crossover point at $n = 6$ (see Fig. 4). We suspect this is due to inherent bias in high-weight approximate spanning trees, since they become worse than randomly sampled strict spanning trees as n grows. This bias may arise because edges in the approximate graph are weighted by the probability that one mutation is ancestral to another, but that edges in fact represent parental rather than ancestral relationships. As such, the root node is likely to have high-weight edges to every other node, even though its probability of being their direct parent may not be as high. We also found that relaxing the sum condition caused a gradual linear decrease in solution quality, from an A-D improvement of 0.54 at $\varepsilon = 0$ to 0.51 at $\varepsilon = 0.05$ when the number of mutations is $n = 10$.

6) Transitive Edge Pruning: We found partial transitive reduction to be a viable method of reducing the number of spanning trees in the ancestry graph without significantly affecting the quality of solutions. We compared the existence and quality of E-VAFFP solution trees resulting from partial transitive reductions of the ancestry graph with those from the non-reduced graph. Using Theorem 2, we also counted the mean and maximum number of spanning trees of the reduced ancestry graph across 10000 trials to quantify the benefit of PTR (see Fig. 5). For this analysis, we used the same default parameters as before.

The total transitive reduction proved to be too extreme. It drastically reduced the mean number of spanning trees, but also reduced the probability of solution existence down to 3%. Additionally, the trees identified from the total transitive reduction were noticeably less similar to the underlying tree, with a mean A-D improvement over random of 0.57 compared to 0.64 for the unpruned ancestry graph (Fig. 5b).

At the other extreme, we found that the 6- and higher PTR had a negligible impact on all measures, reflecting the rarity of $\geq$ 6-transitive edges in 10-node ancestry graphs. The 4-PTR and 5-PTR had no discernible impact on the the fraction of solvable trials, but decreased the maximum number of spanning trees by 43% and 62%, respectively. Meanwhile, the impact on the mean A-D improvement was less than 0.01 for both. Removing 3-transitive edges had a stronger effect on each of these measures. The max and mean number of spanning trees shrank by factors of 9.6 and 7.7 compared to the unpruned graph, while the fraction of solvable trials dropped by two percentage points. In the 3-PTR, the mean A-D improvement decreased slightly from 0.64 to 0.62.

In summary, the 3-, 4-, and 5-PTRs of 10-node ancestry graphs simplify the ancestry graph with minimal impact on solution existence and quality. An ancestry graph with fewer spanning trees results in faster runtime, smaller memory footprint, and allows datasets with more mutations to be analyzed. Selecting between these partial transitive reductions allows for a trade-off between a simpler ancestry graph and higher solution quality. In an ancestry graph with more or fewer nodes, a different value of k would have to be chosen for the desired trade-off. In addition, we note that removing transitive edges disproportionately removes shallow, wide spanning trees from the ancestry graph. If there is a biological reason to suspect that the true evolutionary tree is of this form, PTR may not be appropriate. On the other hand, if it is believed that the underlying tree is likely to be deep and narrow, then PTR becomes even more viable.

B. Real Data

We investigated the strict and approximate ancestry graph approaches on data from patients with chronic lymphocytic leukemia (CLL) [21] and clear cell renal cell carcinoma (ccRCC) [22]. For the CLL patients, we analyzed VAFs from both $100000\times$ coverage targeted deep sequencing and $40\times$ coverage whole genome sequencing (WGS). The ccRCC data was collected using amplicon sequencing with an average coverage over $400\times$ [22]. See Table I for a summary of these datasets. For each dataset, we applied both the approximate and strict ancestry graph methods to identify possible clonal

TABLE I
DATASET SUMMARY

Patient	Samples	Mutations[a]	# Clusters	$\lvert\mathcal{T}(G_F)\rvert$
CLL003 (deep)	5	15/20	4	0
CLL003 (WGS)	5	13/30	4	0
CLL006 (deep)	5	5/10	5	2
CLL006 (WGS)	5	6/16	5	0
CLL077 (deep)	5	12/16	4	1
CLL077 (WGS)	5	16/20	4	0
EV003	8	12/16	4, 5, 6	0
EV005	7	61/64	5, 6	0
EV006	9	52/57	5	0
EV007	8	54/56	4, 5	0
RK26	11	62/62	4, 5, 6	0
RMH002	5	48/48	5, 6	0
RMH004	6	126/126	5, 6	0
RMH008	8	69/71	5, 6	0

[a]After/before filtering out mutations with VAF above 0.5.

trees. When no solutions existed under the standard sum condition, we employed the relaxed sum condition (5), choosing ε as small as possible for one spanning tree to be valid. Before constructing the ancestry graph, we used k-means to cluster mutations by their frequencies across all samples, choosing the number of clusters manually. When the number of mutation clusters was unclear, we performed the analysis with several possible values of k. To eliminate mutations that may have suffered copy number aberrations, we discarded any mutation with a VAF over 0.5.

1) Rarity of Strict Solutions: Of the 11 patients, only CLL006 and CLL077 admitted E-VAFFP solutions, and only in the $100000\times$ coverage targeted sequencing data. EV003 also yielded a valid clonal tree, but only after we removed the R9 sample, a perceived outlier. In all other cases, we had to use the approximate ancestry graph and relax the sum condition in order to find likely clonal trees. This pattern agrees with the finding in simulated data that E-VAFFP solutions are rare and reinforces the importance of coverage in solution existence.

For the datasets in which an E-VAFFP solution existed, we observed one compatible tree in the two patients with four clusters and two trees in the patient with five clusters. For comparison, in simulated data, 19% of the $n = 4$ solvable trials had one tree and 12% of the $n = 5$ solvable trials had two trees.

2) WGS and Targeted Sequencing Agreement: In each of the three CLL patients, the trees identified from WGS data were topologically identical to the trees identified from the deep sequencing data, regardless of whether we used the strict or approximate approaches. The few differences in labeling were due to mutations that were either absent or filtered in one of the two datasets or that were clustered differently due to noise in the WGS data. See Fig. 6 for the CLL077 variant frequencies showing increased noise in the WGS data and see Fig. 7 for the CLL077 trees derived from the deep (left) and WGS (right) datasets.

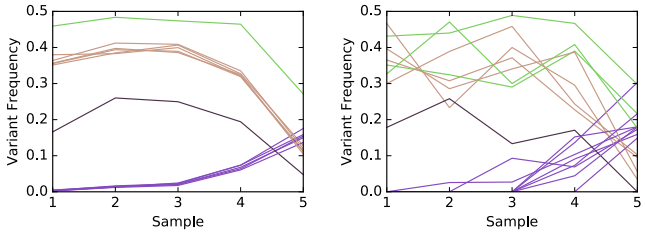


Fig. 6. Variant frequencies over five samples for patient CLL077 in targeted deep sequencing data (left) and whole genome sequencing (right) [21]. Colors of arcs indicate which mutations were clustered together using k-means.

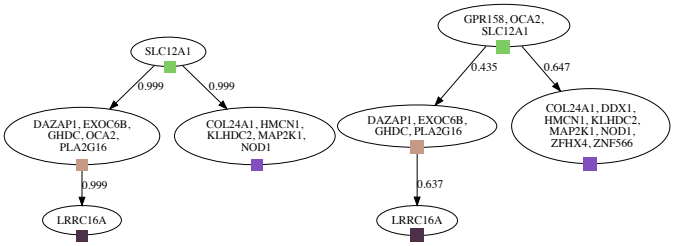


Fig. 7. Clonal trees identified for patient CLL077 from deep sequencing (left) and WGS (right) as the max-weight spanning trees of the respective approximate ancestry graphs. Edge weights are the probability of the relationship and color labels correspond to clusters in Fig. 6. The movement of OCA2 to the root is due to different clustering as a result of noise (see Fig. 6). DDX1, ZFHX4, and ZNF566 were not represented in the deep sequencing data, while GPR158 was filtered out in the deep sequencing data due to VAF over 0.5. The WGS tree required a sum condition relaxation of $\varepsilon = 0.048$.

Additionally, omitting the mutations we filter out due to VAF over 0.5, the CLL trees we identified match exactly with those found by other methods of clonal tree inference, namely CITUP [10] and PhyloSub [12]. Our CLL077 tree also includes the two mutational branches found by AncesTree [9]. In particular, our tree for CLL003, which was generated using the approximate ancestry graph and the relaxed sum condition, agrees exactly with those reported by PhyloSub and CITUP.

Notably, the same trees that obeyed the sum condition in the CLL006 and CLL077 deep sequencing data violated it in the WGS data, by margins of 0.101 and 0.048. This shows that adding noise led to significant violations of the sum condition and to the non-existence of strict solutions. The agreement of the approximate trees generated from noisy WGS data with the trees generated from high-coverage data provides evidence that the relaxed sum condition and approximate ancestry graph allow us to correctly identify likely clonal trees even when noise makes the sum condition unsatisfiable. It is worth noting that the CLL trees had a small number of clusters: either 4 or 5. This is in the regime found in simulated data in which the approximate approach outperforms the strict approach (see Fig. 4).

IV. Discussion

We explored the inference of tumor evolutionary history from SNV frequency data obtained from multi-sample bulk sequencing using the ancestry graph method of [9]. This method is founded on the infinite sites assumption (ISA) and further simplifies the problem by ignoring copy number aberrations. We evaluated the effect of noise on the existence and quality of candidate clonal trees. We also defined the partial transitive reduction of a graph and showed that it can be used to simplify the ancestry graph while on average preserving spanning trees similar to the underlying evolutionary tree. We applied these methods to real cancer datasets, confirming our findings in simulated data about the existence of strict solutions and the viability of the approximate approach.

In simulated data, we confirmed that high noise decreases the probability of strict clonal tree existence. However, in the rare case that trees can be identified in high-noise data, they tend to be better than the more common trees found from low-noise data. This shows that trees similar to the underlying tree are more robust to noise than dissimilar trees. Meanwhile, we showed that the approximate ancestry graph method provides better trees than the strict approach when there are few mutations and worse trees when there are many mutations. We hope that our analysis here will be useful to those analyzing and interpreting real tumor phylogenies constructed using methods that rely on the infinite sites assumption.

Several unanswered questions remain. For instance, we observed that higher coverage decreased the average number of correctly reported ancestral relationships. We are curious to know if this trend continues with more extreme coverages and to understand why this occurs. We are also interested in how the topology of the underlying evolutionary tree affects the ancestry graph method. Early results indicate that wide, shallow evolutionary trees are better represented by $\mathcal{T}(G_F)$, but more investigation is needed to characterize and interpret this effect. Future work should also address the impact of noise on methods that relax the ISA or that consider mutations more complex than SNVs. Finally, the analysis of long-read and single-cell sequencing data will need further attention as these technologies become increasingly feasible, since both show promise in improving phylogeny inference [20].

References

[1] P. C. Nowell, "The clonal evolution of tumor cell populations," *Science*, vol. 194, no. 4260, pp. 23–28, Oct. 1976.

[2] M. Gerstung *et al.*, "The evolutionary history of 2,658 cancers," *bioRxiv*, 2017. [Online]. Available: https://www.biorxiv.org/content/early/2017/08/30/161562

[3] N. McGranahan and C. Swanton, "Clonal heterogeneity and tumor evolution: past, present, and the future," *Cell*, vol. 168, no. 4, pp. 613–628, Feb. 2017.

[4] R. Fisher, L. Pusztai, and C. Swanton, "Cancer heterogeneity: implications for targeted therapeutics," *Brit. J. Cancer*, vol. 1805, no. 1, pp. 105–117, Jan. 2010.

[5] X. Sun and Q. Yu, "Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment," *APS*, vol. 36, no. 10, pp. 1219–1277, Oct. 2015.

[6] H. Zafar, A. Tzen, N. Navin, K. Chen, and L. Nakhleh, "SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models," *Genome Biology*, vol. 18, no. 1, p. 178, Sep. 2017.

[7] K. Jahn, J. Kuipers, and N. Beerenwinkel, "Tree inference for single-cell data," *Genome Biology*, vol. 17, p. 86, May 2016.

[8] M. El-Kebir, "SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error," *Bioinformatics*, vol. 34, no. 17, pp. i671–i679, 2018.

[9] M. El-Kebir, L. Oesper, H. Acheson-Field, and B. J. Raphael, "Reconstruction of clonal trees and tumor composition from multi-sample sequencing data," *Bioinformatics*, vol. 31, no. 12, pp. i62–i70, Jun. 2015.

[10] S. Malikic, A. W. McPherson, N. Donmez, and C. S. Sahinalp, "Clonality inference in multiple tumor samples using phylogeny," *Bioinformatics*, vol. 31, no. 9, pp. 1349–1356, May 2015.

[11] V. Popic, R. Salari, I. Hajirasouliha, D. Kashef-Haghighi, R. B. West, and S. Batzoglou, "Fast and scalable inference of multi-sample cancer lineages," *Genome Biology*, vol. 16, p. 91, May 2015.

[12] W. Jiao, S. Vembu, A. G. Deshwar, L. Stein, and Q. Morris, "Inferring clonal evolution of tumors from single nucleotide somatic mutations," *BMC Bioinformatics*, vol. 15, p. 35, Feb. 2014.

[13] I. Hajirasouliha, A. Mahmoody, and B. J. Raphael, "A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data," *Bioinformatics*, vol. 30, no. 12, pp. i78–i86, 2014.

[14] J. Kuipers, K. Jahn, B. J. Raphael, and N. Beerenwinkel, "Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors," *Genome Research*, 2017.

[15] P. Bonizzoni, S. Ciccolella, G. Della Vedova, and M. Soto, "Beyond perfect phylogeny: Multisample phylogeny reconstruction via ILP," in *Proc. 8th ACM Int. Conf. on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2017, pp. 1–10.

[16] Y. Jiang, Y. Qiu, A. J. Minn, and N. R. Zhang, "Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing," *PNAS*, vol. 113, no. 37, pp. E5528–E5537, Sep. 2016.

[17] A. G. Deshwar, S. Vembu, C. K. Yung, G. H. Jang, L. Stein, and Q. Morris, "PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors," *Genome Biology*, vol. 16, p. 35, Feb. 2015.

[18] M. El-Kebir, G. Satas, L. Oesper, and B. J. Raphael, "Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures," *Cell Systems*, vol. 3, pp. 43–53, Jul. 2016.

[19] Y. Qiao, A. R. Quinlan, A. A. Jazaeri, R. G. Verhaak, D. A. Wheeler, and G. T. Marth, "SubcloneSeeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization," *Genome Biology*, vol. 15, no. 8, p. 443, 2014.

[20] D. Pradhan and M. El-Kebir, "On the non-uniqueness of solutions to the perfect phylogeny mixture problem," in *RECOMB Int. Conf. on Comparative Genomics*. Springer, 2018, pp. 277–293.

[21] A. Schuh *et al.*, "Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns," *Blood*, vol. 120, no. 20, pp. 4191–4196, Nov. 2012.

[22] M. Gerlinger *et al.*, "Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing," *Nature Genetics*, vol. 46, no. 3, pp. 225–233, Feb. 2014.

[23] H. N. Gabow and E. W. Myers, "Finding all spanning trees of directed and undirected graphs," *SIAM J. Comput.*, vol. 7, no. 3, pp. 280–287, Aug. 1978.

[24] W. T. Tutte, "The dissection of equilateral triangles into equilateral triangles," *Proc. Cambridge Philosoph. Soc.*, vol. 44, no. 4, pp. 463–482, Oct. 1948.

[25] A. V. Aho, M. R. Garey, and J. D. Ullman, "The transitive reduction of a directed graph," *SIAM J. Comput.*, vol. 1, no. 2, pp. 131–137, Jun. 1972.

[26] H. T. Hsu, "An algorithm for finding a minimal equivalent graph of a digraph," *J. ACM*, vol. 22, no. 1, pp. 11–16, Jan. 1975.

[27] J. Cook, "Exact calculation of beta inequalities," University of Texas, M. D. Anderson Cancer Center, Tech. Rep., 2005.

[28] P. M. Camerini, L. Fratta, and F. Maffioli, "The k best spanning arborescences of a network," *Networks*, vol. 10, pp. 91–110, 1980.

[29] L. Devroye, *Non-Uniform Random Variate Generation*. New York, NY: Springer-Verlag, 1986, p. 568.

[30] K. Govek, C. Sikes, and L. Oesper, "A consensus approach to infer tumor evolutionary histories," in *Proc. 2018 ACM Int. Conf. on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2018, pp. 63–72.

[31] D. F. Robinson and L. R. Foulds, "Comparison of phylogenetic trees," *Mathematical Biosciences*, vol. 53, no. 1-2, pp. 131–147, Feb. 1981.