

RESEARCH

Open Access



# Parameter, noise, and tree topology effects in tumor phylogeny inference

Kiran Tomlinson<sup>1,2</sup> and Layla Oesper<sup>1\*</sup>

From IEEE International Conference on Bioinformatics and Biomedicine 2018  
Madrid, Spain. 3-6 December 2018

## Abstract

**Background:** Accurate inference of the evolutionary history of a tumor has important implications for understanding and potentially treating the disease. While a number of methods have been proposed to reconstruct the evolutionary history of a tumor from DNA sequencing data, it is not clear how aspects of the sequencing data and tumor itself affect these reconstructions.

**Methods:** We investigate when and how well these histories can be reconstructed from multi-sample bulk sequencing data when considering only single nucleotide variants (SNVs). Specifically, we examine the space of all possible tumor phylogenies under the infinite sites assumption (ISA) using several approaches for enumerating phylogenies consistent with the sequencing data.

**Results:** On noisy simulated data, we find that the ISA is often violated and that low coverage and high noise make it more difficult to identify phylogenies. Additionally, we find that evolutionary trees with branching topologies are easier to reconstruct accurately. We also apply our reconstruction methods to both chronic lymphocytic leukemia and clear cell renal cell carcinoma datasets and confirm that ISA violations are common in practice, especially in lower-coverage sequencing data. Nonetheless, we show that an ISA-based approach can be relaxed to produce high-quality phylogenies.

**Conclusions:** Consideration of practical aspects of sequencing data such as coverage or the model of tumor evolution (branching, linear, etc.) is essential to effectively using the output of tumor phylogeny inference methods. Additionally, these factors should be considered in the development of new inference methods.

**Keywords:** Cancer genomics, Tumor phylogeny, Evolution

## Background

Cancer is caused by somatic mutations in a single founder cell that lead to the unrestrained proliferation of the descendants of that cell. According to the clonal theory of cancer [1], descendants of the founder cell will continue to acquire new somatic mutations that may drive disease progression. Since different descendants acquire distinct mutations, the history of a tumor can be described as a type of phylogenetic tree. In these trees, vertices represent tumor cell populations, or clones, each with their

own set of somatic mutations, and edges represent ancestral relationships between populations. Several different models of tumor evolution have been proposed, including linear, branching, neutral, and punctuated evolution [2–4], describing different patterns of how and when new tumor populations arise. As a result of these evolutionary processes, a tumor itself may be a heterogeneous mix of different tumor cell populations.

A number of recent studies have highlighted the prevalence of such *intra-tumor heterogeneity* [5–7] across many different cancer types. Computational methods for analyzing intra-tumor heterogeneity, including characterization of the populations in a particular tumor and how they evolved, have important implications for understanding

\*Correspondence: [loesper@carleton.edu](mailto:loesper@carleton.edu)

<sup>1</sup>Department of Computer Science, Carleton College, 1 N College St, 55057 Northfield, MN, USA

Full list of author information is available at the end of the article



and, ultimately, treating the disease [8, 9]. For example, cancer types that are typically detected late in the tumor's evolution, such as pancreatic cancer, often have poor prognosis [10]. Intra-tumor heterogeneity may play a key role in therapeutic failure in such instances if the treatment only targets certain tumor cell populations [11]. Treatment strategies that take the evolutionary history of a tumor into account by specifically targeting clonal mutations (those present in every tumor cell) [12] or that combine drugs based on a patient's specific tumor evolutionary history [13] have the potential to be more effective. However, for such approaches to be feasible, there is an imperative need for better approaches to inferring and analyzing the evolutionary history of a single tumor.

There has been an increased recent 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. A number of recent approaches have focused on using single-cell sequencing data to reconstruct tumor phylogenies [14–16]. Ultimately, such methods have the promise to provide improved resolution for such reconstructions. However, currently single-cell sequencing still suffers from both high error rates and high cost. While technological and methodological developments are beginning to alleviate these issues, the majority of the currently available data is still from bulk sequencing experiments. Specifically, most large scale cancer studies such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have made this type of data widely available. Thus, there is still much to be gained from methods that analyze bulk data, while single-cell methods continue to mature. Therefore, we focus here on the data from more economical bulk sequencing. However, there are still many challenges and sources of error in this type of data. In bulk sequencing, collections of potentially heterogeneous cells are sequenced together, which obfuscates the coincidence of mutations. Sources of error include the sequencing process, read alignment, and variant calling algorithms. Thus, specialized methods are required to robustly analyze noisy bulk sequencing data.

Many recent computational methods have been developed to infer tumor phylogenetic trees using multi-sample bulk sequencing data. A large fraction of these methods consider only single nucleotide variants (SNVs) [17–21] and use rules regarding the observed frequencies of each such mutation to identify possible ancestral relationships. In particular, these methods use the infinite sites assumption (ISA), which states that any locus in the genome mutates at most once during the history of tumor, a simplification that makes the underlying computational problem more tractable. For example, AncestryTree [17] constructs a graph called the ancestry graph using mutation frequencies and then finds spanning trees of that graph

adhering to the ISA. However, increasing reports that the ISA is often violated in cancer [22] have led to the development of methods that relax the ISA in some contexts [16, 23]. Some methods also consider structural variants or copy number aberrations [24–27] in addition to SNVs, but this has proven challenging. Finally, several methods allow for multiple tumor evolutionary trees consistent with a given sequencing dataset by enumerating these trees [18, 26, 27]. Along these lines, a recent paper [28] 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 variety of sources of noise present in bulk sequencing data and to what extent these conclusions apply to real sequencing data. Finally, it is not obvious how existing tumor phylogeny inference methods are affected by the distinct tree topologies resulting from different models of tumor evolution such as branching or linear [2].

In this paper, we investigate several extensions to the ancestry graph approach of [17], which relies on the ISA, and quantify when and how well this approach can reconstruct tumor evolutionary histories from multi-sample bulk sequencing data. In particular, we focus on the performance of this method when applied to noisy data. Our specific methodological contributions include: (1) a relaxation of the ancestry graph approach that makes it more robust to noise; and (2) a method for simplifying the ancestry graph that leads to reduced computational costs. Furthermore, our contributions include extensive analysis of the effects of coverage, noise, evolutionary tree topology, and other parameters in reconstructing clonal trees in simulated data. This analysis has numerous potential future implications for both experimental design and algorithm development. Finally, we apply our methods to cancer sequencing datasets from two studies [29, 30].

## Methods

This section is organized as follows. We begin by outlining the existing ancestry graph method [17] and then formalize the new problem of using this method to enumerate all tumor phylogenies consistent with a particular dataset. We then describe a relaxation that improves the method's robustness to noise, and introduce a graph simplification that reduces computational cost. Finally, we describe our data simulation procedure and our tree evaluation metric.

### Problem Formalization

#### Definitions

We use  $s$  to denote the number of samples sequenced from a tumor and  $n$  to denote the number of mutations observed across all samples. We label these mutations  $1, \dots, n$ . The  $s \times n$  variant allele frequency (VAF) matrix  $F$  stores in entry  $F_{ij}$  the fraction of reads from sample  $i$  containing mutation  $j$ . A *clonal tree*  $T$  (or tumor phylogeny)

is a rooted tree on  $n$  nodes with each node labeled by a distinct mutation. Nodes may also be labeled with disjoint sets of mutations, with a corresponding decrease in the number of nodes. Each node represents a tumor cell population that contains all mutations along its root-node path. The infinite sites assumption (ISA) guarantees that a clonal tree is a perfect phylogeny where mutations evolve without homoplasy. Because of this, we can also represent the tree as an  $n \times n$  clonal matrix  $B$ , in which  $B_{\ell j} = 1$  if cell population  $\ell$  contains mutation  $j$  and 0 otherwise. Finally, the  $s \times n$  usage matrix  $U$  stores in  $U_{i\ell}$  the proportion of cells in sample  $i$  that belong to population  $\ell$ .

#### The VAFFP and the Ancestry Graph

The authors of [17] formalized the Variant Allele Frequency Factorization Problem (VAFFP), also called the Perfect Phylogeny Mixture Problem in [28], as follows:

*Given:* A VAF matrix  $F$ .

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

$$F = \frac{1}{2}UB. \quad (1)$$

The  $1/2$  factor appears because we assume that all mutations are heterozygous SNVs (implicitly assuming no copy number aberrations). The VAFFP has been shown to be NP-complete [17], but in practice, many datasets are small enough that finding solutions is feasible.

The authors of [17] describe a method for solving the VAFFP using the *ancestry graph* of  $F$  (see Fig. 1 for a visual overview of this approach). In order to avoid confusion, we will often 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  $F_{ij} \geq F_{ik} \forall i \in \{1, \dots, s\}$ . These edges encode the *ancestry condition*: under the ISA, 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:

$$\sum_{k \in C(j)} F_{ik} \leq F_{ij} \quad \forall i \in \{1, \dots, s\}. \quad (2)$$

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

Every spanning tree  $T$  of  $G_F$  that adheres to the sum condition corresponds to a VAFFP solution (see the right-most part of Fig. 1 for examples). The clonal matrix  $B$  can be constructed from  $T$  by tracing through each root-leaf path in  $T$ . The matrix  $U$  can be efficiently computed using the following equation from [17]:

$$U_{ij} = 2 \left( F_{ij} - \sum_{k \in C(j)} F_{ik} \right). \quad (3)$$

#### The Enumeration Variant Allele Frequency Factorization Problem (E-VAFFP)

Here, we define the focus of our work, the enumeration version of the VAFFP.

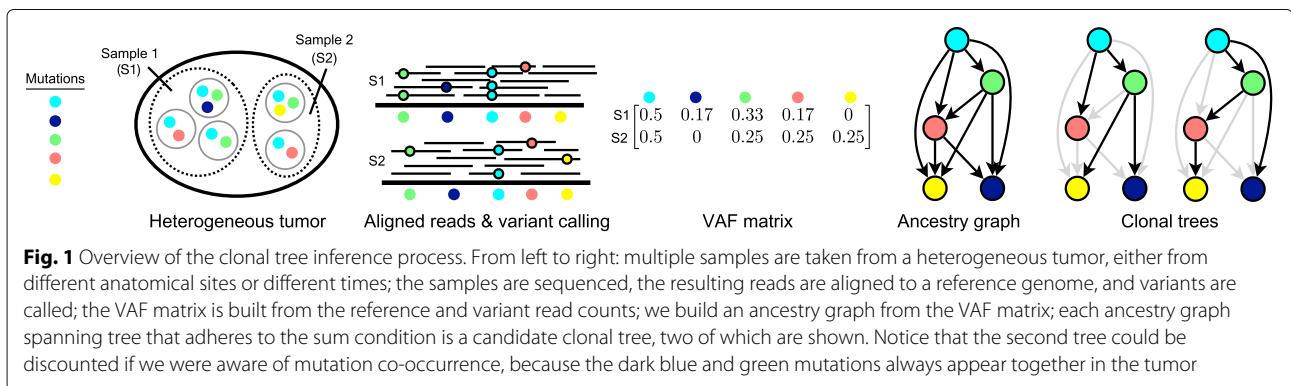
*Given:* A VAF matrix  $F$ .

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

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

#### Finding and Counting E-VAFFP Solutions

In order to solve the E-VAFFP, we employ a modified version of the Gabow-Myers algorithm [31]. Specifically, this algorithm uses a structured depth-first search in order to recursively construct all spanning trees of the graph. It is straightforward to modify this approach to avoid execution branches that violate the sum condition, as has been done previously by [19, 26, 28]. Additionally, we note that the number of such spanning trees of  $G_F$  is the product of its non-root in-degrees [28, 32]. This provides an upper bound on  $|\mathcal{T}(G_F)|$ .



## Relaxing the E-VAFFP

### Approximate Ancestry Graph

Real DNA sequencing data is often quite noisy, but the E-VAFFP assumes that  $F$  is measured exactly. In real data,  $G_F$  often has no spanning trees. To handle less idealized data, we use a method based on the probabilistic approach from [17]. 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 reads as being drawn from a binomial distribution with a flat prior on the proportion parameter. Thus, we model the resulting posterior distribution for the VAF of mutation  $j$  in sample  $i$  with observed variant and reference read counts  $v_{ij}$  and  $r_{ij}$ , respectively, with the beta-distributed random variable  $X_{ij} \sim Beta(v_{ij}+1, r_{ij}+1)$ , as done in [17]. 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:

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

The probabilities on the right hand side of (4) can be calculated from the read counts that generate  $F$  using the approach described in [33], as both of the random variables  $X_{ij}$  and  $X_{ik}$  are beta-distributed.

Just as we did in the strict ancestry graph, we can also use the Gabow-Myers algorithm [31] to enumerate all spanning trees of the approximate ancestry graph whose observed frequencies satisfy the sum condition. In this context, we refer to such trees as *valid* spanning trees. Once these are computed, we can then select the most probable (i.e. max weight) tree. Alternatively, if the graph has too many spanning trees to fully enumerate, we can use the algorithm of [34] to list weighted spanning trees in descending weight order until we find one satisfying the sum condition. Unlike Gabow-Myers, this algorithm is not easy to modify to include the sum condition. Using this method, we can potentially find the most probable clonal tree without the need to enumerate every tree. However, this approach may be significantly slower when no valid spanning trees exist as the method 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 yield more E-VAFFP solutions than the strict ancestry graph. Any tree that violates the sum condition in the strict graph will necessarily violate it in the approximate graph, because the sum condition only depends on the VAF matrix  $F$ . Additionally, any approximate graph spanning tree not present in the strict graph must violate the

ancestry condition (and thus the sum condition), since it includes an edge not present in the strict graph.

Nonetheless, the approximate ancestry graph still provides two key benefits. First, it orders solutions by likelihood, and second, it allows us to explore novel tree topologies not present in the strict graph if we also weaken the sum condition.

### Relaxed Sum Condition

Adding leniency to the sum condition allows the identification of possible clonal trees rendered invalid by noise. For a small *error threshold*  $\varepsilon$ , we can relax the sum condition to require that:

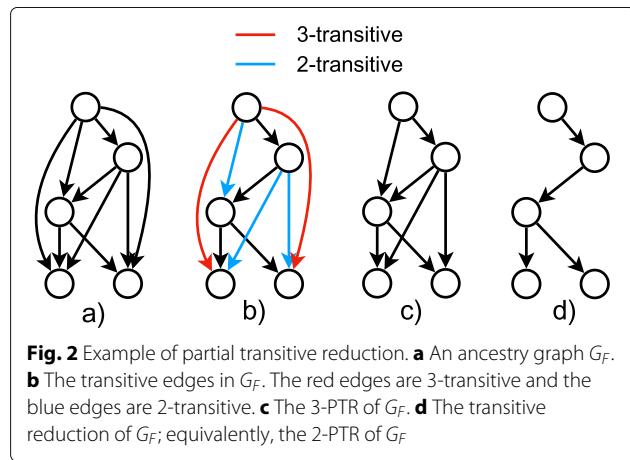
$$\sum_{k \in C(j)} F_{ik} \leq F_{ij} + \varepsilon \quad \forall i \in \{1, \dots, s\} \quad (5)$$

We then can identify the smallest  $\varepsilon$  resulting in one valid spanning tree. In other words, we find the spanning tree with the smallest maximal sum condition violation. We note that [19] also relaxes the sum condition in this way, but does not use it in conjunction with an approximate ancestry graph.

### Pruning Transitive Edges

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  $10^{17}$ , making clonal tree inference extremely slow. We therefore explore the removal of transitive edges from the ancestry graph as a means of reducing the spanning trees present in the graph while maintaining core ancestral relationships (see Fig. 2). This approach may be especially useful if the underlying tumor has a branching rather than a linear topology.

For a directed acyclic graph  $G$ , we say that an edge  $(u, v) \in G$  is  $k$ -transitive if there is a path from  $u$  to  $v$  of length  $k$  (see Fig. 2b). Additionally, we define an edge to be



$\geq k$ -transitive if it is  $i$ -transitive for some  $i \geq k$ . By removing all  $\geq k$ -transitive edges from  $G$  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 [35] 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 [36]. Then, we can easily identify if  $(u, v)$  is  $\geq k$ -transitive by checking the path length from  $u$  to  $v$  in  $R$ . We can do this efficiently by pre-computing the all-pairs shortest path matrix of  $R$  with  $n$  breadth-first searches.

### Simulating Noisy VAF Data

We use simulated data to assess our methods. Our data simulation process consists of four steps: (1) randomly generate an evolutionary tree topology, (2) choose the cellular frequencies, (3) determine the mutation frequencies, and (4) draw variant reads from a binomial distribution, allowing direct computation of  $F$ .

Given the number of mutations  $n$ , the number of samples  $s$ , and the average sequencing coverage  $c$ , we first generate a random tumor phylogeny  $T$ , referred to as the *underlying tree* for the simulation, and an  $s \times n$  VAF matrix consistent with  $T$ . For simplicity, each clone acquires exactly one new somatic mutation, so we also call  $n$  the number of clones. We construct  $T$  iteratively by adding each mutation as the child of a random node already in  $T$ . From  $T$ , we compute the clonal matrix  $B$  described in a previous section. We then generate the cellular frequencies of the  $n$  clones. Clone  $i$  is assigned frequency  $u_i$  such that  $\sum_i u_i = 1$ . To pick  $u_1, \dots, u_n$ , we sample uniformly from all possible frequency values using the standard simplex method from [37].

We then calculate the tumor's mutation frequencies. Using the row vectors  $\vec{f}$  and  $\vec{u}$  to store mutation and cellular frequencies, respectively, we find  $\vec{f}$  using (1):

$$\vec{f} = \frac{1}{2} \vec{u} B \quad (6)$$

Finally, we simulate reads taken from the  $s$  samples. For simplicity, we assume the tumor is completely mixed, so that the expected cellular composition of each sample matches that of the tumor. For each sample  $i$  and for each mutation  $j$ , we simulate  $r_{ij} \sim \text{Poisson}(c)$  reads, where  $c$  is the mean coverage. We then draw the number of variant reads  $v_{ij}$  of mutation  $j$  in sample  $i$  from a binomial distribution:  $v_{ij} \sim \text{Binom}(r_{ij}, f_j)$ . The  $s \times n$  VAF matrix  $F$  then contains entries  $F_{ij} = v_{ij}/r_{ij}$ .

Additionally, we simulate sampling and sequencing noise by adding *overdispersion* to the binomial distribution. We replace  $f_j$  with a beta-distributed random variable

with mean  $f_j$ . The beta distribution parameters  $\alpha$  and  $\beta$  are chosen to be:

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

where  $\rho \in (0, 1)$  is the overdispersion parameter. This results in a beta distribution with mean  $f_j$  and with variance proportional to  $\rho$ . We simulate sequencing data with less noise by setting  $\rho$  closer to 0 and more noise by setting  $\rho$  closer to 1. The case when  $\rho = 0$  corresponds to no overdispersion.

### Evaluation of Reconstructed Trees

To quantify the quality of the clonal trees we enumerate, we use the mean ancestor-descendant (A-D) distance [38] between trees in  $\mathcal{T}(G_F)$  and the underlying tree  $T$ . Note that standard phylogenetic distance measures, including Robinson-Foulds [39], do not apply to clonal trees since they contain internal node labels. To quantify the useful information gained from our solutions, we measure how much more similar trees in  $\mathcal{T}(G_F)$  are to the underlying tree than an equal number of random 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

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

A-D improvement measures the proportional decrease in incorrect ancestral relationships relative to the random baseline. For example, an A-D improvement of 0 means that trees in  $\mathcal{T}(G_F)$  are no better than random, while an A-D improvement of 1 means that  $\mathcal{T}(G_F) = \{T\}$ , the correct tree.

## Results

We investigated strict and approximate E-VAFFP solutions both in simulated noisy data and in two real datasets of 3 chronic lymphocytic leukemia (CLL) patients from [29] and 7 clear cell renal cell carcinoma (ccRCC) patients from [30]. We also separately evaluated the usefulness of pruning transitive edges from the strict ancestry graph.

### E-VAFFP Solutions on Simulated Data

We first present findings on the existence and quality of E-VAFFP solutions in simulated noisy DNA sequencing data. We begin by describing how parameters affect the likelihood of finding compatible trees and then address how similar those inferred trees are to the underlying tree. Lastly, we examine how the topology of the underlying tree affects  $\mathcal{T}(G_F)$ .

#### E-VAFFP Solution Existence

In simulated data, we found that there are typically no E-VAFFP solutions due to sum condition violations. With

more clones, more samples, lower coverage, and higher noise, the probability of finding a solution decreases further. We generated 10000 simulated datasets and ran the ancestry graph method for each parameter value ( $n$  between 3 and 12,  $s$  between 1 and 15, coverage between  $50\times$  and  $200\times$ , and  $\rho$  between 0 and 0.09). We then computed the proportion of trials with at least one E-VAFFP solution, which we call *solvable* trials. We tested each parameter individually, with default values of  $n = 10$ ,  $s = 5$ ,  $60\times$  coverage, and  $\rho = 0$ .

With all parameters at their default settings, the proportion of solvable trials was only 14%. Increasing the coverage caused a dramatic increase in this fraction, up to 47% at  $200\times$  coverage. On the other hand, higher overdispersion had a strong negative effect on solvable trials, with 89 of the 10000 trials solvable at  $\rho = 0.09$ . High sample count had an equivalently strong negative impact, with only 103 trials exhibiting an E-VAFFP solution at  $s = 15$ . Corroborating these findings, we also found that E-VAFFP solutions rarely exist in lower-coverage real data, which we discuss in a later section.

#### E-VAFFP Solution Quality

We found that when using default parameters, the trees in  $\mathcal{T}(G_F)$  showed a mean A-D improvement of 0.64. This corresponds to a retention of 64% of ancestral patterns in the data missed by the random baseline. Increasing the number of clones  $n$  makes valid solutions more rare and further decreases the quality of solutions when they do appear. Conversely, increasing the number of samples  $s$  shows an improvement in the similarity of trees  $\mathcal{T}(G_F)$  to the underlying tree as shown in Fig. 3. The trends we see here for noisy data correspond to those reported by [28] on error-free data.

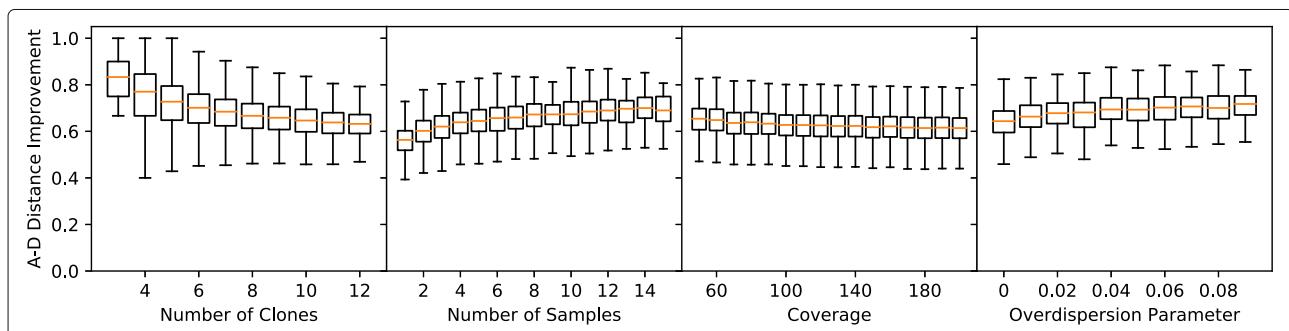
When we conditioned on the existences of solutions, we counter-intuitively found that higher noise improves solution quality (see Fig. 3). For instance, high coverage slightly decreased A-D improvement, from 0.65 at  $50\times$  to 0.61 at  $200\times$ . In the rare case that solutions existed, trials with

more overdispersion also resulted in better-quality trees, with an A-D improvement of 0.64 at  $\rho = 0$  and 0.72 at  $\rho = 0.09$ . These findings suggest that spanning trees more similar to the underlying tree are less likely to be rendered invalid by noise. Therefore, noise preferentially disqualifies bad trees from  $\mathcal{T}(G_F)$ , resulting in a higher mean A-D improvement. Importantly, the decrease in solution existence is so dramatic that it swamps these modest quality gains, making phylogeny inference worse in high-noise data. For example, the total number of correctly inferred ancestral relationships in all trials does actually decrease as we add more overdispersion, since so few trials are solvable at high  $\rho$ .

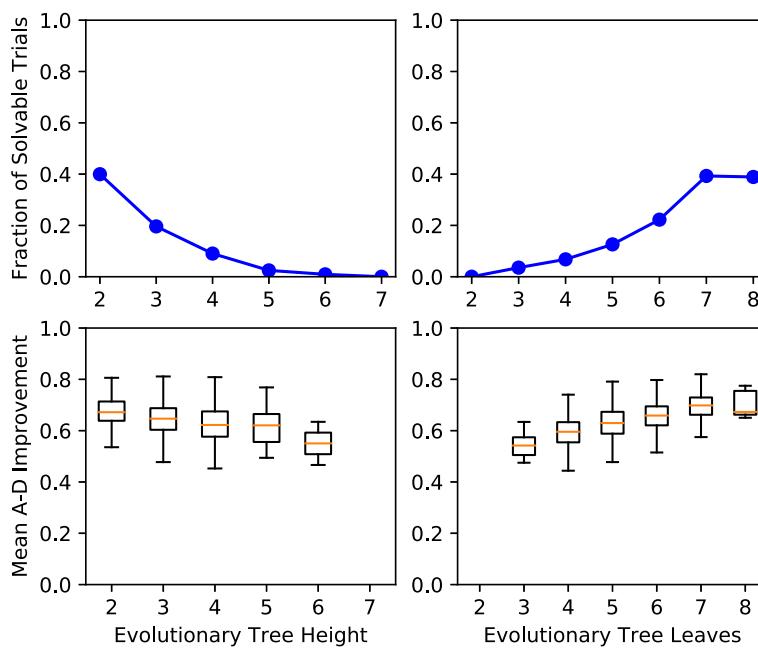
#### Effects of Underlying Tree Topology on E-VAFFP Solutions

The topology of a tumor's underlying evolutionary tree can have a strong effect on the accuracy of reconstruction methods. Trees that are wider (more leaves) and shallower (lower tree height) than average randomly generated trees are said to have a *branching* topology. We find that the features of a branching topology are more likely to result in E-VAFFP solutions and have improved solution quality (see Fig. 4). We also performed these analyses using two additional measures of tree topology, single child fraction and mean subtree height. Similarly, we found that trees that have a low single child fraction and a low mean subtree height (both features of branching trees) also are more likely to yield E-VAFFP solutions and have improved solution quality (see Additional file 1).

The reason why E-VAFFP solutions perform better on branching trees is not immediately obvious. One possible explanation relates to the effect of simulated noise on the resulting ancestry graph. If a descendant mutation and its ancestor have very similar mutation frequencies, then even a small amount of noise could reverse the order of the observed frequencies, violating the ancestry condition. In a totally linear tree, it is possible that each pair of connected vertices has similar frequencies in some sample. Thus, every edge in the ancestry graph has the potential



**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



**Fig. 4** Effects of tree topology on E-VAFFP solution existence and quality. The top row shows the effects of underlying tree height and leaf count on the fraction of trials with any compatible trees. The bottom row shows the effects of these tree metrics on solution quality. Shallow, wide trees yield better reconstructions

to be reversed by noise. In contrast, in a totally branching tree, since all children vertices must adhere to the sum condition, only one child can have a frequency very similar to its parent. Thus, fewer edges in the ancestry graph are likely to be prone to noise. Hence, trees that exhibit more branching may appear more robust under the E-VAFFP than linear trees.

#### Evaluation of Approximate Solutions on Simulated Data

Just as with strict E-VAFFP solutions, we examined the quality of solutions derived from the approximate ancestry graph. We also determined the viability of relaxing the sum condition and investigated the validity of the edge weighting function used to construct the approximate graph.

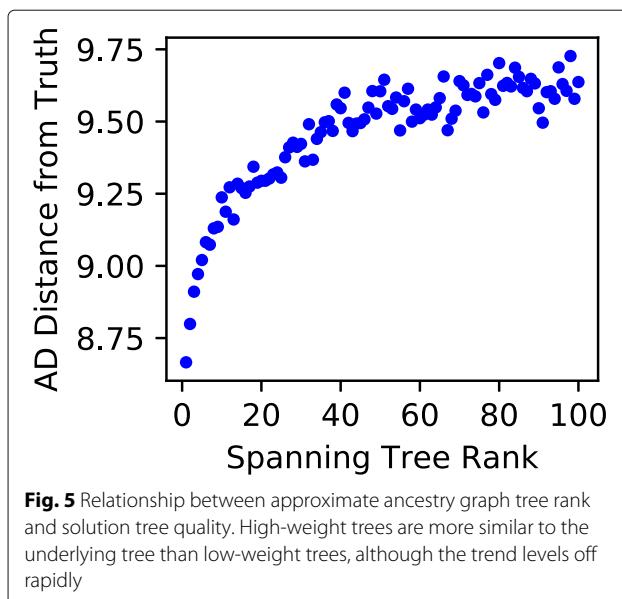
#### Approximate Solution Existence

Even when the error threshold  $\varepsilon$  is small, the relaxed sum condition (5) results in a significant increase in the proportion of solvable trials. We let  $\varepsilon$  range from 0 and 0.05, since real data indicated that sum condition overflows are typically small in practice (Table 2). As we increased  $\varepsilon$  in this range, we observed a proportional increase in the fraction of solvable trials from 14% to 64%. However, there was also a significant increase in the mean number of trees in  $\mathcal{T}(G_F)$  from 2000 to 69000, which dramatically slows down inference. Thus, there is a trade off between the probability of finding a valid tree and computational cost of enumerating these trees.

#### Approximate Solution Quality

The approximate ancestry graph method is founded on the assumption that the weighting function in Eq. 4 accurately represents the probability that the corresponding edge exists in the underlying evolutionary tree. If this is the case, the total weight of a solution tree should be a measure of its quality, and the max-weight tree should be the most similar to the underlying tree. We verified this by comparing the ranks of approximate ancestry graph spanning trees to their the mean A-D distance from the underlying tree. We say that the max-weight valid spanning tree has rank 1 and that the  $i$ th highest weight valid tree has rank  $i$ . We selected the 1104/10000 trials with at least 100 valid spanning trees and sorted the top 100 trees in descending weight order. Then, we aggregated statistics for each tree rank across trials. As expected, we found that high-weight trees are in fact more similar to the underlying tree than lower weight trees (Fig. 5). Specifically, we find that the average rank 1 tree has 6.9% smaller A-D distance to the underlying tree than the average rank 25 tree. This effect begins to level off as rank increases: the average rank 25 tree has just 3.4% smaller A-D distance to the underlying tree than the average rank 100 tree.

We also examined the effects of parameters on the quality of approximate solutions. Solution quality responds in the same way to changes in sample count, coverage, and overdispersion in the approximate ancestry graph as in the strict ancestry graph. However, we found an intriguing difference in the response to number of clones  $n$ . Choosing



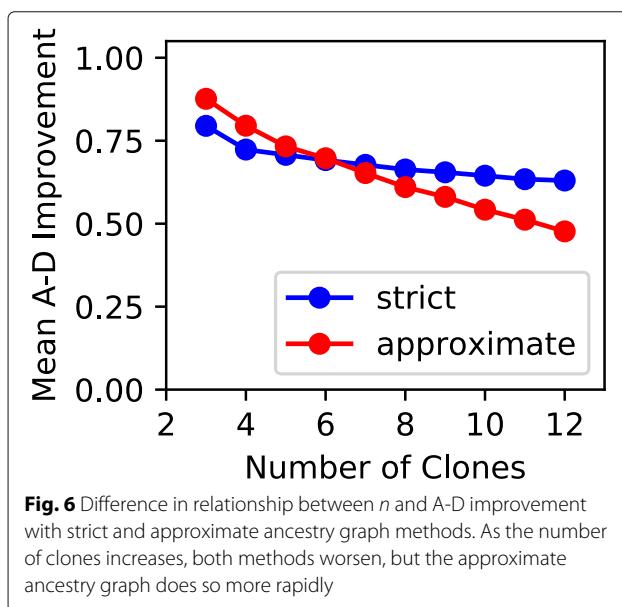
the max-weight valid spanning tree of the approximate graph provides noticeably better solutions than the strict approach for small  $n$ . However, the approximate method drops off more sharply in quality as  $n$  grows, with the crossover point at  $n = 6$  (see Fig. 6). We suspect this is due to inherent bias in high-weight approximate spanning trees, since they become worse than randomly sampled strict spanning trees (as measured by A-D improvement) as  $n$  grows. We investigate this phenomenon in depth in the following section. We also found that relaxing the sum condition caused a gradual linear decrease in the approximate solution quality, from an A-D improvement of 0.54 at  $\varepsilon = 0$  to 0.51 at  $\varepsilon = 0.05$  when the number of clones is

$n = 10$ . The negative effect on quality of relaxing the sum condition lessens when there are fewer clones, and the relaxed sum condition may even improve mean solution quality when  $n < 6$  (see Additional file 3).

#### Tree Rank in the Approximate Ancestry Graph

As we saw in Fig. 6, the quality of solutions derived from the approximate ancestry graph falls off more quickly than the strict E-VAFFP enumeration method as the number of clones increases. We believe this is due to systematic bias in high-weight spanning trees. This bias may arise because edges in the approximate graph are weighted by the probability that one clone 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. This would result in high-weight spanning trees that tend to be shallow and wide.

To assess this conjecture, we gathered data on the relationship between spanning tree rank in the approximate graph and the four topology metrics from the previous section (height, leaf count, single child fraction, and mean subtree height). We found a strong and consistent trend that high-weight trees do in fact tend to be shallower and wider than lower weight trees (Fig. 7). This effect is most pronounced at low ranks, with average heights of 2.65 at rank 1, 2.88 at rank 25, and 3.00 at rank 100. Moreover, the average underlying tree yielding at least 100 solutions has height 3.09. This shows that high-weight spanning trees are biased towards branching topologies. We found the same trend using leaf count, mean subtree height, and single child fraction. Despite this bias, the max-weight tree is still, on average, the best choice available among approximate ancestry graph spanning trees, as seen in the previous section.

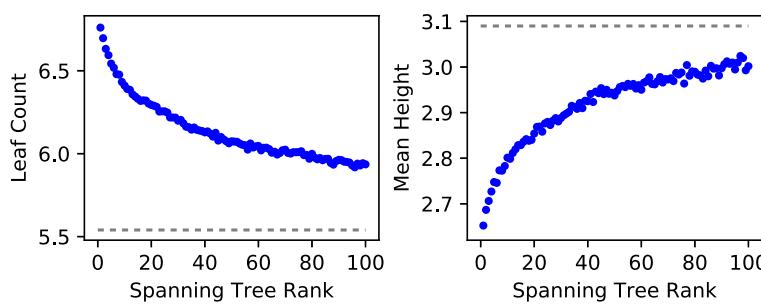


#### Evaluation of Transitive Edge Pruning

We found that partial transitive reduction (PTR) successfully reduces the size of  $\mathcal{T}(G_F)$  while preserving solution quality. We first compared the solution quality and existence that result from applying PTRs to the standard ancestry graph method. Next, we counted the average and maximum number of ancestry graph spanning trees as a measure of performance improvement due to PTR (Fig. 8). Our default parameters were unchanged from the previous experiment.

The 2-PTR (i.e. the canonical transitive reduction) was too extreme to be useful, as it decreased the fraction of solvable trials to 3%. Moreover, 2-PTR also decreased solution quality as measured by mean A-D improvement from 0.64 in the standard ancestry graph to 0.57 (Fig. 8).

On the other hand, higher-order PTR (6+) had almost no effect, as  $\geq 6$ -transitive edge are exceedingly rare in



**Fig. 7** Relationships between approximate ancestry graph tree rank and solution tree topology. The dashed lines show the average values for underlying trees yielding at least 100 spanning trees. On average, high-weight solutions are wider and shallower than lower height solutions. Additionally, they are significantly wider and shallower than the underlying evolutionary trees

ancestry graphs with 10 nodes. However, 4- and 5-PTR showed more promise. Neither had a noticeable impact on the proportion of solvable trials, but they reduced the maximum spanning tree count by 43% and 62%, respectively. At the same time, both 4- and 5-PTR decreased the mean A-D improvement by less than 0.01. The 3-PTR had a correspondingly stronger impact on these quantities, decreasing the mean and maximum spanning tree counts by factors of 7.7 and 9.6 relative to the standard ancestry graph. The proportion of solvable trials shrank by two percentage points with 3-PTR, while the mean A-D improvement was 0.02 worse.

To summarize, we were able to reduce the number of edges in 10-node ancestry graphs without harming solution quality and existence using 3-, 4-, and 5-PTR. Fewer edges results in fewer spanning trees, and thus lower runtime, less memory usage, and the potential to handle more clones. Picking different partial transitive reductions allows us to control the trade-off between these benefits and better solutions. With a different number of nodes, we would have to pick a different PTR to achieve the desired balance.

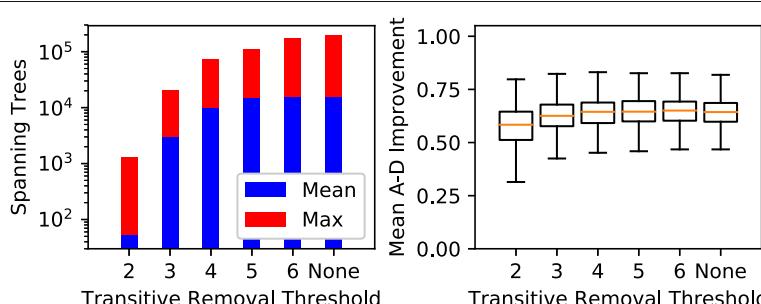
#### Topology Effects of PTR

Removing highly transitive edges from the ancestry graph disproportionately removes wide spanning trees from  $\mathcal{T}(G_F)$  (Fig. 9). We considered 2-, 3-, 4-, and 5-PTR across

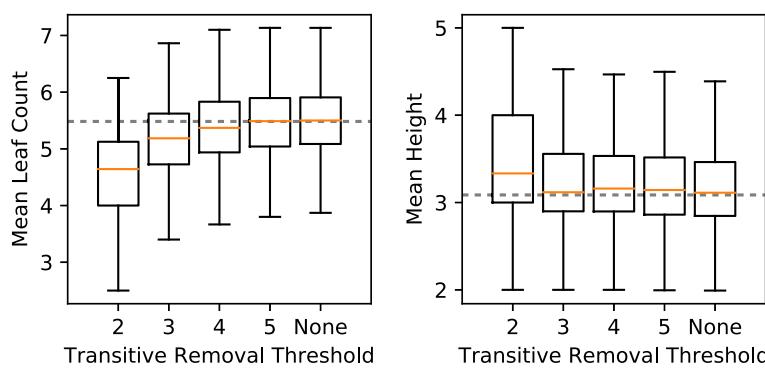
10000 trials of 10-node ancestry graphs. In particular, we only report results across trials in which solutions existed after pruning transitive edges (267, 1183, 1360, and 1409/10000 for 2-, 3-, 4-, and 5-PTR, respectively). We found that 2-PTR (the most extreme reduction) results in valid trees with 0.80 fewer leaves on average, while 3-PTR reduces the mean number of leaves by 0.33. In contrast, the mean height of solution trees only seems to be significantly affected by 2-PTR, which increased the mean height of trees by 0.25. For 3- and higher-order PTR, the mean height of trees was affected by less than 0.06. Single child fraction and mean subtree height both display similar trends to leaf count (see Additional file 1).

#### Real Data

We evaluated the strict and approximate ancestry graph methods using a chronic lymphocytic leukemia (CLL) dataset [29] and a clear cell renal cell carcinoma (ccRCC) dataset [30]. For the CLL data, we examined VAFs from  $100000 \times$  coverage targeted deep sequencing and from  $40 \times$  coverage whole genome sequencing (WGS). The ccRCC dataset used amplicon sequencing, with over  $400 \times$  average coverage [30]. An overview of the two datasets can be found in Table 1. For both datasets, we used the approximate and strict ancestry graph approaches to enumerate candidate clonal trees. When the standard sum condition yielded no solutions, we instead applied the relaxed sum



**Fig. 8** Effect of partial transitive reduction on the number and quality of solutions. ‘None’ represents the unpruned ancestry graph



**Fig. 9** Effect of partial transitive reduction on topological features of  $\mathcal{T}(G_F)$ . PTR disproportionately removes wide trees, causing the mean leaf count to decrease with more extreme pruning. The effect on tree height is less clear, although 2-PTR clearly removes shallow trees. The dashed line shows the mean value for underlying trees

condition (5), picking the smallest  $\varepsilon$  that resulted in at least one ancestry graph spanning tree. In the CLL data, we clustered mutations by observed frequency across all samples using  $k$ -means, and manually chose the number of clusters. For the ccRCC dataset, we instead used the clusters found by LICHeE, which uses mutation occurrence to enhance VAF-based clustering [19]. We note that we could have chosen to use a different method for mutation clustering (e.g. PyClone [40]) for this analysis. However, we choose the clusters produced by LICHeE as this allowed a direct comparison of our reconstructed trees with those reported in the LICHeE paper, which also analyzed this dataset. Furthermore, we note that PyClone is designed for more deeply sequenced mutations than we had available here. For both datasets, these clusters represent hypothesized clones in the tumor. To remove sites that may have undergone copy number aberrations, we ignored all mutations with a VAF above 0.5.

#### Rarity of Strict Solutions

Of the 11 patients we analyzed, only the  $100000 \times$  coverage targeted sequencing data for CLL006 and CLL077 admitted E-VAFFP solutions. 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 CLL077 data (with four clones) and two trees in the CLL006 data (with five clones). 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.

#### WGS and Targeted Sequencing Agreement in CLL Data

The trees identified from both WGS and deep sequencing data for all three CLL patients were topologically identical, regardless of whether we had inferred them using

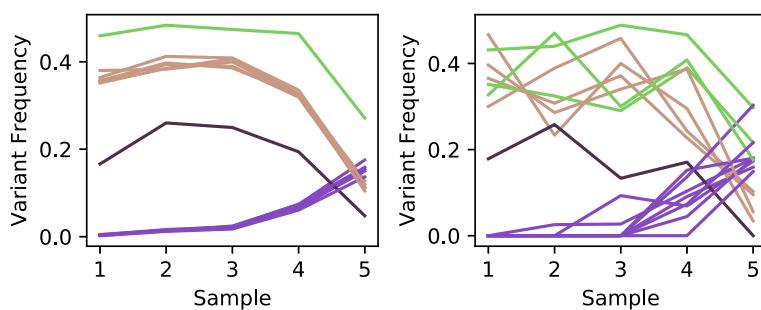
the strict or approximate methods. All minor labeling differences were the result of mutations that were filtered or simply absent in one of the datasets or that were differently clustered because of noise in the WGS data. Figure 10 displays the variant frequencies in patient CLL077, which showcases high WGS data noise. See Fig. 11 for the trees inferred from the deep and WGS CLL077 data.

Furthermore, setting aside any mutations filtered out because of possible copy number aberrations, the CLL trees we found agree entirely with the trees identified by two other inference methods, CITUP [18] and PhyloSub [20]. Moreover, our CLL077 tree displays the two major branches inferred by AncestryTree [17]. Most significantly, our CLL003 tree, which we generated with the

**Table 1** Dataset Summary

Patient	Samples	Mutations	Clones	$ \mathcal{T}(G_F) $
CLL003 (D)	5	15/20	4	0
CLL003 (W)	5	13/30	4	0
CLL006 (D)	5	5/10	5	2
CLL006 (W)	5	6/16	5	0
CLL077 (D)	5	12/16	4	1
CLL077 (W)	5	16/20	4	0
EV003	8	33/49	8	0
EV005	7	58/75	8	0
EV006	9	57/72	7	0
EV007	8	48/60	10	0
RK26	11	58/62	12	0
RMH002	5	44/51	8	0
RMH008	8	69/77	10	0

Mutation counts are displayed after/before filtering out mutations with VAF above 0.5. Mutations in CLL patients were clustered by VAF using  $k$ -means to identify clones, while we used clusters from [19] for the ccRCC patients. (D) indicates deep sequencing data and (W) indicates WGS data



**Fig. 10** Variant allele frequencies over five samples for patient CLL077. The left panel shows VAFs from targeted deep sequencing and the right panel shows VAFs from whole genome sequencing [29]. The colors of arcs indicate which mutations were clustered together using  $k$ -means

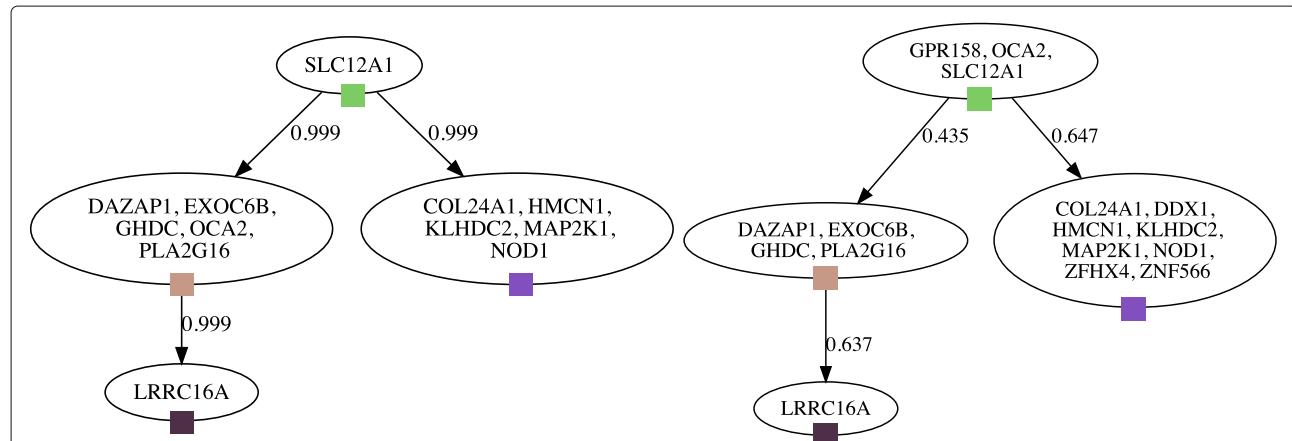
approximate ancestry graph and the relaxed sum condition, precisely matches the trees found by PhyloSub and CITUP.

It is worth emphasizing that when we relaxed the sum condition in the CLL006 and CLL077 WGS data, we recovered the same trees that had obeyed the sum condition in the deep sequencing data. Noise in the WGS data introduced sum condition violations of 0.101 and 0.048 in the CLL006 and CLL077 data, respectively. This is evidence that our sum condition relaxation, in concert with the approximate ancestry graph, allows us to successfully infer likely trees despite noise rendering the sum condition unsatisfiable. It is also worth mentioning that the CLL trees had few clusters, only 4 or 5. This places us within the regime we found in simulated data where the approximate method performs better than the strict method (see Fig. 6).

#### Approximate Solutions in ccRCC Data

As noted earlier, none of the eight ccRCC patients' data admitted strict E-VAFFP solutions. However, relaxing the sum condition and using the approximate ancestry graph nonetheless allows us to find candidate clonal trees. We selected the smallest sum condition relaxation  $\varepsilon$  that resulted in a single valid tree. In the case that several trees were found with the same sum condition relaxation, we picked the one with the highest weight in the approximate ancestry graph.

The trees we found in this way for patients EV003, EV005, EV006, EV007, RMH002, RMH008, and RK26 display strong agreement with those found by LICHeE [19]. See Table 2 for the tolerance  $\varepsilon$  needed to find these trees and for notes on their agreement with LICHeE (the trees themselves can be found in Additional file 2). We did not compare our results on RMH004 to those reported by



**Fig. 11** Clonal trees identified for patient CLL077. The left panel shows the tree derived from deep sequencing and the right panel shows the tree from WGS data. These trees were 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. 10. The movement of OCA2 to the root is due to different clustering as a result of noise (see Fig. 10). 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  $\varepsilon = 0.048$

**Table 2** ccRCC tree comparison with LICHeE

Patient	$\epsilon$	Notes
EV003	0.037	Exact match.
EV005	0.046	One node different.
EV006	0.078	Exact match.
EV007	0.042	Exact match.
RK26	0.028	One node different.
RMH002	0.086	Exact match.
RMH008	0.027	Exact match.

The second column shows the sum condition relaxation required. The third column notes the degree of similarity between our inferred tree and that of LICHeE

LICHeE due to an apparently malformed data file used to create those results. The sum condition overflows in the ccRCC data were relatively small (the largest  $\epsilon$  required was 0.086) but consistently present across patients. However, it is difficult to determine whether these overflows are due to legitimate ISA violations, such as the occurrence of convergent mutations, or simply due to noise in the measured VAFs.

## Discussion

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. Additionally, we found that the topology of the underlying tree has a strong impact on the quality and ease of phylogeny inference. While our analysis here focuses on the ancestry graph approach introduced in [17], the sum condition that underlies that method (which results from the ISA) is shared by a number of other approaches, such as [18–20] and others. Therefore, our conclusions here may likely apply to other methods—including new phylogenetic inference methods continuing to be developed. Thus, we claim that patterns of tumor evolution (linear, branching, etc.) should be more explicitly considered when developing and applying inference methods. This may become increasingly important as large-scale studies look across patients to identify common patterns of evolution within and across cancer types.

Meanwhile, we showed that the approximate ancestry graph method provides better trees than the strict approach when there are few clones and worse trees when there are many clones. This is likely connected to the relationship we found between tree rank and topology, with high-weight trees likely to be wide and shallow. Moreover, the approximate graph produces trees significantly skewed in this direction.

We also found several results that bear on the validity and applicability of the ISA. Despite the fact that our simulated data procedure adhered to the ISA, the majority of resulting VAF data broke the sum condition due to the noise added to the simulation. We found the same kind of violations in the ccRCC and 400 $\times$  coverage CLL data (with the notable exception of the ultra-high 100000 $\times$  coverage CLL data). However, we still found clonal trees in agreement with existing literature using only small sum condition overflows  $\epsilon$ , no higher than 0.09. This indicates that some violations of strict frequency assumptions are to be expected even if the ISA largely holds in practice. These findings encourage the exploration of methods that relax the ISA, although it is not clear that we should abandon it entirely.

We hope that our analysis here will be useful to those analyzing and interpreting real tumor phylogenies constructed using methods that rely on the ISA. 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. Future work should also address the impact of noise, tree topology, and other parameters on methods that relax the ISA or that consider mutations more complex than SNVs, such as copy number aberrations. Furthermore, our data simulation procedure did not include complex effects such as regional tumor heterogeneity or distinguish between driver and passenger mutations. These other factors could effect phylogeny inference and merit additional investigation. Finally, while we focused on methods applicable to multi-sample bulk sequencing data, the analysis of these issues with regard to 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 [28].

## Conclusions

We explored the inference of tumor evolutionary history from SNV frequency data obtained from multi-sample bulk sequencing using the ancestry graph method of [17]. This method is founded on the infinite sites assumption (ISA) and further simplifies the problem by ignoring copy number aberrations. Our contributions here include introduction and exploration of two methods of loosening the strict ISA assumption that allowed phylogenies to be found even in non-idealized data. We evaluated the effects of parameters, noise, and evolutionary tree topology on the existence and quality of candidate clonal trees. We found that these factors can significantly influence phylogeny inference, often in non-obvious ways (e.g. the counterintuitive effects of high coverage and high noise on solution quality). Methodically, we 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.

Tumor phylogeny inference has the potential to yield insight into how tumors develop and potentially to inform personalized cancer treatment [8, 9], which will become increasingly viable as sequencing methods continue to improve and become cheaper. As such, it is important not only to develop new and more accurate inference methods, but also to understand how those methods are impacted by the data they take as input. However, this issue has not been thoroughly explored in the existing literature [28]. Our work here addresses this oversight explicitly and has numerous potential implications. Our findings on the effects of controllable factors like sequencing coverage and number of sequenced samples can help inform practical decisions in real-world phylogeny inference experiments. For instance, we found that higher coverage does not necessarily improve the quality of inferred trees. Additionally, our results on uncontrollable factors like tumor evolution patterns and clone count can assist in interpreting trees reconstructed using ISA-based approaches such as [17–20]. Finally, our results provide strong motivation for additional work in exploring the performance of inference methods under different situations, since we showed that factors like tumor evolution pattern and noise levels exert significant pressure on inference results.

## Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12920-019-0626-0>.

**Additional file 1:** Single child fraction and mean subtree height plots. This PDF file contains corresponding plots for each of our topology results using these additional measures of tree topology.

**Additional file 2:** CLL and ccRCC trees. This PDF file contains the trees we reconstructed from real data, as well as the trees we obtained by running LICHeE.

**Additional file 3:** Joint effect of clone count and sum condition relaxation. This PDF file contains plots showing the interaction effects of the clone count  $n$  and the sum condition relaxation parameter  $\varepsilon$ .

## Abbreviations

(E-)VAFFP: (Enumeration) variant allele frequency factorization problem; A-D: ancestor-descendant; ccRCC: Clear cell renal cell carcinoma; CLL: Chronic lymphocytic leukemia; DAG: Directed acyclic graph; ISA: Infinite sites assumption; PTR: Partial transitive reduction; SNV: Single nucleotide variant; VAF: Variant allele frequency; WGS: Whole genome sequencing

## About this supplement

This article has been published as part of *BMC Medical Genomics Volume 12 Supplement 10, 2019: Selected articles from the IEEE BIBM International Conference on Bioinformatics & Biomedicine (BIBM) 2018: medical genomics*. The full

contents of the supplement are available online at <https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-12-supplement-10>.

## Authors' contributions

KT and LO conceived of the project. KT implemented the approach and carried out the analysis under the guidance of LO. KT and LO drafted and revised the manuscript. Both authors read and approved the final manuscript.

## Funding

This project and publication costs were funded by NSF CRIF award IIS-1657380. This project is also supported by Elledge, Eugster, and Class of '49 Fellowships from Carleton College.

## Availability of data and materials

The simulated data and the SNV read counts from the real datasets supporting the conclusions of this article are available here: <https://bitbucket.org/oesperlab/inference-effects-data>.

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Author details

<sup>1</sup>Department of Computer Science, Carleton College, 1 N College St, 55057 Northfield, MN, USA. <sup>2</sup>Department of Computer Science, Cornell University, 402 Gates Hall, 14853 Ithaca, NY, USA.

Published: 23 December 2019

## References

1. Nowell PC. The clonal evolution of tumor cell populations. *Science*. 1976;194(4260):23–8.
2. Davis A, Gao R, Navin N. Tumor evolution: Linear, branching, neutral or punctuated? *Biochim Biophys Acta Rev Cancer*. 2017;1867(2):151–61.
3. Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J, et al. A Big Bang model of human colorectal tumor growth. *Nat Genet*. 2015;47(3):209–16.
4. Sun R, Hu Z, Curtis C. Big Bang Tumor Growth and Clonal Evolution. *Cold Spring Harb Perspect Med*. 2018;8(a028381):1–13.
5. Gerstung M, Jolly C, Leshchiner I, Dentro SC, Gonzalez S, Mitchell TJ, et al. The evolutionary history of 2658 cancers. *bioRxiv*. 2017.
6. McGranahan N, Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*. 2017;168(4):613–28.
7. Dentro S, Leshchiner I, Haase K, Wintersinger J, Deshwar A, Tarabichi M, et al. Pervasive intra-tumour heterogeneity and subclonal selection across cancer types. Chicago: AACR; 2018.
8. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. *Brit J Cancer*. 2010;100(1):105–17.
9. Sun X, Yu Q. Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. *APS*. 2015;36(10):1219–77.
10. Yachida S, Jones S, Bozic I, Antal T, Leary R, Fu B, et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *467*. 2010;7319:1114–7.
11. Greaves M. Evolutionary determinants of cancer. *Cancer Discov*. 2015;5(8):806–20.
12. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*. 2017;02;168(4):613–28.
13. Al-Lazikani B, Banerji U, Workman P. Combinatorial drug therapy for cancer in the post-genomic era. *Nat Biotechnol*. 2012;30(7):679–92.
14. Zafar H, Tzen A, Navin N, Chen K, Nakhleh L. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biol*. 2017;18(1):178.
15. Jahn K, Kuipers J, Beerenwinkel N. Tree inference for single-cell data. *Genome Biol*. 2016;17:86.
16. El-Kebir M. SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics*. 2018;34(17):i671–9.

17. 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):i62–70.
18. Malikic S, McPherson AW, Donmez N, Sahinalp CS. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*. 2015;31(9):1349–56.
19. Popic V, Salari R, Hajirasouliha I, Kashef-Haghghi D, West RB, Batzoglou S. Fast and scalable inference of multi-sample cancer lineages. *Genome Biol*. 2015;16:91.
20. Jiao W, Vembu S, Deshwar AG, Stein L, Morris Q. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*. 2014;15:35.
21. Hajirasouliha I, Mahmoodi A, Raphael BJ. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*. 2014;30(12):i78–86.
22. Kuipers J, Jahn K, Raphael BJ, Beerenwinkel N. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Res*. 2017;27(11):1885–94.
23. Bonizzoni P, Ciccolella S, Della Vedova G, Soto M. Beyond perfect phylogeny: Multisample phylogeny reconstruction via ILP. In: Proc. 8th ACM Int. Conf. on Bioinformatics, Computational Biology, and Health Informatics. Boston: ACM; 2017. p. 1–10.
24. Jiang Y, Qiu Y, Minn AJ, Zhang NR. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *PNAS*. 2016;113(37):E5528–37.
25. 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 Biol*. 2015;16:35.
26. El-Kebir M, Satas G, Oesper L, Raphael BJ. Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. *Cell Syst*. 2016;3:43–53.
27. Qiao Y, Quinlan AR, Jazaeri AA, Verhaak RG, Wheeler DA, Marth GT. SubcloneSeeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization. *Genome Biol*. 2014;15(8):443.
28. Pradhan D, El-Kebir M. On the Non-uniqueness of Solutions to the Perfect Phylogeny Mixture Problem. In: RECOMB Int. Conf. on Comparative Genomics. Quebec: Springer; 2018. p. 277–93.
29. Schuh A, Becq J, Humphray S, Alexa A, Burns A, Clifford R, et al. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood*. 2012;120(20):4191–6.
30. Gerlinger M, Horswell S, Larkin J, Rowan AJ, Salm MP, Varela I, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet*. 2014;46(3):225–33.
31. Gabow HN, Myers EW. Finding all spanning trees of directed and undirected graphs. *SIAM J Comput*. 1978;7(3):280–7.
32. Tomlinson K, Oesper L. Examining Tumor Phylogeny Inference in Noisy Sequencing Data. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Madrid: IEEE; 2018. p. 36–43.
33. Cook J. Exact calculation of beta inequalities. Houston: University of Texas, M. D. Anderson Cancer Center; 2005.
34. Camerini PM, Fratta L, Maffioli F. The k best spanning arborescences of a network. *Networks*. 1980;10:91–110.
35. Aho AV, Garey MR, Ullman JD. The transitive reduction of a directed graph. *SIAM J Comput*. 1972;1(2):131–7.
36. Hsu HT. An algorithm for finding a minimal equivalent graph of a digraph. *J ACM*. 1975;22(1):11–6.
37. Devroye L. Non-Uniform Random Variate Generation. New York: Springer-Verlag; 1986. p. 568.
38. Govek K, Sikes C, Oesper L. A Consensus Approach to Infer Tumor Evolutionary Histories. In: Proc 2018 ACM Int. Conf. on Bioinformatics, Computational Biology, and Health Informatics. Washington D.C.: ACM; 2018. p. 63–72.
39. Robinson DF, Foulds LR. Comparison of phylogenetic trees. *Math Biosci*. 1981;53(1-2):131–47.
40. Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, et al. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods*. 2014;11(4):396.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](http://biomedcentral.com/submissions)

