A Consensus Approach to Infer Tumor Evolutionary Histories

Kiya Govek
Carleton College
Northfield, Minnesota
kiya.govek@gmail.com

Camden Sikes
Carleton College
Northfield, Minnesota
camdensikes@gmail.com

Layla Oesper
Carleton College
Northfield, Minnesota
loesper@carleton.edu

ABSTRACT

Inspired by recent efforts to model cancer evolution with phylogenetic trees, we consider the problem of finding a consensus tumor evolution tree from a set of conflicting input trees. In contrast to traditional phylogenetic trees, the tumor trees we consider contain features such as mutation labels on internal vertices (in addition to the leaves) and allow multiple mutations to label a single vertex.

We describe several distance measures between these tumor trees and present an algorithm to solve the consensus problem called GraPhyC. Our approach uses a weighted directed graph where vertices are sets of mutations and edges are weighted using a function that depends on the number of times a parental relationship is observed between their constituent mutations in the set of input trees. We find a minimum weight spanning arborescence in this graph and prove that the resulting tree minimizes the total distance to all input trees for one of our presented distance measures.

We evaluate our GraPhyC method using both simulated and real data. On simulated data we show that our method outperforms a baseline method at finding an appropriate representative tree. Using a set of tumor trees derived from both whole-genome and deep sequencing data from a Chronic Lymphocytic Leukemia patient we find that our approach identifies a tree not included in the set of input trees, but that contains characteristics supported by other reported evolutionary reconstructions of this tumor.

CCS CONCEPTS

• **Applied computing → Computational biology**;

KEYWORDS

cancer; phylogeny; evolutionary history; consensus

ACM Reference Format:
Kiya Govek, Camden Sikes, and Layla Oesper. 2018. A Consensus Approach to Infer Tumor Evolutionary Histories. In *ACM-BCB'18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, August 29-September 1, 2018, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3233547.3233584

1 INTRODUCTION

A tumor is the result of an evolutionary process where somatic mutations – those that occur during the lifetime of the individual – accumulate and lead to the growth of a tumor [19]. The *evolutionary history* of a tumor describes the order in which these mutations appeared and includes information such as the ancestral relationships between mutations. Thus, the evolutionary history of a tumor is often described using a rooted tree [26] where vertices represent different populations of cells, each with a unique complement of somatic mutations, that have existed during the tumor's evolution and the root represents the founding cell population.

A better understanding of the evolutionary history underlying tumor growth may provide important insights into how and why tumors develop [29]. Thus, in recent years there has been tremendous advances in methods that aim to infer such a tree describing a tumor's evolutionary history from DNA sequencing data. For instance, one of the earlier methods, TRaP [28], uses bulk sequencing data from a single tumor sample to infer this history. A number of other methods that utilize multi-sample bulk sequencing data such as [5, 7, 8, 12, 13, 15, 20, 24, 33] and many others have been developed. As single-cell sequencing methods have matured, a number of methods that utilize this type of data have also appeared [11, 22, 32]. There are also recent hybrid methods that utilize both bulk sequencing data and single cell data [14].

While the recent algorithmic progress has led to improved inference of tumor evolutionary histories, there are a number of situations where more than one evolutionary history for a single tumor may be inferred. The first such situation is when a single method returns multiple possible trees. This may occur for stochastic methods, such as PhyloSub [13] or PhyloWGS [5] that use a Bayesian approach and report a collection of the most likely trees sampled. Multiple plausible trees are also possible for methods that use a discrete optimization criterion, such as AncesTree [7] which returns the largest tree that adheres to a specific criterion, and multiple such trees satisfy the optimization criteria. Some algorithms, such as SPRUCE [8] also explicitly enumerate a set of possible tree reconstructions. The second situation where different reconstructions of a single tumor exist is when different algorithms are applied to the same dataset. It's not uncommon for different methods to produce similar, but not identical results. These disparities may result from different underlying assumptions made by each approach, or the application of alternative algorithmic approaches. The last situation is when different types of data exists for a single tumor sample [10, 25], such as bulk sequencing, targeted deep sequencing and single cell data. Computational approaches applied to such different datasets may yield a set of distinct possible evolutionary histories. While some hybrid methods are being developed that incorporate several data types [14, 23], there are only a few such methods and they are designed for specific data types.

Given a set of disparate tumor evolutionary histories, a natural question is whether information across these histories can be combined to infer a better evolutionary history? This type of *consensus* approach has been useful when applied to traditional phylogenetic

ACM-BCB'18, August 29-September 1, 2018, Washington, DC, USA

ACM ISBN 978-1-4503-5794-4/18/08.
https://doi.org/10.1145/3233547.3233584

trees that are used to show the evolutionary relationships between different species. The leaves of a phylogenetic tree represent the extant species and the internal vertices represent the most recent common ancestor of all its descendants. The topologies of such trees are typically inferred by starting with the set of extant species and inferring the set of internal nodes in the tree.

A number of consensus methods exist for finding a consensus, or representative, phylogenetic tree from a set of conflicting trees. One of the simplest such methods is *strict consensus* [21] which creates a tree that contains all the same groupings of species that occur in all the input trees. The *majority-rule consensus* tree constructs a tree that contains groupings that exist in a majority of the input trees [16]. The *Adams consensus* looks at which species are often clustered together across the set of input trees [1]. Many other such consensus algorithms exist. For a review see [4]. These methods typically rely on the input trees being traditional phylogenetic trees where a set of species label just the leaves of the tree. This is not the case for tumors where ancestral populations (internal nodes) still exist at the present time. Furthermore, in tumor evolutionary histories the set of leaves (species) can be different between different input trees. Thus, novel algorithmic techniques are needed to perform consensus on a set of tumor evolutionary histories.

We formalize the problem of finding a consensus tumor evolutionary history from a set of possible tumor histories as the m-Tumor Tree Consensus Problem (m-TTCP). The input to the problem is a set of potential tumor evolutionary histories, represented as a specific type of rooted tree, along with a distance measure between tumor trees. The problem aims to find a rooted tree that minimizes the total distance from the consensus tree to the set of input trees. We also present and analyze variations of this problem for four different distance measures which allow us to capture different aspects of tumor histories.

We propose a graph-based algorithm, GraPhyC. We prove that our approach optimally solves one variant of the m-TTCP when given a specific distance measure. On simulated data we show that our GraPhyC method outperforms a baseline method for all four distance measures, and at recovering the true underlying tree used to create the simulated data, even when that tree is not part of the set of input trees. This indicates that our approach may be able to identify a more accurate tumor evolutionary history from a set of noisy input trees. We apply GraPhyC to real DNA sequencing data and show that it is able to recover a consensus tree not included in the input data, but is supported by other existing reconstructions of this tumor's history.

2 METHODS

2.1 Phylogenetic Trees and Tumor Trees

The following definition may be used to describe traditional phylogenetic trees where the leaves of the tree represent n extant species.

Definition 2.1. An n-**phylogenetic tree** is a rooted tree T with n leaves. Each leaf has exactly one unique label from the set $\{1, \ldots, n\}$.

We note that any two n-phylogenetic trees T_1 and T_2 will have the same set of leaves. Each leaf may be further described as a set of mutations, and those sets of mutations remain the same between any two such trees.

A tumor is the result of an evolutionary process, and therefore its history may also be represented using a similar tree structure where vertices represent different tumor populations that existed during the history of the tumor and edges indicate direct ancestral relationships. Each tumor population is distinguished by a unique complement of somatic mutations that exist in the population. Unlike a phylogenetic tree, the set of extant species is not known a priori. Instead, this tree is constructed using information about the set of mutations that arose during the history of the tumor. We will make two assumptions about how a tumor evolves. The first assumption is the *infinite sites assumption*, which states that no mutation occurs more than once during the history of a tumor. While some recent methods allow limited violations to this rule [3, 8] this has been a common assumption for many methods that infer tumor evolution such as [5, 7, 12, 15, 24, 28]. Therefore, we maintain this assumption for this work. The second assumption we make is that the tumor is the result of monoclonal evolution, that the tumor can be traced back to a single founder population. Thus, the history of a tumor can be described using the following definition.

Definition 2.2. An m-**tumor tree** is a rooted tree T where: (i) each vertex in the tree is labeled by one or more mutations from the mutation set $[m] = \{1, \ldots, m\}$; (ii) every mutation labels some vertex; and (iii) no mutation appears more than once.

The mutations labeling each vertex in an m-tumor tree indicate the mutations that first appeared in the corresponding tumor population. We note that one could make an alternative, but ultimately equivalent, interpretation where the mutations label the edge incoming to the vertex instead. Figure 1 shows an example of a 4-phylogenetic tree and a 4-tumor tree.

Each vertex v in an m-tumor tree T represents a population of tumor cells that existed at some point during the tumor's evolution. The particular set of mutations that existed in the population represented by vertex v are the mutations that label all vertices on the unique path from the root to v. For any vertex v, we denote this set of mutations as $clone(v)$. We can also define the **set of clones** in $T = (V, E)$ as $clone(T) = \{clone(v) | v \in V\}$ (see Figure 1 for an example). Thus, unlike n-phylogenetic trees, any two m-tumor trees T_1 and T_2 may have a different set of clones (species) labeling their leaves, in addition to different sets of clones across the entire tree (i.e. $clone(T_1) \neq clone(T_2)$).

We also note that a more general model that includes polyclonal evolution could be achieved by adding a default root vertex to the m-tumor tree that is unlabeled and represents the germline. For simplicity, we will restrict our attention to the monoclonal case for the remainder of this work.

Finally, let $\mathcal{T}_m$ be the set of all m-tumor trees. We note that any $T \in \mathcal{T}_m$ defines a partition (clustering) of the mutation set $[m]$ by considering the labels assigned to the vertices of the tree. Let $P(m)$ represent an arbitrary partition of the mutation set $[m]$. We define $\mathcal{T}_{P(m)} \subseteq \mathcal{T}_m$ to be the subset of m-tumor trees that induce the partition $P(m)$. A relevant special case is when we consider the partition $\mathbb{1}(m) = 1|2|3|\ldots|m$ where each mutation is partitioned into its own cluster. Therefore, $\mathcal{T}_{\mathbb{1}(m)} \subseteq \mathcal{T}_m$ is the set of m-tumor trees where each node in the tree is labeled by exactly one mutation. We will refer to $\mathcal{T}_{\mathbb{1}(m)}$ as the set of **single-label** m-**tumor trees**.

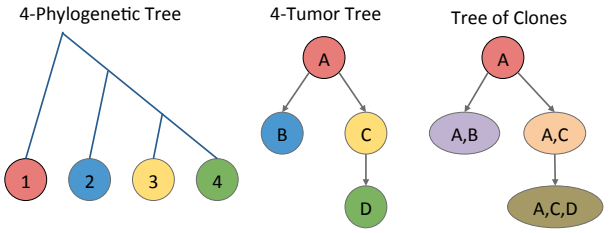

Figure 1: (left) A 4-phylogenetic tree on with species labeled 1, 2, 3 and 4. (middle) A 4-tumor tree with mutations labeled A, B, C and D. (right) A tree of the clones in the 4-tumor tree in the middle of the figure.

2.2 Consensus Trees

Given a set of discordant m-tumor trees representing possible tumor evolutionary histories, our goal is to find a **consensus** tree that is better able to describe the evolutionary history of the underlying tumor. Specifically, we define the following problem.

The m-Tumor Tree Consensus Problem (m-TTCP): Given a set $\mathcal{S} = \{T_1, T_2, \ldots, T_n\} \subseteq \mathcal{T}_m$ of m-tumor trees, and a distance measure $dist(\cdot, \cdot)$ between m-tumor trees, find a consensus tree T^* such that

$$T^* = \underset{T \in \mathcal{T}_m}{\operatorname{argmin}} \sum_{i=1}^{n} dist(T, T_i).$$

Similar consensus problems have been studied in the realm of n-phylogenetic trees [2]. We will explicitly consider several variations on this problem in later sections.

2.3 Distance Between Trees

The m-TTCP relies upon having a distance measure between m-tumor trees. A number of different distance measures between n-phylogenetic trees have been defined including nearest-neighbor interchange (NNI) [30], quartets distance [9], path distance [27, 31], and others. We also note that [17] defines a distance between tumor evolutionary trees, but they do so in the context of comparing the structure of clonal evolution across patients, rather than comparing a set of trees on a fixed set of mutations. Here we present several different distance measures between m-tumor trees (Figure 2).

2.3.1 Path Distance. For any m-tumor tree T and pair of mutations $i, j \in [m]$, let $path(T, i, j)$ be the length of the unique path in T from the vertex with label i to the vertex with label j. Note, this path ignores any directionality associated with edges in the graph. This type of distance measure has been previously to describe distances between n-phylogenetic trees. We define the **path distance** $PD(T_1, T_2)$ between two m-tumor trees, T_1 and T_2 as the sum of the absolute value of the difference between the path lengths in T_1 and T_2 for all pairs of mutations (Figure 2(a)). Formally,

$$PD(T_1, T_2) = \sum_{i<j} |path(T_1, i, j) - path(T_2, i, j)|. \quad (1)$$

As presented here, path distance is a version of the distance with the same name defined on n-phylogenetic trees (where all labels occur on the leaves of the tree). We include this measure to show

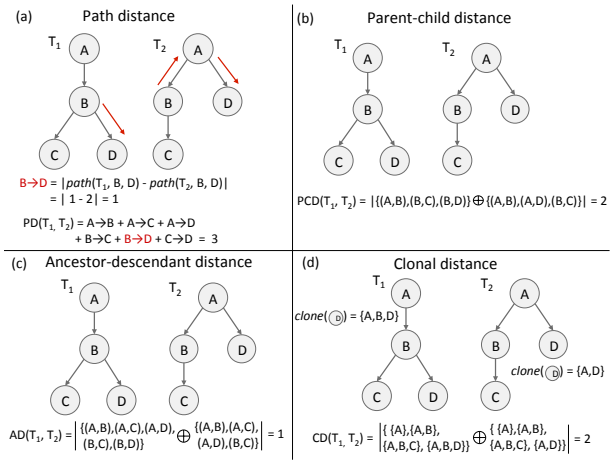

Figure 2: Examples of different distance measures for two m-clonal trees. (a) Path distance. (b) Parent-child distance. (c) Ancestor-descendant distance. (d) Clonal distance.

that measures defined on n-phylogenetic trees can be used in this situation. However, this distance has a number of limitations when applied to m-tumor trees. For example, since path distance only takes into account distances between mutations in the m-tumor tree, it loses information about the ancestral relationships between these mutations. For example, suppose A is parental to mutation B in one tree, but mutation B is parental to mutation A in another tree. The path distance between these mutations is both 1 in each tree, and therefore these mutations do not contribute to the total path distance between the trees. Because of this and other limitations, we define three additional distance measures on m-clonal trees that allow us to more directly consider the structure of these trees.

2.3.2 Parent-Child Distance. Given an m-Tumor Tree T and two mutations i and j, we say that mutation i is **parental** to mutation j in T if: (i) there is an edge connecting the vertex labeled with mutation i to the vertex labeled with mutation j in T, and (ii) the vertex labeled with mutation i lies on the path from the root to the vertex labeled with j in T. If mutation i is parental to mutation j, then we say that mutation j is a **child** of mutation i. Given an m-Tumor Tree T we define $\phi_{PC}(T) = \{(i, j) | i \text{ is parental to } j \text{ in } T\}$, that is $\phi_{PC}(T)$ is the set of all ordered pairs of mutations (i, j) such that i is parental to j in T. Finally, we can use these definitions to define the **parent-child distance** $PCD(T_1, T_2)$ between two m-Tumor Trees T_1 and T_2 to be the number of parent-child relationships that exist in one tree but not the other (Figure 2(b)). Formally:

$$PCD(T_1, T_2) = |\phi_{PC}(T_1) \oplus \phi_{PC}(T_2)|. \quad (2)$$

Observation 2.1. *$PCD(\cdot, \cdot)$ is a distance metric when we restrict the domain to $\mathcal{T}_{\mathbb{1}}(m) \times \mathcal{T}_{\mathbb{1}}(m)$.*

Proof. By definition, parent-child distance always returns non-negative values. Therefore, we need to show the following three things in order to prove that it is a valid distance metric.
Identity of indiscernibles: Suppose that $T_1, T_2 \in \mathcal{T}_{\mathbb{1}(m)}$ such that $PCD(T_1, T_2) = 0$. This implies that T_1 and T_2 have the exact same set of parent-child relationships. Since parent-child relationships

can be used to uniquely reconstruct a tree, this means that $T_1 = T_2$. Similarly, if $T_1 = T_2$, then by definition $PCD(T_1, T_2) = 0$.
Symmetry: Suppose $T_1, T_2 \in \mathcal{T}_{\mathbb{1}(m)}$. Then, $PCD(T_1, T_2) = |\phi_{PC}(T_1) \oplus \phi_{PC}(T_2)| = |\phi_{PC}(T_2) \oplus \phi_{PC}(T_1)| = PCD(T_2, T_1)$.
Triangle Inequality: Suppose $T_1, T_2, T_3 \in \mathcal{T}_{\mathbb{1}(m)}$. Let $(i, j) \in \phi_{PC}(T_1) \oplus \phi_{PC}(T_3)$. Without loss of generality we may assume that $(i, j) \in \phi_{PC}(T_1)$ and $(i, j) \notin \phi_{PC}(T_3)$. We now consider the two possible scenarios for (i, j).

Case 1: Assume $(i, j) \in \phi_{PC}(T_2)$. In this case, then $(i, j) \in \phi_{PC}(T_2) \oplus \phi_{PC}(T_3)$, and hence contributes one to $PC(T_2, T_3)$.

Case 2: Assume $(i, j) \notin \phi_{PC}(T_2)$. In this case, then $(i, j) \in \phi_{PC}(T_1) \oplus \phi_{PC}(T_2)$, and hence contributes one to $PC(T_1, T_2)$.

Thus, for every $(i, j) \in \phi_{PC}(T_1) \oplus \phi_{PC}(T_3)$ then either $(i, j) \in \phi_{PC}(T_2) \oplus \phi_{PC}(T_3)$ or $(i, j) \in \phi_{PC}(T_1) \oplus \phi_{PC}(T_2)$. Hence,

$$\begin{aligned} PCD(T_1, T_3) &= |\phi_{PC}(T_1) \oplus \phi_{PC}(T_3)| \\ &\leq |\phi_{PC}(T_1) \oplus \phi_{PC}(T_2)| + |\phi_{PC}(T_2) \oplus \phi_{PC}(T_3)| \\ &= PCD(T_1, T_2) + PCD(T_2, T_3) \end{aligned}$$

and therefore the triangle inequality holds.

□

A distance function d that fulfills the same properties as a distance metric except that it relaxes the definition to allow $d(x, y) = 0$ when $x \neq y$ is a **distance psuedometric**.

Observation 2.2. *$PCD(\cdot, \cdot)$ is a distance pseudometric.*

Proof. This proof is nearly identical to the proof of Observation 2.1 with the following modification.
Relaxed identity of indiscernibles: Suppose $T \in \mathcal{T}_m$, then by definition $PCD(T, T) = 0$. Note, the reverse is not true. Suppose T_1 is a tree with a root vertex labeled with mutation m_0 and three children labeled each labeled with one mutations, m_1, m_2 and m_3 respectively, and T_2 is a tree with a root label with mutation m_0 and one child, labeled with three mutations m_1, m_2 and m_3. In this instance $T_1 \neq T_2$, but certainly $PCD(T_1, T_2) = 0$.

□

2.3.3 Ancestor-Descendant Distance. In the previous section we defined a distance between m-tumor trees that looks at direct relationships between vertices in the tree. Here we look at a more generalized version that considers longer range relationships between vertices.

Given an m-Tumor Tree T and two mutations i and j, we say that mutation i is **ancestral** to mutation j in T if the vertex labeled with mutation i lies on the path from the root of T to the vertex labeled with j. Note that under this definition i is considered ancestral to j (and vice versa) if both mutations label the same vertex in T. We make this choice since it is unclear which mutation is ancestral to another in this instance (we could also have chosen to label neither mutation as ancestral), and this provides flexibility if these mutations label different vertices in a different tree. If mutation i is ancestral to mutation j, then we say that mutation j is a **descendant** of mutation i. Given an m-Tumor Tree T we define $\phi_{AD}(T) = \{(i, j) | i \text{ is ancestral to } j \text{ in } T\}$, that is $\phi_{AD}(T)$ is the set of all ordered pairs of mutations (i, j) such that i is ancestral to j. Finally, we can use these definitions to define the **ancestor-descendant distance** $AD(T_1, T_2)$ between two m-Tumor Trees T_1 and T_2 to be the number of ancestor-descendant relationships that exist in one tree but not the other (Figure 2(c)). Formally,

$$AD(T_1, T_2) = |\phi_{AD}(T_1) \oplus \phi_{AD}(T_2)|. \tag{3}$$

Observation 2.3. *$AD(\cdot, \cdot)$ is a distance metric when we restrict the domain to $\mathcal{T}_{\mathbb{1}}(m) \times \mathcal{T}_{\mathbb{1}}(m)$.*

Observation 2.4. *$AD(\cdot, \cdot)$ is a distance pseudometric.*

The proofs of Observation 2.3 and Observation 2.4 are nearly identical the corresponding statements about the parent-child distance and are therefore omitted for space.

2.3.4 Clonal Distance. Underlying each vertex v in an m-tumor tree is a set of mutations $clone(v)$ representing all the mutations labeling all vertices from the root to v. This set of mutations, represent a collection of mutations that are predicted to exist (or have existed) in a collection of cells in the tumor. We define a **clonal distance** measure $CD(T_1, T_2)$ that allows us to compare the sets of clones underlying two m-tumor trees, by counting the number of clones that are unique to either T_1 or T_2 (Figure 2(d)). Formally,

$$CD(T_1, T_2) = |clone(T_1) \oplus clone(T_2)|. \tag{4}$$

Clonal distance more strongly penalizes inconsistencies closer to the root of the tree than at the leaves. For instance, consider a linear tree where the child of the root has two mutations and a nearly identical tree except that the child of the root has been split into two children. These trees will have near maximal clonal distance, despite being similar appearing trees. However, the clonal distance will decrease if the branch occurs farther down the tree.

2.4 Consensus with Labels on Internal Vertices

There are two main difference between n-phylogenetic trees and m-tumor trees: (1) An m-tumor tree has labels on all vertices, including internal vertices; and (2) A vertex in an m-tumor tree may have multiple labels. In this section we consider a version of the m-TCCP where we restrict both our input and output and only consider single-label m-tumors trees from the set $\mathcal{T}_{\mathbb{1}(m)}$. This allows us to address just the first difference where mutations label internal vertices, without yet considering multiple labels on a vertex (a problem we will consider later). Specifically, we define the following problem.

The Single Label m-Tumor Tree Consensus Problem (SL-m-TTCP): Given a set $\mathcal{S} = \{T_1, T_2, \ldots, T_n\} \subseteq \mathcal{T}_{\mathbb{1}(m)}$ of m-tumor trees, and a distance measure $dist(\cdot, \cdot)$ between m-tumor trees, find a consensus tree T^* such that

$$T^* = \operatorname*{argmin}_{T \in \mathcal{T}_{\mathbb{1}(m)}} \sum_{i=1}^{n} dist(T, T_i).$$

2.4.1 Parent Child Graph. To solve the SL-m-TTCP we will first need to see how to build a particular graph. Given a set $\mathcal{S} = \{T_1, T_2, \ldots, T_n\} \subseteq \mathcal{T}_{\mathbb{1}(m)}$ of single-label m-tumor trees, we can build a graph that represents all parent-child relationships that exist in $\mathcal{S}$. Specifically, the **parent-child-graph** $G = (V, E)$ is the weighted directed graph with vertices $\{1, 2, \ldots, m\}$ and directed edges $E = \{(u, v) | (u, v) \in \cup_{i=1}^{n} \phi_{PC}(T_i)\}$. Each edge (u, v)

is weighted with $|\mathcal{S}| - 2 \cdot count(u, v)$ where $count(u, v)$ is number of trees $T \in \mathcal{S}$ where vertex u is a parent of vertex v. Note, for mutations a and b the value $|\mathcal{S}| - 2 \cdot count(\mathcal{S}, a, b))$ is negative if a is parental to b in at least half of the trees in $\mathcal{S}$ and positive otherwise. We also note that the relative order of weighted edges would be the same if we had weighted edges with $-count(u, v)$ instead.

A **spanning arborescence** of the parent-child graph G is a subgraph $G' = (V, E')$ with $E' \subseteq E$ such that there exists a unique path from some root vertex v_r to every vertex $v \in V$. In particular, we note that every $T \in \mathcal{S}$ defines some spanning arborescence in G, but that other novel spanning arborescences, not corresponding to some $T \in \mathcal{S}$ may exist within G.

Theorem 2.3. *Given a set $\mathcal{S} = \{T_1, T_2, \ldots, T_n\} \subseteq \mathcal{T}_{\mathbb{1}(m)}$ of single-label m tumor trees, the minimum weight spanning arborescence of the corresponding parent-child graph G defines a tree T^* that is a solution to the SL-m-TTCP when the distance measure is parent-child distance.*

We omit a proof of 2.3 here as it follows directly from a more general theorem we will state in the following section.

2.4.2 The SL-GraPhyC Algorithm. Algorithm 1 shows an approach we call the **S**ingle-**L**abel **Gra**ph-based **Ph**ylogenetic **C**onsensus (SL-GraPhyC) which allows us to solve the SL-m-TTCP in one instance.

Algorithm 1: SL-GraPhyC Algorithm

Input: $\mathcal{S} = \{T_1, T_2, \ldots, T_n\} \subseteq \mathcal{T}_{\mathbb{1}(m)}$
Output: A consensus tree $T^* \in \mathcal{T}_{\mathbb{1}(m)}$

```
Build the parent-child graph G = (V, E) for the set S.
T* ⟵ ∅, best ⟵ 0
foreach v ∈ V do
    Find T, a minimum spanning arborescence of G rooted at v.
    w ⟵ total weight of T.
    if w < best then
        T* ⟵ T, best ⟵ w
    end
end
return T*
```

We first note that efficient algorithms, such as Edmonds/Chu-Liu [6], exist for finding the minimum spanning arborescence of a directed graph $G = (V, E)$ given a a root vertex $r \in V$. Second, we note that Theorem 2.3 tells us that the consensus tree output by SL-GraPhyC yields a solution to the SL-m-TCCP when we consider a distance measure of parent-child distance. In section 3 we show that even when a different distance measure is used, the output of our approach outperforms a baseline consensus approach.

2.5 Consensus with Clustered Mutations

In this section we build on our approach from the previous section to also incorporate the second main difference between m-tumor trees and n-phylogenetic trees - a vertex may have multiple labels. One challenge with trying to solve the m-TTCP in this instance is that different m-tumor trees $T_1, T_2 \in \mathcal{T}_m$ may have different clustering of mutations over the vertices in the tree, and may therefore even have a different number of vertices. Thus, we consider the following variation on the m-TTCP, where we restrict our output m-tumor tree to induce a specific partition on the set of mutations $\{1, 2, \ldots, m\}$.

The Clustered Mutation m-Tumor Tree Consensus Problem (CM-m-TTCP): Given a set $\mathcal{S} = \{T_1, T_2, \ldots, T_n\} \subseteq \mathcal{T}_{(m)}$ of m-tumor trees, a partition over mutations $P(m)$, and a distance measure $dist(\cdot, \cdot)$ between m-tumor trees, find a consensus tree T^* such that

$$T^* = \underset{T \in \mathcal{T}_{P(m)}}{\operatorname{argmin}} \sum_{i=1}^{n} dist(T, T_i).$$

2.5.1 Identifying a Partition. Every tree $T \in \mathcal{T}_m$ defines a partition (alternatively, a clustering) of the mutation set $[m]$ by considering the labels assigned to the vertices in T. Therefore a set of input trees $\mathcal{S} = \{T_1, T_2, \ldots, T_n\} \subseteq \mathcal{T}_m$ defines a collection of n partitions of the mutation set $[m]$, one for each input tree. Our goal is to use these partitions to identify a representative partition of $[m]$ that we will use to restrict our attention to while constructing a consensus tree. We use the approach of *consensus clustering*, an established problem with a number of proposed solutions [18]. Given a mutation set $[m]$ and an arbitrary partition over those mutations $P(m)$ we define the a function $cluster(a, b, P(m))$ which is 1 if $a, b \in [m]$ are clustered together in $P(m)$ and -1 if they are not clustered together. We note that given a set of partitions $\mathcal{P} = \{P_1, P_2, \ldots, P_n\}$ over $[m]$, the value $\sum_{i=1}^{n} cluster(a, b, P_i)$ will be positive if a and b and clustered together in at least half of the partitions in $\mathcal{P}$ and will be negative otherwise.

Algorithm 2 below is a greedy approach to consensus clustering which is a variation on hierarchical clustering that uses the previous observation to decide when to stop merging clusters.

Algorithm 2: Greedy Consensus Clustering

Input: Partitions $\mathcal{P} = \{P_1, P_2, \ldots, P_n\}$ over a set of mutations $\{1, 2, \ldots, m\}$.
Output: A consensus partition P over the mutation set.

```
P = {{1}, {2}, ..., {m}}
while |P| > 1 do
    foreach A, B ∈ P where A ≠ B do
        d_AB = Σ_{a∈A} Σ_{b∈B} Σ_{i=1}^{n} cluster(a, b, P_i)
    end
    A*, B* = argmax_{A,B∈P} d_AB
    if d_{A*B*} > 0 then
        P = P − A* − B* + (A* ∪ B*)
    else
        return P
    end
end
return P
```

2.5.2 Parent Child Graph. Given a set $\mathcal{S} = \{T_1, T_2, \ldots, T_n\} \subseteq \mathcal{T}_m$ of m-phylogenetic trees we can build a modified version of our **parent-child graph** that represents parent-child relationships in $\mathcal{S}$. Since we now allows multiple mutations to label any vertex in any of our input trees, we need to make a few modifications to

the approach presented in section 2.4.1 where we restricted our consideration to single-label m tumor trees. Previously the set of vertices in the graph were just the set of mutations. But since we are allowing multiple mutations to be clustered together, we also need as input a partition P over the mutation set $[m]$. We construct a directed graph $G = (V, E)$ where $|V| = |P|$ and each vertex is labeled with a set mutations from P. Directed edges are added for every parent-child relationship existing in some tree in $\mathcal{S}$.

Previously we weighted each directed edge (u, v) with $|\mathcal{S}| - 2 \cdot count(u, v)$ where $count(u, v)$ is number of trees $T \in \mathcal{S}$ where vertex u is a parent of vertex v. Since multiple mutations may label each vertex, we need to consider a weighting scheme that allows us to account for vertices with multiple mutations, without unnecessarily giving these edges more weight. Let $A, B \in P$ be distinct sets of mutations in P with corresponding vertices in G of v_A and v_B. We weight the edge between v_A and v_B as follows: $w(v_A, v_B) = \sum_{a \in A} \sum_{b \in B} (|\mathcal{S}| - 2 \cdot count(\mathcal{S}, a, b))$. See Figure 3 for an example of a parent-child graph.

Theorem 2.4. *Given a set $\mathcal{S} = \{T_1, T_2, \ldots, T_n\} \subseteq \mathcal{T}_m$ of m tumor trees, and a partition of the mutations set $P(m)$, the minimum weight spanning arborescence of the corresponding parent-child graph G defines a tree T^* that is a solution to the CM-m-TTCP when the distance measure used is parent-child distance.*

Proof Sketch. Due to space constraints, we provide here only a sketch of the main ideas underlying the proof.

We start by creating a completely disconnected graph $G = (V, E)$ where V is the set of partitions in $P(m)$ and $E = \emptyset$. We then can compute the total distance from this graph to all input trees as $\sum_{i=1}^{n} PCD(G, T_i)$. We then show that adding any edge between cluster A and B in G will result in a *change of total distance* equal to $\Delta PCD(G, T_i) = \sum_{a \in A} \sum_{b \in B} (|S| - 2 \cdot count(\mathcal{S}, a, b))$. Since this change does not depend on the structure of G at the time, each edge can be added greedily as long as it does not create a cycle in G. This process stops once we get a tree T^*, a minimum spanning arborescence of the parent-child graph. This tree must also be a tree that minimizes the total distance $\sum_{i=1}^{n} PCD(T^*, T_i)$.

2.5.3 The GraPhyC Algorithm. Algorithm 3 is our general Graph-based Phylogenetic Consensus (GraPhyC) approach that allows us to find a consensus tree given a set of input m-phylogenetic trees $\mathcal{S} = \{T_1, T_2, \ldots, T_n\}$. Figure 3 shows an overview of the method, including both clustering of mutations and construction of the parent-child graph.

Theorem 2.4 tells us that the consensus tree output by Algorithm 3 yields a solution to the CM-m-TCCP when we use a distance measure of parent-child distance. We also note that any algorithm for partitioning mutations can be used in place of our greedy consensus clustering algorithm. If all input trees share the same root or a priori information is known about the root, the approach can be modified to only find potential trees with that root.

2.5.4 Multiple Solutions. We note that multiple distinct maximum spanning arborescences may exist in a parent-child graph $G = (V, E)$. Our approach presented in Algorithm 3 will only return one such solution. We modified the Edmonds/Chu-Liu algorithm to return the set of minimum spanning arborescences, although this has a significant impact on the running time of the approach. Furthermore, a distance measure $dist : \mathcal{T}_m \times \mathcal{T}_m \longrightarrow \mathbb{R}^{\geq 0}$ may be supplied as an optional parameter which can be used to rank the set of minimum spanning arborescences returned by the algorithm.

Algorithm 3: GraPhyC Algorithm

Input: $\mathcal{S} = \{T_1, T_2, \ldots, T_n\} \subseteq \mathcal{T}_m$
Output: A consensus tree $T^* \in \mathcal{T}_m$

$\mathcal{P} \longleftarrow \{\}$
foreach $T_i \in \mathcal{S}$ **do**
  $\mathcal{P} = \mathcal{P} \cup$ partition over $[m]$ induced by T_i
end
$P \longleftarrow$ Greedy Consensus Clustering($\mathcal{P}$)
Build the parent-child graph $G = (V, E)$ for the set $\mathcal{S}$ and partition P.
$T^* \longleftarrow \emptyset$, $best \longleftarrow 0$
foreach $v \in V$ **do**
  Find T, a minimum spanning arborescence of G rooted at v.
  $w \longleftarrow$ total weight of T.
  if $w < best$ **then**
    $T^* \longleftarrow T, best \longleftarrow w$
  end
end
return T^*

3 RESULTS

We implemented GraPhyC in Java using our own modified implementation of Edmond's/Chu-Liu algorithm to search for all minimum spanning trees in the parent-child graph. We analyze our GraPhyC method on both simulated data and real sequencing data.

3.1 Results on Simulated Data

Using simulated data we first evaluate how well our approach actually solves the variations of the m-TTCP for our four distance measures. We then analyze how well our inferred consensus tree reflects the true underlying tree used to create the set of input trees.

3.1.1 Simulated Data Creation. We created each set of input trees by first creating a random m-tumor tree by iteratively adding mutations with random parents from the existing tree. This tree represents the "true history" of the tumor. We then randomly sample a frequency for each mutation that adheres to the *sum rule* [5, 7] which states that the frequency of any mutation must be greater than or equal to the sum of frequencies of its children. This rule follows from the infinite sites assumption. To create each input tree, we then iterate through all nodes x in the tree and with a fixed probability of 0.3 do one of the two following edits: (1) Randomly choose a new parent y (that is not a descendant of x and so that the sum rule is not violated) and move the subtree rooted at x to be a child of y. (2) For the current parent y of x, if y and x have the same frequency, swap these nodes so that x is now the parent of y and y is the parent of $x's$ former children. Move 1 occurs with probability 0.8 and move 2 occurs when move 1 does not happen. We force these

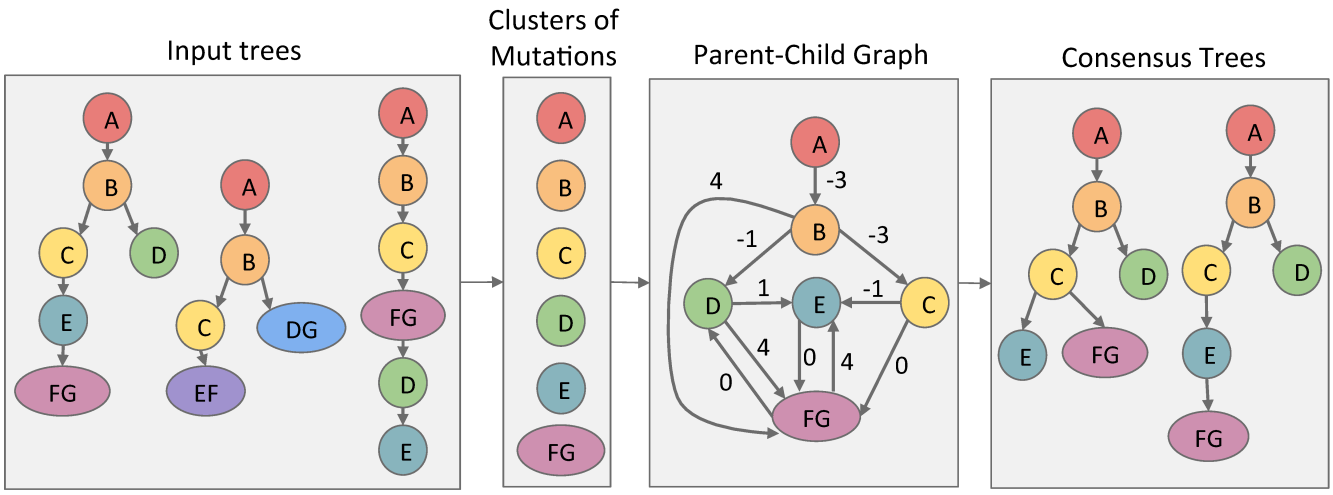

Figure 3: Overview of the GraPhyC method (described in Algorithm 3). First Algorithm 2 is used to identify clusters of mutations. Then, the parent-child graph G is built for these clusters of mutations. Finally, potential consensus trees are found as minimum spanning arborescences of graph G.

moves to maintain the sum rule since most algorithms that compute tumor evolutionary histories aim to adhere to this constraint.

3.1.2 Distance Measures and the SL-m-TCCP. We first evaluated the efficacy of our GraPhyC algorithm at solving the SL-m-TTCP compared to a baseline approach for all of our distance measures. This allows us to consider our consensus paradigm separately from our approach to mutation clustering. For this experiment we created 80 different sets of 5 input trees, each with 11 mutations. We kept the number of mutations low so that we could use a brute force search to determine the true optimal solution to the SL-m-TCCP for each set of input trees. This is also a similar number of mutations to the real dataset we analyze.

We compare our results on these simulated datasets to a baseline comparison we call "Best Input" which returns the input tree that minimizes the total distance to all other trees in the input set. For an inferred consensus tree T from a set of input trees $\{T_1, T_2, \ldots, T_n\}$ having a true optimal consensus tree T^* (found using brute force search) we compute the following normalized measure of error for a given distance measure $dist(\cdot, \cdot)$.

$$err(T) = \frac{\sum_{i=1}^{n} dist(T, T_i) - \sum_{i=1}^{n} dist(T^*, T_i)}{\sum_{i=1}^{n} dist(T^*, T_i)} \quad (5)$$

We use this measure of error rather than directly computing $dist(T, T^*)$ since multiple distinct optimal trees may exist or be returned by one of the tested methods. We compute this error for our four different distance measures (Path, Parent-Child, Ancestor-Descendant and Clonal) for the results output by both GraPhyC and Best Input. Figure 4 shows violin plots of these results. For each distance measure we see that the GraPhyC algorithm outperforms the the Best Input method. Specifically we see empirical verification of Theorem 2.3 that GraPhyC always returns the optimal solution ($err(T) = 0.0$) when the distance measure is Parent-Child distance, in contrast to Best Input which has a mean $err(T) = 0.075$. We also

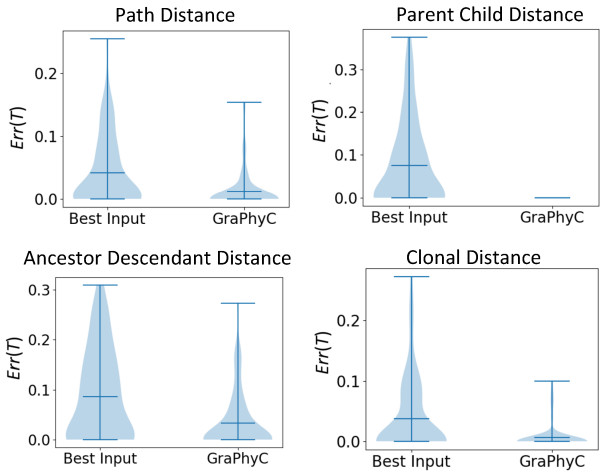

Figure 4: Violin plots comparing GraPhyC to a baseline Best Input method for four different distance measures in terms of error from the true optimal consensus tree (Equation (5)).

note that for all distance measures GraPhyC returns a tree with $err(T) = 0$ in over half of the trials (Best Input only achieves this for Parent-Child distance).

3.1.3 GraPhyC Performs Well at Uncovering the Original Tree. In the previous section we showed that GraPhyC does a good job at solving the SL-m-TTCP, but we have not yet shown that problem is a good way to uncover the original tree the data was derived from. We created simulated datasets that contain a range of mutations from the set $\{10, 15, 20, 25, 30, 35, 40\}$, while keeping track of the original tree T^* that was used to create the dataset. For each number of mutations we created 100 simulated datasets of 5 input trees and found the consensus trees identified by GraPhyC and the

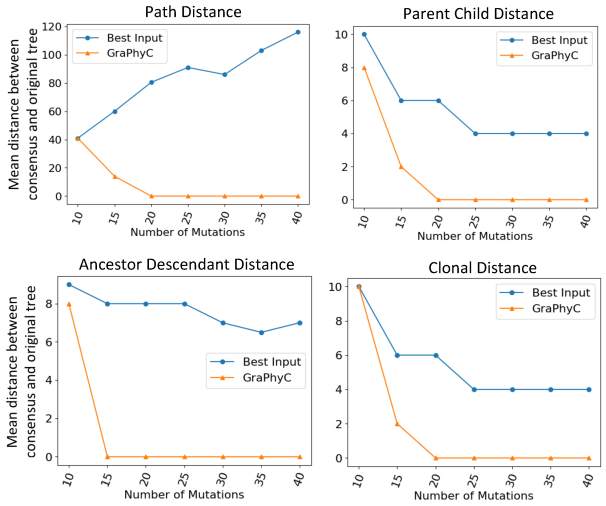

Figure 5: Mean distances between original underlying tree and either a consensus tree inferred by GraPhyC or Best Input for each of our distance measures.

Best Input method. We compare these trees to the original tree T^* using our four distance measures and find that GraPhyC outperforms the Best Input method (Figure 5). Furthermore, we also see that GraPhyC shows improved inference as the number of mutations increases. Thus, we can conclude that GraPhyC has better potential for uncovering the real underlying tree. Furthermore, in a number of these trials GraPhyC is able to uncover the true original tree even when that tree is not included in the set of input trees.

3.2 Results on Real data

We also analyze the results of our GraPhyC algorithm using chronic lymphocytic leukemia (CLL) sequencing data from Schuh *et. al* [25]. Sample CLL077 from this study has been extensively analyzed in terms of subclonal evolution. In particular, a number of methods for inferring tumor evolutionary histories have used this sample in their analysis[5, 7, 13, 20]. Data for this sample includes both ~40x average coverage whole-genome sequencing from 5 time points (plus a matched normal sample) and ~100,000X average coverage targeted deep amplicon sequencing of 16 mutations from those same 5 time points. So, while ground truth cannot be verified in real data such as this, the variety of data types available for this sample, and its extensive previous analysis, make it an ideal candidate for analysis with our consensus approach.

PhyloWGS [5] is a method for inferring tumor evolutionary histories that utilizes a Markov Chain Monte Carlo (MCMC) approach to sample possible histories from a posterior distribution. Since it is a stochastic approach PhyloWGS may sample different phylogenetic histories on different runs of the algorithm. Thus, a consensus approach applied to the different outputs produced by the method, may yield a better estimate of the underlying evolutionary history. We ran the PhyloWGS algorithm using default parameters multiple times on sample CLL077 using both the ultra deep sequencing data of the 16 selected mutations and the whole genome sequencing data for those same 16 mutations. PhyloWGS produced 4 distinct phylogenetic trees for the deep sequencing data ($d1$, $d2$, $d3$ and $d4$) and 4 distinct phylogenetic trees for the whole sequencing data ($w1$, $w2$, $w3$ and $w4$). Figure 6(a) and Figure 6(b) show these sets of inferred trees. We note that there is variation among the set of trees produces both in terms of topology and mutations placement. We quantify this variation by using our four distance measures to do pairwise comparisons across all created trees (Figure 6(c)). In particular we note that the set of trees inferred from the whole-genome data are much more similar as a group than the trees inferred from the deep sequencing data. For example, with the parent-child distance measure the average pairwise distance between the whole-genome trees is 62.13 compared to 83.63 for the deep sequencing trees.

We applied GraPhyC to the complete set of input trees obtained from PhyloWGS and identified a new tree (which we call $d3^*$), not contained in our input set, as the optimal consensus tree and shown in Figure 6(d). This tree is very similar to input tree $d3$ with one change - the mutation to gene *SLC12A1* moved from its cluster in $d3$ to one of the child nodes of that cluster. This new child cluster is in fact the cluster that *SLC12A1* is in for all of the whole-genome sequencing input trees. The PhyloWGS paper also reports this same location for *SLC12A1* and explicitly notes that this placement differs from that reported by an expert generated tree constructed using deep sequencing data. Furthermore, our inferred consensus tree ($d3^*$) is very similar to the tree reported in the PhyloWGS paper which reports tree $w1$. The one difference between these trees is that in our reconstruction mutation *NOD1* has been pulled out of a cluster and forms a new child node. This alteration matches that reported by a different method [7] which also pulled mutation *NOD1* out of this cluster and inferred it to occur in a leaf node of the tree. Thus, even though the ground truth is not known for this tumor, there is support for our reconstructed tree outside of the input trees used by the method.

We also wanted to test how sensitive the output of GraPhyC was to the particular set of trees selected as input trees. We therefore ran GraPhyC using different subsets of the whole-genome and deep sequencing trees for this same CLL077 sample. We tested with 12 different subsets of the 8 input trees (each subset tested had at least 3 input trees and included at least one whole-genome and one deep sequencing tree) and found that tree $d3^*$ was returned in 9 of these trials (and in every trial with at least 5 input trees) and the very similar tree $d3$ was returned in 2 trials (see Table 1). All trials where a different tree was returned contained fewer input trees (4 or less). Finally, we note that for all tested subsets, the tree returned by GraPhyC has a lower total distance from the set of input trees than the consensus tree returned by Best Input method for path, parent-child, and ancestor-descendant distances, except for path distance and the subset consisting of trees $w2, w3, w4, d1, d2, d4$.

4 CONCLUSIONS AND DISCUSSION

In this work we formalize the problem of finding a consensus tumor evolutionary history from a set of possible tumor histories as the m-TTCP. Novel consensus methods are needed to solve to this problem as existing methods are designed for traditional phylogenetic trees where extant species only label the leaves of the input trees. We define four distance measures between tumor trees, and develop

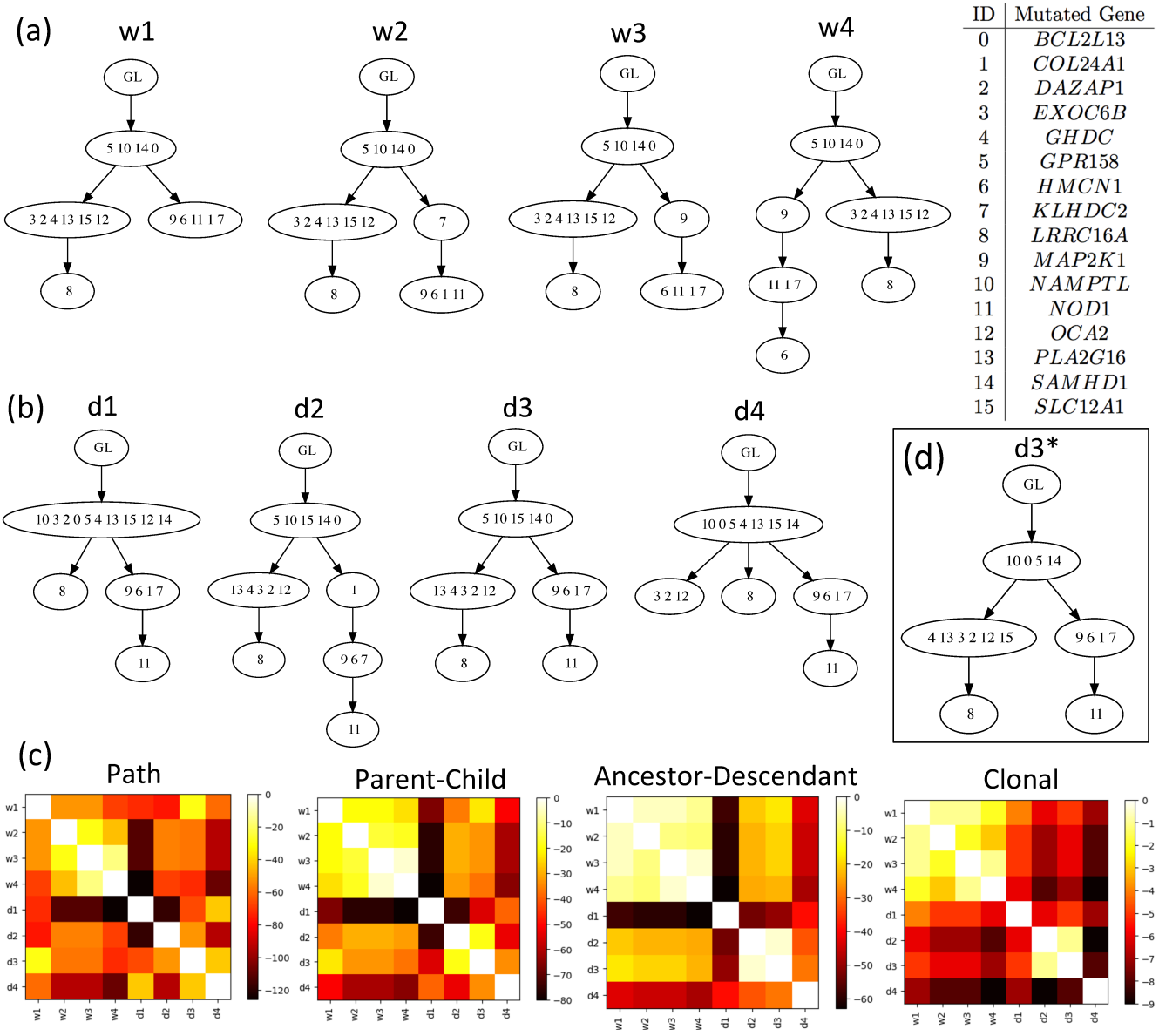

Figure 6: (a) Potential tumor histories of sample CLL077 found by PhyloWGS using whole-genome data. GL stands for germline. (b) Same as part (a) but using deep-sequencing data. (c) Pair-wise distances between all trees in parts (a) and (b) using four distance measures. (d) Consensus tree found by GraPhyC using all whole-genome and deep sequencing trees as input.

the GraPhyC algorithm which optimally solves a version of the *m*-TTCP for one of our distance measures. We demonstrate the efficacy of our GraPhyC approach compared to a baseline approach on both simulated trees and trees derived from real DNA sequencing data.

There are a number of important avenues for continuing this work. For example, we note that our proposed GraPhyC approach clusters mutations first, and then infers a consensus tree over these clustered mutations. The effect of clustering was a not a focus in this work and can certainly be further expanded on. Additionally, in a more ideal situation we would be able to perform clustering and consensus jointly rather than sequentially. We will explore different methods for combining these steps together. We could also expand our method to allow for more flexibility in terms of weighting confidence in the set of input trees or even ancestral relationships within those trees. We presented four possible distance measures between tumor evolutionary histories that capture different aspects of tumor evolution. We also plan to analyze these distance measures further, and perhaps to define new distance measures for computing the distance between tumor evolutionary histories.

Finally, there are a number of different applications where inferring a consensus tumor history from a set of plausible histories maybe useful. We have only partially explored these applications.

Table 1: Subsets of input trees for sample CLL077 and consensus tree inferred by GraPhyC.

Input Trees								Consensus Tree
$w1$	$w2$	$w3$	$w4$	$d1$	$d2$	$d3$	$d4$	
x	x	x	x	x	x	x	x	$d3^*$
x	x	x	x	x	x	x		$d3^*$
x	x	x	x		x	x	x	$d3^*$
x	x	x			x	x	x	$d3^*$
x	x	x		x	x	x		$d3^*$
	x	x	x	x	x		x	$d3^*$
x	x		x	x	x		x	$d3^*$
	x		x	x	x		x	$d3^*$
x		x		x		x		$d3^*$
x		x			x	x		$d3$
				x	x	x		$d3$
x		x				x		$w1$

Additional analysis including consensus across histories inferred by different methods, applied to both simulated and real sequencing data, would be useful at demonstrating the applicability of our GraPhyC method. Verification of the accuracy of such methods on real data remains challenging, but may be improved as more datasets with both bulk and single cell sequencing data of the same samples becomes available. We also note that our approach may also be useful in problems outside of tumor phylogenetics. In particular, our approach may be useful when considering the movement of transposable elements in a genome, a process which also may be described using a node-labeled tree.

ACKNOWLEDGEMENTS

This project is supported by NSF CRII award IIS-1657380.

REFERENCES

[1] Edward N Adams III. 1972. Consensus techniques and the comparison of taxonomic trees. *Systematic Biology* 21, 4 (1972), 390–397.

[2] Jean-Pierre Barthélemy and Fred R McMorris. 1986. The median procedure for n-trees. *Journal of Classification* 3, 2 (1986), 329–334.

[3] Paola Bonizzoni, Anna Paola Carrieri, Gianluca Della Vedova, and Gabriella Trucco. 2014. Explaining evolution via constrained persistent perfect phylogeny. *BMC Genomics* 15 Suppl 6 (2014), S10. https://doi.org/10.1186/1471-2164-15-S6-S10

[4] David Bryant. 2003. A classification of consensus methods for phylogenetics. *DIMACS series in discrete mathematics and theoretical computer science* 61 (2003), 163–184.

[5] Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. 2015. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol* 16 (Feb 2015), 35. https://doi.org/10.1186/s13059-015-0602-8

[6] Jack Edmonds. 1967. Optimum branchings. *Journal of Research of the National Bureau of Standards B* 71, 4 (1967), 233–240.

[7] Mohammed El-Kebir, Layla Oesper, Hannah Acheson-Field, and Benjamin J Raphael. 2015. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics* 31, 12 (Jun 2015), i62–70. https://doi.org/10.1093/bioinformatics/btv261

[8] Mohammed El-Kebir, Gryte Satas, Layla Oesper, and Benjamin J Raphael. 2016. Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures. *Cell Syst* 3, 1 (Jul 2016), 43–53. https://doi.org/10.1016/j.cels.2016.07.004

[9] George F. Estabrook, F. R. McMorris, and Christopher A. Meacham. 1985. Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units. *Systematic Biology* 34, 2 (1985), 193–200. https://doi.org/10.2307/sysbio/34.2.193

[10] Charles Gawad, Winston Koh, and Stephen R Quake. 2014. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc Natl Acad Sci U S A* 111, 50 (Dec 2014), 17947–52. https://doi.org/10.1073/pnas.1420822111

[11] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. 2016. Tree inference for single-cell data. *Genome Biol* 17 (May 2016), 86. https://doi.org/10.1186/s13059-016-0936-x

[12] Yuchao Jiang, Yu Qiu, Andy J Minn, and Nancy R Zhang. 2016. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc Natl Acad Sci U S A* 113, 37 (09 2016), E5528–37. https://doi.org/10.1073/pnas.1522203113

[13] Wei Jiao, Shankar Vembu, Amit G Deshwar, Lincoln Stein, and Quaid Morris. 2014. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics* 15 (Feb 2014), 35. https://doi.org/10.1186/1471-2105-15-35

[14] Salem Malikic, Katharina Jahn, Jack Kuipers, S. Cenk Sahinalp, and Niko Beerenwinkel. 2018. Integrative Inference of Subclonal Tumour Evolution from Single-Cell and Bulk Sequencing Data. In *Research in Computational Molecular Biology*, B.J. Raphael (Ed.). Springer, 269–270.

[15] Salem Malikic, Andrew W McPherson, Nilgun Donmez, and Cenk S Sahinalp. 2015. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics* 31, 9 (May 2015), 1349–56. https://doi.org/10.1093/bioinformatics/btv003

[16] Timothy Margush and Fred R McMorris. 1981. Consensusn-trees. *Bulletin of Mathematical Biology* 43, 2 (1981), 239–244.

[17] Yusuke Matsui, Atsushi Niida, Ryutaro Uchi, Koshi Mimori, Satoru Miyano, and Teppei Shimamura. 2017. phyC: Clustering cancer evolutionary trees. *PLoS Comput Biol* 13, 5 (May 2017), e1005509. https://doi.org/10.1371/journal.pcbi.1005509

[18] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. 2003. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning* 52, 1-2 (2003), 91–118.

[19] P C Nowell. 1976. The clonal evolution of tumor cell populations. *Science* 194, 4260 (Oct 1976), 23–8.

[20] Victoria Popic, Raheleh Salari, Iman Hajirasouliha, Dorna Kashef-Haghighi, Robert B West, and Serafim Batzoglou. 2015. Fast and scalable inference of multi-sample cancer lineages. *Genome Biol* 16 (May 2015), 91. https://doi.org/10.1186/s13059-015-0647-8

[21] F James Rohlf. 1982. Consensus indices for comparing classifications. *Mathematical Biosciences* 59, 1 (1982), 131–144.

[22] Edith M Ross and Florian Markowetz. 2016. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biol* 17 (Apr 2016), 69. https://doi.org/10.1186/s13059-016-0929-9

[23] Sohrab Salehi, Adi Steif, Andrew Roth, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah. 2017. ddClone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. *Genome Biol* 18, 1 (03 2017), 44. https://doi.org/10.1186/s13059-017-1169-3

[24] Gryte Satas and Benjamin J Raphael. 2017. Tumor phylogeny inference using tree-constrained importance sampling. *Bioinformatics* 33, 14 (Jul 2017), i152–i160. https://doi.org/10.1093/bioinformatics/btx270

[25] Anna Schuh, Jennifer Becq, Sean Humphray, Adrian Alexa, Adam Burns, Ruth Clifford, Stephan M Feller, Russell Grocock, Shirley Henderson, Irina Khrebtukova, et al. 2012. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood* 120, 20 (2012), 4191–4196.

[26] Russell Schwartz and Alejandro A Schäffer. 2017. The evolution of tumour phylogenetics: principles and practice. *Nat Rev Genet* 18, 4 (04 2017), 213–229. https://doi.org/10.1038/nrg.2016.170

[27] Mike A Steel and David Penny. 1993. Distributions of tree comparison metrics—some new results. *Systematic biology* 42, 2 (1993), 126–141.

[28] Francesco Strino, Fabio Parisi, Mariann Micsinai, and Yuval Kluger. 2013. TrAp: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Res* 41, 17 (Sep 2013), e165. https://doi.org/10.1093/nar/gkt641

[29] Charles Swanton. 2014. Cancer evolution: the final frontier of precision medicine? *Ann Oncol* 25, 3 (Mar 2014), 549–51. https://doi.org/10.1093/annonc/mdu005

[30] M S Waterman and T F Smith. 1978. On the similarity of dendrograms. *J Theor Biol* 73, 4 (Aug 1978), 789–800.

[31] W. T. Williams and H. T. Clifford. 1971. On the Comparison of Two Classifications of the Same Set of Elements. *Taxon* 20, 4 (1971), 519–522. http://www.jstor.org/stable/1218253

[32] Hamim Zafar, Anthony Tzen, Nicholas Navin, Ken Chen, and Luay Nakhleh. 2017. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biol* 18, 1 (Sep 2017), 178. https://doi.org/10.1186/s13059-017-1311-2

[33] Habil Zare, Junfeng Wang, Alex Hu, Kris Weber, Josh Smith, Debbie Nickerson, ChaoZhong Song, Daniela Witten, C Anthony Blau, and William Stafford Noble. 2014. Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput Biol* 10, 7 (Jul 2014), e1003703. https://doi.org/10.1371/journal.pcbi.1003703