# Distance Measures for Tumor Evolutionary Trees

Zach DiNardo*[1], Kiran Tomlinson*[1], Anna Ritz[2], and Layla Oesper†[1]

[1]Department of Computer Science, Carleton College, Northfield, MN, USA.
[2]Department of Biology, Reed College, Portland, OR, USA.

## Abstract

In recent years, there has been increased interest in studying cancer by using algorithmic methods to infer the evolutionary tree underlying a tumor's developmental history. Quantitative measures that compare such trees are then vital to benchmarking these algorithmic tree inference methods, understanding the structure of the space of possible trees for a given dataset, and clustering together similar trees in order to evaluate inheritance patterns. However, few appropriate distance measures exist, and those that do exist have low resolution for differentiating trees or do not fully account for the complex relationship between tree topology and how the mutations that label that topology are inherited. Here we present two novel distance measures, **C**ommon **A**ncestor **Set** distance (CASet) and **D**istinctly **I**nherited **S**et **C**omparison distance (DISC), that are specifically designed to account for the subclonal mutation inheritance patterns characteristic of tumor evolutionary trees. We apply CASet and DISC to two simulated and two breast cancer datasets and show that our distance measures allow for more nuanced and accurate delineation between tumor evolutionary trees than existing distance measures. Implementations of CASet and DISC are available at: `https://bitbucket.org/oesperlab/stereodist`.

*Joint first author

†Correspondence: loesper@carleton.edu

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

# 1 Introduction

A tumor is the result of an evolutionary process and its history can therefore be described as a rooted tree [1]. Specifically, the tree's root can either represent a healthy cell or the original founding tumor population (assuming monoclonal tumor evolution), and every other vertex represents a distinct tumor population that existed at some point during the tumor's evolution. Directed edges represent direct ancestral relationships between populations. In recent years, a number of methods have been developed to infer the tree describing a tumor's evolution from single nucleotide variants in bulk sequencing data [2–9] and single cell sequencing data [10–14] (see [15] for a more complete listing). The goal of tree inference is to gain a better understanding of tumor development, which may in turn reveal insights about the mutations that drive a tumor's growth [16, 17] and may be targeted for patient treatment [18, 19].

Due to the ongoing development of tumor evolution inference methods, the similarity of two potential tumor histories often needs to be quantified. First, new methods need to be benchmarked against other methods or against a ground truth tree, and the ad hoc measures that have typically been used in these situations [2–4] have not been rigorously studied. Second, some methods themselves rely on the use of a distance measure when inferring a tumor's evolutionary history. For example, the GraPhyC method [20] uses a distance measure to create a consensus tumor history from several input histories. Lastly, there have been growing questions about the structure of the space of possible evolutionary histories consistent with the underlying sequence data [21–23] and how tumor evolutionary histories across patients can be used to identify patterns of tumor evolution [24, 25]. Further analysis of these questions would be aided by distance measures tuned to the intricacies of tumor evolution histories.

For traditional phylogenetic trees, there are several well-known distance measures such as Robinson-Foulds [26] and triplet distance [27], among others. However, the substantial evolutionary differences between species and cancer subpopulations prevent the direct application of traditional phylogenetic distance methods [15]. The primary difference between the two types of trees is that tumor evolutionary trees have internal node labels, which represent mutations rather than extant species. Furthermore, the mutations in a tumor evolutionary tree are inherited by all descendant tumor populations, creating a complex underlying relationship between topology and mutation labeling. In addition, nodes in tumor evolutionary trees may contain multiple labels (indicating mutations whose order of appearance cannot be readily identified) whereas nodes in a traditional phylogeny contain only one label each [20, 28]. There is then a critical need for specially calibrated distance methods that account for the intricacies of tumor evolutionary trees.

Despite this need for tumor tree distance measures, a limited number of such measures have been rigorously developed and comprehensively evaluated. Several simple distance measures described by [20] generalized earlier ad hoc approaches that relied on the existence of a ground truth tree [2–4], but were not the focus of that work, and their effectiveness was not analyzed in depth. Another recently proposed approach called MLTED uses an edit distance-based measure focused on handling multi-labeled nodes to count the minimum number of moves to convert both trees into a specific common tree [28]. However, this distance measure does not explicitly consider that mutation labels are inherited by all descendant populations. We argue that distance measures that consider how mutation placement affects a tree's global structure will allow for much more precision when comparing tumor evolutionary trees.

In this work, we formalize the definition of a *tumor evolution distance measure* by precisely defining the input and output of such a function and describing what features of such a measure are desirable in the context of tumor evolution. We then describe two novel tumor evolution distance measures, CASet and DISC, that are specifically designed to account for structure of mutation inheritance by subsequent tumor populations and extend these measures to trees that do not share the same set of mutations. We apply our distance measures to two simulated datasets and two breast cancer datasets. We find that CASet and DISC allow for more precision when comparing tumor evolution histories and are better able to distinguish groups of similar trees than existing distance measures when applied in a clustering scenario.

# 2 Methods

## 2.1 Tumor Evolutionary Trees

We first make two common assumptions about tumor evolution. The first is the *infinite sites assumption* (ISA) which states that no mutation occurs more than once during a tumor's history, and that once gained, a mutation is never lost. This has been a common assumption made by many methods that infer tumor evolutionary histories ([2–4, 6, 8] and many others). While some recent phylogeny inference methods do allow for minor violations of the ISA (e.g. [29]), our work here will be most widely applicable if we assume the ISA. The second assumption is that all tumor cells are descended from a single founding tumor cell, and hence the tumor's evolution can be described as *monoclonal*. This assumption is non-essential to our approaches and can easily be dropped by rooting evolution trees with healthy cells instead of founding tumor mutations. Nonetheless, we make the monoclonal assumption to simplify our definitions. We now formally describe the evolutionary history of a monoclonal tumor adhering to the ISA as a clonal tree.

**Definition 2.1.** *A* ***clonal tree*** *is a rooted, directed tree $T$ in which: (i) each vertex in the tree is labeled by one or more mutations; and (ii) no mutation appears more than once.*

We also define $\mathcal{T}$ to be the set of all clonal trees. Given a tree $T \in \mathcal{T}$, we define $M(T)$ to be the set of all mutations (i.e. vertex labels) in $T$. In this representation, every vertex represents a distinct tumor clone (or population) that existed at some point during the tumor's evolution. Directed edges represent direct ancestral relationships between tumor clones. The mutation labels indicate the clone in which the mutation first appeared. Thus, the complete set of mutations that exist in any particular clone, represented by vertex $v$, is the set of mutations that label all vertices on the path from the root to vertex $v$.

We sometimes wish to restrict our attention to a predetermined set of mutations, so we also define $m$-clonal trees for this purpose.

**Definition 2.2.** *An* ***m-clonal tree*** *is a clonal tree $T$ with $M(T) = \{1, ..., m\}$.*

We emphasize that this definition uses the variable $m$ to refer to the set of mutations rather than the number of clones in the tree. We also define $\mathcal{T}_m$ to be the set of all $m$-clonal trees that share the same mutation set $[m] = \{1, 2, \ldots, m\}$.

## 2.2 Tumor Evolution Distances

In this section, we define a tumor evolution distance measure on clonal trees and then analyze what particular features are desirable for such a measure.

**Definition 2.3.** *A* ***tumor evolution distance measure*** *is a function $d : \mathcal{T} \times \mathcal{T} \longrightarrow \mathbb{R}^{\geq 0}$ for which a value of $d(\cdot, \cdot)$ that is close to 0 indicates that the two input clonal trees are very similar and progressively larger values of $d(\cdot, \cdot)$ indicate the clonal trees are more dissimilar.*

A tumor evolution distance measure must give us a quantitative evaluation of how different two tumor histories are from each other, but how best to define "different" is not immediately obvious. There are two main aspects of tumor evolutionary trees that should contribute to a distance measure: (i) the topology of the tree; and (ii) the labels present in the vertices of the trees. Topology can be separated from labeling by simply ignoring all labels, and the labels can be separated from topology by considering only the the set(s) of labels that appear or appear together. Thus, simple distance measures could certainly consider each of these aspects separately. However, these two tree attributes are inherently intertwined.

Since mutation labelings indicate in which vertex a mutation was first acquired, all descendants of that vertex also inherit that mutation. A difference in a vertex with many descendants should then contribute more to a distance measure than one in a vertex with few descendants, since it affects more clonal populations. Thus, a tumor evolution distance measure that simply counts the differences between trees (often referred to as a tree edit distance, as proposed by [28, 30], and others) does not address the impact any given label change may have. A distance measure should assign different weights to disagreements in different locations in order to appropriately address the relationship between topology and mutation labeling. These observations and analyses form the basis for the distance measures presented in the following section.

## 2.3 Two New Distance Measures on *m*-Clonal Trees

In this section, we first present some useful notation and then describe two novel tumor evolution distance measures restricted to $m$-clonal trees. In Section 2.4, we extend these measures to clonal trees in general.

### 2.3.1 Notation

Suppose $T \in \mathcal{T}$ is a clonal tree. We will need to translate between mutations and the vertices in a clonal tree that are labeled with those mutations. Therefore, given a mutation $i \in M(T)$, we will denote the vertex in $T$ that is labeled with mutation $i$ as $v_i$. We are interested in sets of mutations that label certain paths in a clonal tree. In particular, given $i \in M(T)$, we define the *ancestral set* $A(i)$ as the set of mutations that label the path from the root vertex $r$ to $v_i$ in $T$. This definition means that $A(i)$ gives the set of mutations that exist in the clone represented by vertex $v_i$. Given $i, j \in M(T)$, we define the *common ancestor set* $C(i,j)$ to be $A(i) \cap A(j)$. That is, $C(i,j)$ is the set of mutations that are ancestral to both mutations $i$ and $j$. Given $i, j \in M(T_k)$, we also define the *distinctly inherited set* $D(i,j)$ to be $A(i) \backslash A(j)$. That is, $D(i,j)$ is the set of mutations that are ancestral to mutation $i$ but not mutation $j$ in $T$. Note that under this definition it is almost always the case that $D(i,j) \neq D(j,i)$. When we have more than one tree, we use subscripts to distinguish between them. For instance, $A_k(i), C_k(i,j)$, and $D_k(i,j)$ all refer to $T_k$. See Figure 1 for examples of ancestral sets, common ancestor sets, and distinctly inherited sets of a tree $T_1$. Given two sets of mutations $A$ and $B$, we note that the *Jaccard distance* between them is defined as $\text{Jacc}(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$ and $\text{Jacc}(\emptyset, \emptyset) = 0$.

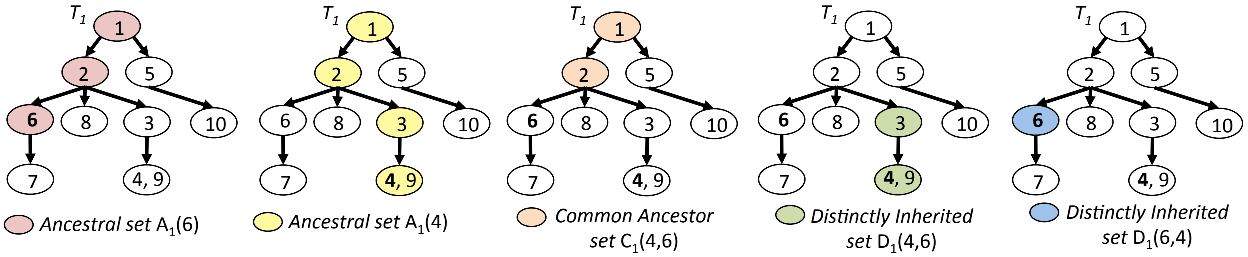

T1
1
2
5
6
8
3
10
7
4, 9
Ancestral set A1(6)
Ancestral set A1(4)
Common Ancestor set C1(4,6)
Distinctly Inherited set D1(4,6)
Distinctly Inherited set D1(6,4)

Figure 1: Examples of ancestral sets, common ancestor sets and distinctly inherited sets on one tree.

### 2.3.2 Common Ancestor Set Distance

Given two $m$-clonal trees $T_k, T_\ell \in \mathcal{T}_m$, we define a new tumor evolutionary tree distance measure called ***Common Ancestor Set** distance* (CASet) that computes a distance between $T_k$ and $T_\ell$. Informally, CASet distance is the average Jaccard distance between all corresponding common ancestor sets in $T_k$ and $T_\ell$. Equation (1) gives a formal definition of CASet distance.

$$\text{CASet}(T_k, T_\ell) = \frac{1}{\binom{m}{2}} \sum_{\{i,j\} \subseteq [m]} \text{Jacc}(C_k(i,j), C_\ell(i,j)) \tag{1}$$

**Observation 2.1.** *The running time to compute* $\mathrm{CASet}(T_k, T_\ell)$ *is* $O(m^3)$.

A proof of Observation 2.1 can be found in the appendix. See Figure 2 for an example of CASet distance.

### 2.3.3 Distinctly Inherited Set Comparison

CASet distance compares the common ancestor sets of all pairs of mutations, which emphasizes differences close to the root. However, we might also want to emphasize differences in more recently acquired mutations. Given two $m$-clonal trees $T_k, T_\ell \in \mathcal{T}_m$, we define a new tumor evolution distance measure between $T_k$ and $T_\ell$ called ***D**istinctly **I**nherited **S**et **C**omparison distance* (DISC) that accounts for mutation differences in the more recent tumor clones. Informally, DISC distance is the average Jaccard distance between all corresponding distinctly inherited ancestor sets in $T_k$ and $T_\ell$. Equation (2) gives a formal definition of the DISC distance.

$$\mathrm{DISC}(T_k, T_\ell) = \frac{1}{m(m-1)} \sum_{\substack{(i,j)\in[m]^2 \\ i\neq j}} \mathrm{Jacc}(D_k(i,j), D_\ell(i,j)) \tag{2}$$

Note that this range of summation is different from CASet, which only considers unordered pairs $\{i, j\}$.

**Observation 2.2.** *The running time to compute* $\mathrm{DISC}(T_k, T_\ell)$ *is* $O(m^3)$.

The proof of Observation 2.2 is similar to the CASet runtime proof and can also be found in the appendix. Figure 2 also contains an example DISC computation.

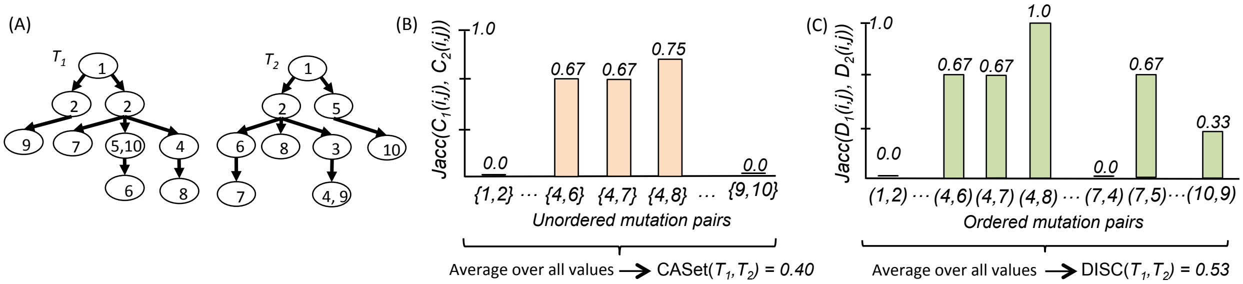


Figure 2: (A) A pair of 10-clonal trees. (B) Example of CASet distance applied to the 10-clonal trees in part (A). (C) Example of DISC distance applied to the 10-clonal trees in (A).

### 2.3.4 Algorithm Implementations

It is straightforward to implement these distance measures using set data structures to store all ancestor sets in the two trees. However, under certain circumstances, it may be faster or more convenient to compute CASet and DISC distances using matrix operations instead, which is described in Appendix A.2.

## 2.4 Extending CASet and DISC to Clonal Trees

Thus far, we have assumed that any two tumor evolutionary trees to be compared have the exact same set of mutation labels (that is, that they are both $m$-clonal). However, there are many scenarios in which this may not be the case. For instance, some methods such as [2] may not use all available mutations when creating a tumor evolutionary history tree. Further, different trees may be reconstructed from different data types for the same tumor (e.g. single cell and bulk sequencing in [31]) that do not share the same set of observed mutations. In this section, we present two extensions to both of our distance measures that allow for the comparison of clonal trees with different sets of mutation labels.

### 2.4.1 Intersection of Mutation Sets

In the first extension to clonal trees, we consider the intersection of the mutation sets for the input trees. This allows us to compute a distance between two trees by only considering pairs of mutations that the two trees share. Let $I_{k,\ell} = M(T_k) \cap M(T_\ell)$ be the intersection of the sets of mutations labeling $T_k$ and $T_\ell$. Thus, we can modify both CASet and DISC distances as follows:

$$\mathrm{CASet}_\cap(T_k, T_\ell) = \frac{1}{\binom{|I_{k,\ell}|}{2}} \sum_{\{i,j\} \subseteq I_{k,\ell}} \mathrm{Jacc}(C_k(i,j), C_\ell(i,j)) \tag{3}$$

and

$$\mathrm{DISC}_\cap(T_k, T_\ell) = \frac{1}{|I_{k,\ell}|(|I_{k,\ell}|-1)} \sum_{\substack{(i,j) \in {I_{k,\ell}}^2 \\ i \neq j}} \mathrm{Jacc}(D_k(i,j), D_\ell(i,j)). \tag{4}$$

We note that while we restrict attention to pairs of mutations that exist in both trees, the actual sets compared using the Jaccard distance may themselves contain mutations that exist in only one of the trees. Thus, these distances are not the same as removing all non-shared mutations from the trees, contracting the tree topology, and computing the original version of the distances.

This variation is most useful when differences in mutation labelings between trees should not strongly contribute to their distance value. For example, this method could be used when comparing trees reconstructed using two different methods, one of which does not use all input mutations. A degenerate case arises if this approach is applied to two trees with disjoint mutation sets. Since we sum over pairs of shared mutations, $\mathrm{CASet}_\cap$ and $\mathrm{DISC}_\cap$ are both $0$ in this case. If we want mutation set differences to be penalized, we can instead sum over pairs of mutations in the union of the two trees' mutation sets, as described in the following section.

### 2.4.2 Union of Mutation Sets

In the second extension to clonal trees, we consider the union of the mutation sets for the input trees. To do so, we need to address how to handle mutations that exist in only one tree. Let $U_{k,\ell} = M(T_k) \cup M(T_\ell)$ be the union of the sets of mutations labeling $T_k$ and $T_\ell$. If $i \notin M(T_k)$, then we define $A_k(i) = \emptyset$ and compute $C_k(i,j)$ and $D_k(i,j)$ as usual. Thus, we can modify both CASet and DISC distances as follows:

$$\mathrm{CASet}_\cup(T_k, T_\ell) = \frac{1}{\binom{|U_{k,\ell}|}{2}} \sum_{\{i,j\} \subseteq U_{k,\ell}} \mathrm{Jacc}(C_k(i,j), C_\ell(i,j)) \tag{5}$$

and

$$\mathrm{DISC}_\cup(T_k, T_\ell) = \frac{1}{|U_{k,\ell}|(|U_{k,\ell}|-1)} \sum_{\substack{(i,j) \in {U_{k,\ell}}^2 \\ i \neq j}} \mathrm{Jacc}(D_k(i,j), D_\ell(i,j)). \tag{6}$$

This variation allows differences in the sets of mutation labels to contribute to the distance computed between two trees. Thus, this variation may be most useful for comparing tumor evolutionary trees generated by different data types, across samples taken at different times, or even across patients. Notice that because $\mathrm{Jacc}(X, \emptyset) = 1$ if $X \neq \emptyset$, we no longer have distance $0$ between trees with disjoint labels. In the appendix, we describe a formula that relates $\mathrm{CASet}_\cup$ and $\mathrm{CASet}_\cap$, allowing us to compute $\mathrm{CASet}_\cup$ with fewer operations.

# 3 Results

We compare CASet and DISC to existing distance measures on two simulated datasets and two real datasets. We find that our methods allow for more granularity in comparing tumor evolutionary trees and outperform other methods when used in a clustering context.

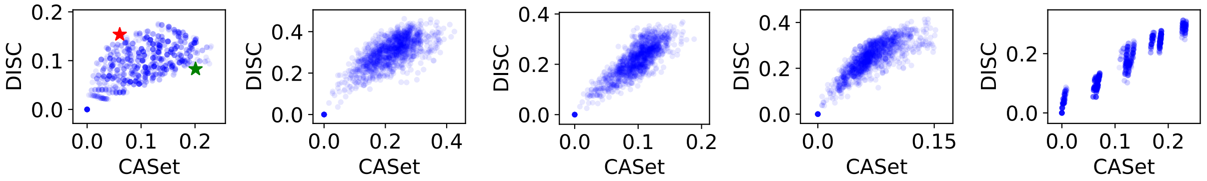
## 3.1 Results on Simulated Datasets

### 3.1.1 Dataset Generation

We created two different simulated datasets for our analysis. In the first dataset, we manually constructed five "base" clonal trees each with 15 mutations but different topologies and labelings (Figure A.1), and generated five variants of each base tree for a total of 25 trees. The second dataset was generated using the OncoLib [32] tree generation tool, which simulates tumor evolutionary trees and read count data. We took the read count data from five OncoLib simulations (each with 3 sequenced samples and 10–20 mutations) and reconstructed clonal trees that were consistent with the simulated data [22]. Each OncoLib simulation yielded a "tree family" of between 50 and 10,000 clonal trees, of which 50 were sampled to create a dataset of 250 trees from five families (labeled A–E). All trees within a family have the same set of mutations, but mutation sets differ across families. The true underlying clonal trees produced by OncoLib are in Figure A.5.

### 3.1.2 Correlation Between CASet and DISC

In each of the five tree families from the OncoLib dataset, the underlying tree structure has a strong effect on the correlation between CASet and DISC distances (Figure 3). In general, two trees with a large CASet distance also have a large DISC distance; however, family A includes some pairs of trees whose CASet and DISC distances differ significantly.



(a) Tree Family A ($r = 0.59$)

(b) Tree Family B ($r = 0.75$)

(c) Tree Family C ($r = 0.85$)

(d) Tree Family D ($r = 0.82$)

(e) Tree Family E ($r = 0.97$)

Figure 3: Correlation of CASet and DISC on tree families A-E on the OncoLib dataset. Pearson correlation coefficients are shown in parentheses. The red and green stars in (a) mark the tree pairs $T_{14}, T_{26}$ and $T_5, T_{20}$, respectively. Note that $x$ and $y$ scales are not consistent across all plots in order for all patterns to be visible.

The different patterns of CASet and DISC correlation are related to the height of the underlying true tree. For example, the true tree underlying family A (Figure A.5) is the shallowest of the set and has the lowest correlation between CASet and DISC, while the true tree of family E is the deepest tree and has the highest correlation. Shallow trees will tend to be wider, and therefore the ancestral set comparisons used by CASet will frequently only contain the root (whose labels are consistent due to the tree reconstruction process). As such, the Jaccard distance between the two sets will be 0 more frequently than in a deeper tree, in which more labels are in the ancestral set on average, causing any given label change to have an impact across more sets. In contrast, changes in shallow trees (with many leaves) will be weighted slightly more heavily by DISC by a similar argument. The relationship between tree height and distance correlation can also be seen for the manually constructed dataset (Figure A.2).

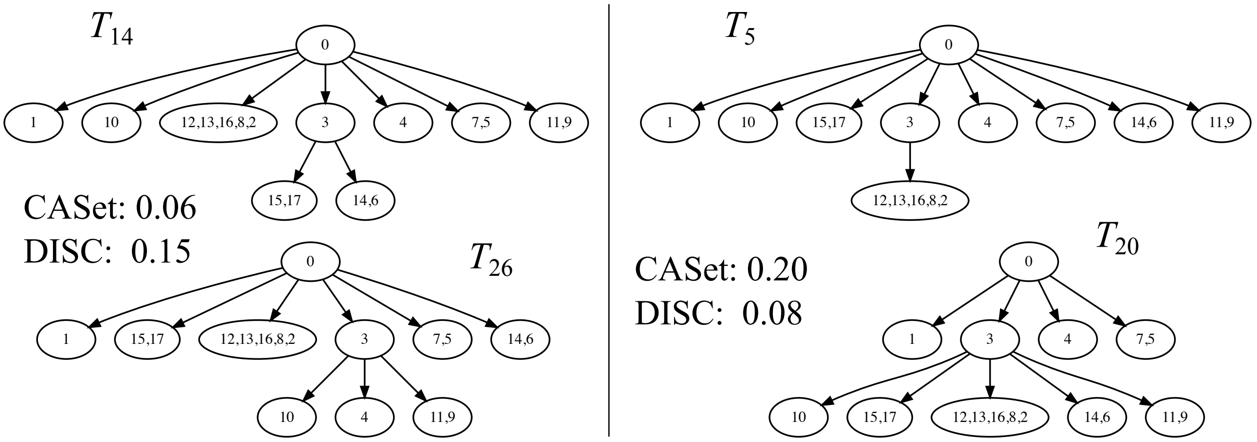


Figure 4: CASet and DISC calculations for trees $T_{14}$ and $T_{26}$ (left) and $T_5$ and $T_{20}$ (right) from the OncoLib dataset. These pairs of trees are starred in Figure 3(a).

We then examined two pairs of trees in the OncoLib dataset A for which CASet and DISC disagreed significantly, $\{T_{14}, T_{26}\}$ and $\{T_5, T_{20}\}$, highlighted in Figure 3a. According to CASet, $T_{14}$ and $T_{26}$ are more similar than $T_5$ and $T_{20}$, while DISC reports the opposite (Figure 4). The CASet distance between $T_{14}$ and $T_{26}$ is small, since ancestral relationships in the two trees are nearly identical (label 3 being the exception). However, the larger DISC distance between $T_{14}$ and $T_{26}$ reflects the significant differences in leaf locations in the two trees. Nine mutations are direct children of the root in both $T_{14}$ and $T_{26}$, but eight mutations are children of 3 in exclusively one tree. This results in a large number of mutation pairs with different sets of distinctly inherited ancestors, causing the DISC distance to be comparatively large. On the other hand, the CASet distance between $T_5$ and $T_{20}$ is large, reflecting the significant ancestral differences between the trees. Many more mutations have 3 as an ancestor in $T_{20}$ than in $T_5$. However, the DISC distance is relatively small because enough siblings move together between the two trees that their sets of distinct ancestors are mostly similar. Thus we see that CASet emphasizes ancestral similarity and performs especially well on deeper trees, while DISC prioritizes similarity in leaf relationships and is more granular with shallower trees.

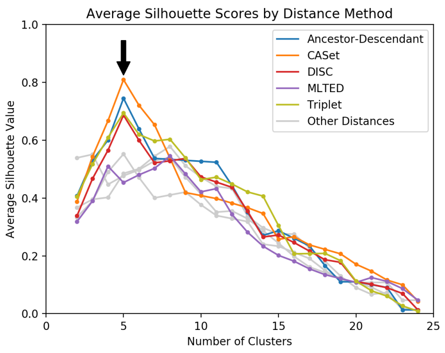


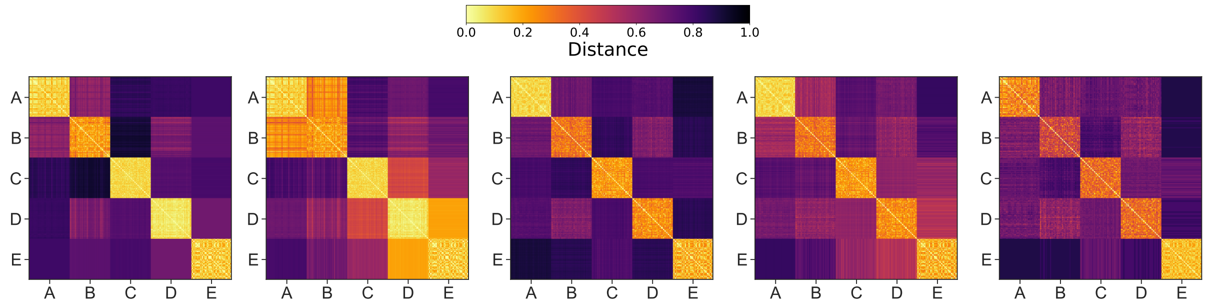
Figure 5: Plot of average silhouette values for different cuts of the hierarchical clustering with average linkage tree for each distance method on the manually created dataset. Three distances (clonal, parent-child, and path) did not have a large silhouette score for 5 clusters and are shown in gray. See Figure A.3 for a fully labeled plot.

### 3.1.3 Clustering Clonal Trees

Previous work has shown that many different hypothesized mutation trees can be consistent with data from a single patient [21, 22]. Clustering these trees is a compelling use case for tumor evolution distance measures as it has the potential to reveal structure in the space of compatible trees of a single dataset. Clustering trees inferred from different patients can also be used to identify shared evolutionary patterns. We compare our distance measures to MLTED [28], the four distance measures introduced in [20] (ancestor-descendant, parent-child, clonal, and path distances), and triplet distance [33] (a modified version of a distance designed for phylogenetic trees, described in A.4) using a clustering scenario on both simulated datasets.

We performed hierarchical clustering with average linkage on both the manual and OncoLib datasets. On the manual dataset, we compute the average silhouette value [34] for different cuts of the resulting tree on all tested distance measures and use the cut with the highest such value to produce a clustering of the data (Figure 5). Heatmaps showing all pairwise distances are in Figure A.4. We found that several of the methods (including CASet and DISC) produce the correct number of clusters, five.

On the OncoLib dataset, we applied $\mathrm{CASet}_\cup$, $\mathrm{CASet}_\cap$, $\mathrm{DISC}_\cup$, $\mathrm{DISC}_\cap$, and MLTED [28] when doing hierarchical clustering. Figure 6 shows heatmaps of all pairwise tree distances for the five tree families A–E and the average silhouette score for five clusters, which all but one method identified as the optimal hierarchical clustering cut ($\mathrm{CASet}_\cap$ had an optimal cut at three clusters). When cut at five clusters, all five distance measures correctly clustered the trees, but they did so with varying degrees of tightness. $\mathrm{CASet}_\cup$ performed best in distinguishing trees belonging to different datasets, with an average silhouette score of 0.81 over the five tree families. While it performs worse at separating different families, $\mathrm{CASet}_\cap$ identifies that the pairs A, B and D, E have strong agreement about ancestral relationships among their shared mutations (see Figure A.5). This highlights the different useful features of $\mathrm{CASet}_\cup$ and $\mathrm{CASet}_\cap$. In comparison, MLTED did a worse job of separating trees from different groups, with an average silhouette score of 0.54 over the five tree families. At the same time, it also does not recognize that the relationships between mutations shared by the D and E tree families are very similar.

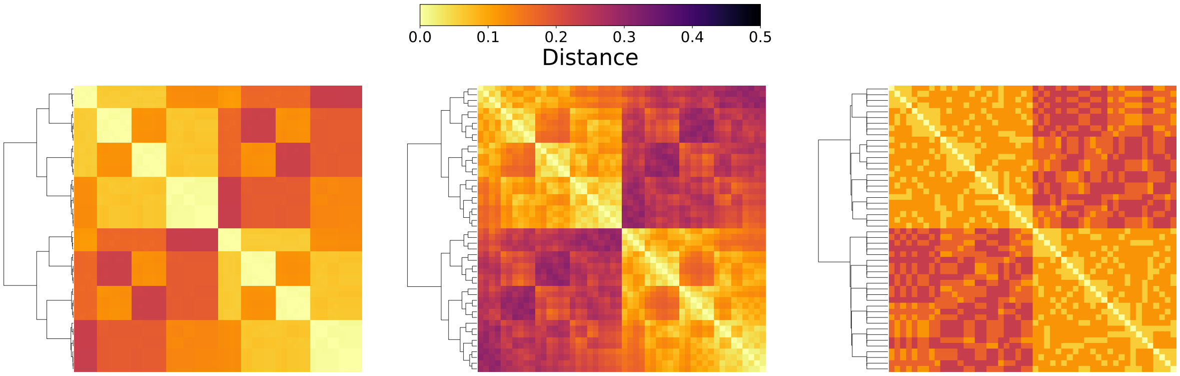


(a) $\mathrm{CASet}_\cup$ (0.81) (b) $\mathrm{CASet}_\cap$ (0.57) (c) $\mathrm{DISC}_\cup$ (0.70) (d) $\mathrm{DISC}_\cap$ (0.62) (e) MLTED (0.54)

Figure 6: Inter-dataset distance heatmaps of five tree families in the OncoLib dataset. The color of each cell represents the distance between two trees. Average silhouette scores for five clusters are displayed in parentheses to quantify clustering tightness and separation.

### 3.1.4 Intra-family Clustering Structure

Figures 3e and 6 show that tree family E from the OncoLib dataset might have internal structure. Therefore we took a closer look at the internal clustering structure for this tree family (see Figure 7). In this dataset, CASet, DISC, and MLTED all identify two primary clusters of trees, but CASet and DISC are both able to resolve more complex substructure. In particular, CASet distinguishes between eight strongly defined subfamilies, with an average silhouette score of 0.94. The same eight subfamilies are visible in the DISC heatmap along with a finer-grained resolution within each of these clusters. In contrast, the MLTED heatmap shows less resolution within the two primary categories. For a similar set of plots of family A, see Figure A.7, but none of the other four tree families had internal structure as well-defined as family E under any of the three distances.



(a) CASet

(b) DISC

(c) MLTED

Figure 7: Clustering of dataset E. Note that the colormap range has been reduced to provide more contrast.

## 3.2 Results on Real Datasets

We apply our distance measures to two different breast cancer datasets [36, 37]. We first apply CASet and DISC to the three potential tumor evolutionary histories reported in [28] recovered using different methods applied to single cell sequencing data from a triple negative breast cancer patient [36] (see Figure 8). Since CASet and DISC are designed to evaluate the topology of trees in addition to their labels and inheritance, both measures are able to provide more granular information about the similarity of the trees making it possible to conclude that $T_2$ is more similar to $T_1$ than it is to $T_3$. This is in contrast to MLTED [28], which considers these pairs of trees to have the same similarity, and, furthermore, computes the distance between $T_1$ and $T_3$ as 0 since they can be clonally expanded to match. While it may be useful to evaluate the similarity of trees with regard to such clonal expansion, a tumor evolutionary tree is categorized in part by subclonal populations that represent the evolutionary patterns of the tumor. Ignoring these does not fully take into account the information represented in a tumor evolutionary tree.

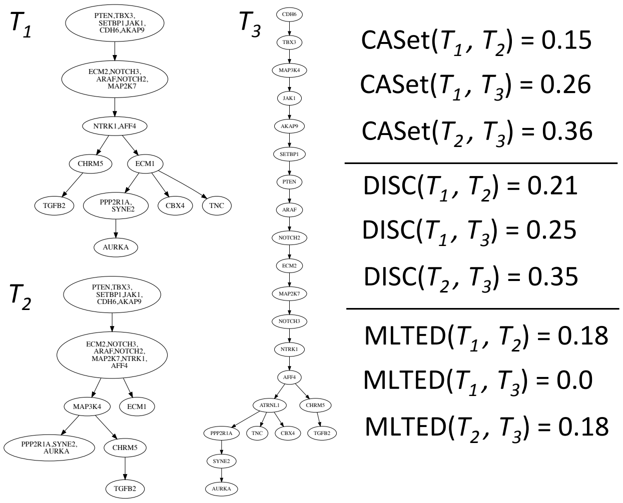

$T_1$
PTEN,TBX3, SETBP1,JAK1, CDH6,AKAP9
ECM2,NOTCH3, ARAF,NOTCH2, MAP2K7
NTRK1,AFF4
CHRM5
ECM1
TGFB2
PPP2R1A, SYNE2
CBX4
TNC
AURKA
$T_2$
PTEN,TBX3, SETBP1,JAK1, CDH6,AKAP9
ECM2,NOTCH3, ARAF,NOTCH2, MAP2K7,NTRK1, AFF4
MAP3K4
ECM1
PPP2R1A,SYNE2, AURKA
CHRM5
TGFB2
$T_3$
CDH6
TBX3
MAP3K4
JAK1
AKAP9
SETBP1
PTEN
ARAF
NOTCH2
ECM2
MAP2K7
NOTCH3
NTRK1
AFF4
ATRNL1
CHRM5
PPP2R1A
TNC
CBX4
TGFB2
SYNE2
AURKA
CASet($T_1$, $T_2$) = 0.15
CASet($T_1$, $T_3$) = 0.26
CASet($T_2$, $T_3$) = 0.36
DISC($T_1$, $T_2$) = 0.21
DISC($T_1$, $T_3$) = 0.25
DISC($T_2$, $T_3$) = 0.35
MLTED($T_1$, $T_2$) = 0.18
MLTED($T_1$, $T_3$) = 0.0
MLTED($T_2$, $T_3$) = 0.18

Figure 8: Tumor evolutionary trees inferred by PHiSCS [35] ($T_1$), SciFit [13] ($T_2$), and SCITE [10] ($T_3$) from a triple negative breast cancer patient [36] as reported by [28] and corresponding pairwise distances.

We also emphasize the importance of using quantitative measures when comparing reconstructed trees. The authors of [8] introduced a new tree reconstruction method MIPUP, and when they compare their results to those of another method, LICHeE [3], on breast cancer xenoengraftment data sample SA501 [37], they use only qualitative analysis to claim their method produces phylogenies closer to those in the original publication. We assessed this claim quantitatively by running CASet, DISC, and MLTED on the SA501 tree from [37] and the corresponding trees reconstructed by MIPUP and LICHeE, as shown in [8]. These trees had around 180 SNV mutations each in 5-7 nodes with slightly different mutation sets between trees. We found that the LICHeE tree was more similar to the phylogeny proposed by [37] than the MIPUP tree according to both CASet and DISC. The CASet/DISC distances were $0.837$/$0.397$ for the MIPUP tree and $0.781$/$0.377$ for the LICHeE tree. However, MLTED reports the opposite by a small margin, with distances of $0.798$ for the MIPUP tree and $0.811$ for the LICHeE tree compared to the proposed phylogeny. We note that with trees of this size, distances of this magnitude are more common than of the magnitudes represented in the OncoLib dataset, for example. However, the purpose of this analysis is to determine the relative distances of two trees to a ground truth tree and the raw distance is less relevant. Therefore, although the MIPUP tree may be marginally more similar to the originally proposed tree from a tree edit distance standpoint, it appears to be less similar when we account for the ancestral patterns of the disagreements.

## 4 Discussion and Conclusion

In this work, we argue that distance measures designed for tumor evolutionary trees are needed for assessing phylogeny inference methods and for exploring the relationships between sets of evolutionary trees. To this end, we introduce two new tumor evolution distance measures, CASet and DISC. By comparing the common ancestors of all mutation pairs, CASet incorporates differences in both mutation labeling and tree topology. In particular, CASet uses the number of clones that inherit common mutations when weighting the effect of mutation labeling differences between trees. In contrast, DISC pays special attention to the set of mutations that distinguish clones from each other, placing comparatively more emphasis on recently acquired mutations. We extend both distance measures to apply clonal trees with different sets of mutations.

We demonstrate the differences between CASet and DISC on simulated data and use a clustering application to show that that CASet is better able to distinguish groups of trees than existing distance measures. Moreover, we find that both CASet and DISC can identify complex clustering structure in a space of trees that is missed by the MLTED distance measure [28].

Using a breast cancer dataset, we show that CASet and DISC are able to differentiate between trees with differing clonal makeups, demonstrating their topological acuity. In addition, we use CASet and DISC to assess trees reconstructed by MIPUP [8] and LICHeE [3] from a breast cancer xenograft dataset [37]. Our results suggest that the MIPUP tree may not more closely resemble the original hypothesized tree from [37] than the LICHeE tree, as is suggested in [8] based solely on qualitative analysis. This highlights the importance of using quantitative measures such as CASet and DISC for these types of assessments.

Future work is needed to determine the benefits of using CASet or DISC in methods like GraPhyC [20] that rely explicitly on distance measures. Our findings on the internal clustering structure of the OncoLib tree families also invite further investigation. It remains to be seen whether the same kind of structure is found in real cancer data—either within sets of trees consistent with data from a single patient or across sets of patients with the same type of cancer. If so, this structure could provide insight into improved tumor phylogeny inference methods or common mutational patterns across patients.

# References

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

[2] Mohammed El-Kebir, Layla Oesper, Hannah Acheson-Field, and Benjamin J. Raphael. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70, Jun. 2015.

[3] Victoria Popic, Raheleh Salari, Iman Hajirasouliha, Dorna Kashef-Haghighi, Robert B West, and Serafim Batzoglou. Fast and scalable inference of multi-sample cancer lineages. *Genome Biol*, 16:91, May 2015.

[4] Salem Malikic, Andrew W McPherson, Nilgun Donmez, and Cenk S Sahinalp. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31(9):1349–56, May 2015.

[5] Gryte Satas and Benjamin J Raphael. Tumor phylogeny inference using tree-constrained importance sampling. *Bioinformatics*, 33(14):i152–i160, Jul 2017.

[6] Wei Jiao, Shankar Vembu, Amit G Deshwar, Lincoln Stein, and Quaid Morris. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*, 15:35, Feb 2014.

[7] Hosein Toosi, Ali Moeini, and Iman Hajirasouliha. BAMSE: Bayesian model selection for tumor phylogeny inference among multiple tumor samples. In *2017 IEEE 7th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, pages 1–1. IEEE, 2017.

[8] Edin Husić, Xinyue Li, Ademir Hujdurović, Miika Mehine, Romeo Rizzi, Veli Mäkinen, Martin Milanič, and Alexandru I Tomescu. MIPUP: Minimum perfect unmixed phylogenies for multi-sampled tumors via branchings and ILP. *Bioinformatics*, 2018.

[9] Noushin Niknafs, Violeta Beleva-Guthrie, Daniel Q Naiman, and Rachel Karchin. Subclonal hierarchy inference from somatic mutations: Automatic reconstruction of cancer evolutionary trees from multi-region next generation sequencing. *PLoS Comput Biol*, 11(10):e1004416, Oct 2015.

[10] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome Biol*, 17:86, May 2016.

[11] Sayaka Miura, Louise A Huuki, Tiffany Buturla, Tracy Vu, Karen Gomez, and Sudhir Kumar. Computational enhancement of single-cell sequences for inferring tumor evolution. *Bioinformatics*, 34(17):i917–i926, Sep 2018.

[12] Edith M Ross and Florian Markowetz. Onconem: inferring tumor evolution from single-cell sequencing data. *Genome Biol*, 17:69, Apr 2016.

[13] Hamim Zafar, Anthony Tzen, Nicholas Navin, Ken Chen, and Luay Nakhleh. Sifit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biol*, 18(1):178, 09 2017.

[14] Mohammed El-Kebir. SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics*, 34(17):i671–i679, Sep 2018.

[15] R. Schwartz and A. A. Schaffer. The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.*, 18(4):213–229, 04 2017. [PubMed Central:PMC5886015] [DOI:10.1038/nrg.2016.170] [PubMed:28190876].

[16] Benjamin J Raphael, Jason R Dobson, Layla Oesper, and Fabio Vandin. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med*, 6(1):5, 2014.

[17] Clemency Jolly and Peter Van Loo. Timing somatic events in the evolution of cancer. *Genome Biol*, 19(1):95, 07 2018.

[18] Collin M Blakely, Thomas B K Watkins, Wei Wu, Beatrice Gini, Jacob J Chabon, Caroline E McCoach, Nicholas McGranahan, Gareth A Wilson, Nicolai J Birkbak, Victor R Olivas, Julia Rotow, Ashley Maynard, Victoria Wang, Matthew A Gubens, Kimberly C Banks, Richard B Lanman, Aleah F Caulin, John St John, Anibal R Cordero, Petros Giannikopoulos, Andrew D Simmons, Philip C Mack, David R Gandara, Hatim Husain, Robert C Doebele, Jonathan W Riess, Maximilian Diehn, Charles Swanton, and Trever G Bivona. Evolution and clinical impact of co-occurring genetic alterations in advanced-stage EGFR-mutant lung cancers. *Nat Genet*, 49(12):1693–1704, Dec 2017.

[19] Nabil Amirouchene-Angelozzi, Charles Swanton, and Alberto Bardelli. Tumor evolution as a therapeutic target. *Cancer Discovery*, 7(8):805–817, 2017.

[20] Kiya Govek, Camden Sikes, and Layla Oesper. A consensus approach to infer tumor evolutionary histories. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '18, pages 63–72, New York, NY, USA, 2018. ACM.

[21] Dikshant Pradhan and Mohammed El-Kebir. On the non-uniqueness of solutions to the perfect phylogeny mixture problem. In *RECOMB Int. Conf. on Comparative Genomics*, RECOMB-CG '18, pages 277–293. Springer, 2018.

[22] Kiran Tomlinson and Layla Oesper. Examining tumor phylogeny inference in noisy sequencing data. In *2018 IEEE Int. Conf. on Bioinformatics and Biomedicine*, BIBM '18, pages 36–43. IEEE, 2018.

[23] Mohammed El-Kebir, Gryte Satas, Layla Oesper, and Benjamin J Raphael. Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. *Cell Syst*, 3(1):43–53, 07 2016.

[24] Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner, Stefan C Dentro, Santiago Gonzalez Rosado, Daniel Rosebrock, Thomas J Mitchell, Yulia Rubanova, Pavana Anur, Kaixan Yu, et al. The evolutionary history of 2,658 cancers. *bioRxiv*, page 161562, 2018.

[25] Yusuke Matsui, Atsushi Niida, Ryutaro Uchi, Koshi Mimori, Satoru Miyano, and Teppei Shimamura. phyc: Clustering cancer evolutionary trees. *PLoS Comput Biol*, 13(5):e1005509, 05 2017.

[26] David F Robinson and Leslie R Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147, 1981.

[27] Douglas E Critchlow, Dennis K Pearl, and Chunlin Qian. The triples distance for rooted bifurcating phylogenetic trees. *Systematic Biology*, 45(3):323–334, 1996.

[28] Nikolai Karpov, Salem Malikic, Md. Khaledur Rahman, and S. Cenk Sahinalp. A multi-labeled tree edit distance for comparing "clonal trees" of tumor progression. In Laxmi Parida and Esko Ukkonen, editors, *18th International Workshop on Algorithms in Bioinformatics (WABI 2018)*, volume 113 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 22:1–22:19, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[29] Paola Bonizzoni, Simone Ciccolella, Gianluca Della Vedova, and Mauricio Soto Gomez. Does relaxing the infinite sites assumption give better tumor phylogenies? An ILP-based comparative approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2018.

[30] Matt McVicar, Benjamin Sach, Cédric Mesnage, Jefrey Lijffijt, Eirini Spyropoulou, and Tijl De Bie. Sumoted: An intuitive edit distance between rooted unordered uniquely-labelled trees. *Pattern Recognition Letters*, 79:52 – 59, 2016.

[31] Charles Gawad, Winston Koh, and Stephen R Quake. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc Natl Acad Sci USA*, 111(50):17947–52, Dec 2014.

[32] El-Kebir Group. OncoLib: Library for tumor heterogeneity, 2018.

[33] Douglas E. Critchlow, Dennis K. Pearl, and Chunlin Qian. The triples distance for rooted bifurcating phylogenetic trees. *Systematic Biology*, 45(3):323–334, 09 1996.

[34] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.

[35] Salem Malikic, Simone Ciccolella, Farid Rashidi Mehrabadi, Camir Ricketts, Md Khaledur Rahman, Ehsan Haghshenas, Daniel Seidman, Faraz Hach, Iman Hajirasouliha, and S Cenk Sahinalp. Phiscs-a combinatorial approach for sub-perfect tumor phylogeny reconstruction via integrative use of single cell and bulk sequencing data. *BioRxiv*, page 376996, 2018.

[36] Yong Wang, Jill Waters, Marco L Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam, and Nicholas E Navin. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155–60, Aug 2014.

[37] Peter Eirew, Adi Steif, Jaswinder Khattra, Gavin Ha, Damian Yap, Hossein Farahani, Karen Gelmon, Stephen Chia, Colin Mar, Adrian Wan, Emma Laks, Justina Biele, Karey Shumansky, Jamie Rosner, Andrew McPherson, Cydney Nielsen, Andrew J. L. Roth, Calvin Lefebvre, Ali Bashashati, Camila de Souza, Celia Siu, Radhouane Aniba, Jazmine Brimhall, Arusha Oloumi, Tomo Osako, Alejandra Bruna, Jose L. Sandoval, Teresa Algara, Wendy Greenwood, Kaston Leung, Hongwei Cheng, Hui Xue, Yuzhuo Wang, Dong Lin, Andrew J. Mungall, Richard Moore, Yongjun Zhao, Julie Lorette, Long Nguyen, David Huntsman, Connie J. Eaves, Carl Hansen, Marco A. Marra, Carlos Caldas, Sohrab P. Shah, and Samuel Aparicio. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, 518:422, November 2014.

# A Appendix

## A.1 Runtime Proofs

*Proof of Observation 2.1.* We first do some pre-processing by computing $A_k(i)$ for all $i \in [m]$ using a depth first traversal of $T_k$. We keep track of the set of mutations on the path from the root to the current vertex while executing the algorithm. When visiting vertex $v$ during the traversal, we can then store $A_k(i)$ for all mutations $i$ that label vertex $v$. Since we store $m$ sets of mutations, each of which may contain up to $m$ mutations, this step takes $O(m^2)$. Next we sort each of these $m$ sets which takes $O(m^2 \log m)$. Next we can compute the $\binom{m}{2}$ common ancestor sets, each of which takes $O(m)$ to find since the Ancestor sets are sorted. This step is then $O(m^3)$. We do the same steps for $T_\ell$. Finally we need to compute the Jaccard distance between each of the $\binom{m}{2}$ common ancestor sets for $T_k$ and $T_\ell$. Since all these sets are in sorted order, each of these computations is $O(m)$. Thus, computing these Jaccard distances and summing them up to get the CASet distance is $O(m^3)$. Therefore, the overall runtime to compute CASet distance is $O(m^3)$. □

*Proof of Observation 2.2.* We first do some pre-processing by computing $A_k(i)$ for all $i \in [m]$ by doing a depth first traversal of $T_k$. We keep track of the set of mutations on the path from the root to the current vertex while executing the algorithm. When visiting vertex $v$ during the traversal, we can then store $A_k(i)$ for all mutations $i$ that label vertex $v$. Since we store $m$ sets of mutations, each of which may contain up to $m$ mutations, this step takes $O(m^2)$. Next we sort each of these $m$ sets which takes $O(m^2 \log m)$. Next we can compute the $m^2$ distinctly inherited sets, each of which takes $O(m)$ to find since the Ancestor sets are sorted. This step is then $O(m^3)$. We do the same steps for $T_\ell$. Finally we need to compute the Jaccard distance between each of the $m^2$ distinctly inherited ancestor sets for $T_k$ and $T_\ell$. Since all these sets are in sorted order, each of these computations is $O(m)$. Thus, computing these Jaccard distances and summing them up to get the DISC distance is $O(m^3)$. Therefore, the overall runtime to compute DISC distance is $O(m^3)$. □

## A.2 Matrix Implementation

Let $A_k$ denote the $m \times m$ binary *ancestor matrix* of $T_k$, which contains a 1 in entry $A_k[i, j]$ if mutation $i$ is an ancestor of mutation $j$. To compute CASet distance, we first have to find the common ancestors of every pair of columns in $A_k$ and in $A_\ell$. This can be done quickly by taking boolean ANDs of columns. Similarly, distinctly inherited sets can be found using XORs of columns. Then, the Jaccard distances in the algorithm can be computed by performing boolean ANDs and ORs on the common ancestor/distinctly inherited vectors to find intersections and unions, respectively, and finally summing the entries. Note that the process of computing all common ancestors within a tree should be performed as a pre-processing step if a large number of trees are being compared to each other. Lastly, we note that the asymptotic running time of this approach is $O(m^3)$, but it may still be faster in practice depending on available hardware and programming language specifics.

## A.3 A Direct Relationship Between CASet$_\cup$ and CASet$_\cap$

In Theorem 1, we describe a direct relationship that allows us to compute $\mathrm{CASet}_\cup$ based on $\mathrm{CASet}_\cap$. In particular, $\mathrm{CASet}_\cup$ is a weighted sum of $\mathrm{CASet}_\cap$ with terms that quantify the number and proportion of mutations unique to each tree. Using Theorem 1, we can compute $\mathrm{CASet}_\cap(T_k, T_\ell)$ and then use the formula to compute $\mathrm{CASet}_\cup(T_k, T_\ell)$ if desired. This allows us to consider fewer pairs $i, j$, which may be helpful in practice when the trees have significantly different sets of mutations. The relationships between mutations considered by DISC are more complex, which prevents us from performing a similarly clean expansion of $\mathrm{DISC}_\cup$.

**Theorem 1.**

$$\begin{aligned}\mathrm{CASet}_\cup(T_k,T_\ell) = \frac{1}{\binom{|U_{k,\ell}|}{2}}\Bigg(\binom{|I_{k,\ell}|}{2}\,\mathrm{CASet}_\cap(T_k,T_\ell) + |I_{k,\ell}||U_{k,\ell}\setminus I_{k,\ell}| \\ + \binom{|M(T_\ell)\setminus M(T_k)|}{2} + \binom{|M(T_k)\setminus M(T_\ell)|}{2}\Bigg)\end{aligned}$$

*Proof of Theorem 1.* By definition,

$$\mathrm{CASet}_\cup(T_k,T_\ell) = \frac{1}{\binom{|U_{k,\ell}|}{2}} \sum_{\{i,j\}\subseteq U_{k,\ell}} \mathrm{Jacc}(C_k(i,j),C_\ell(i,j))$$

We can break this sum into parts (a)-(d), based on the cases for $i$ and $j$:

(a) $i,j \in I_{k,\ell}$

(b) $i \in I_{k,\ell}$ and $j \notin I_{k,\ell}$

(c) $i,j \in M(T_\ell)\setminus M(T_k)$ or $i,j \in M(T_k)\setminus M(T_\ell)$

(d) $i \in M(T_\ell)\setminus M(T_k)$ and $j \in M(T_k)\setminus M(T_\ell)$

Notice that if either $i$ or $j$ does not appear in $T_k$, we will have $C_k(i,j) = \emptyset$, and likewise for $T_\ell$. Thus, in cases (b) and (c), we have $\mathrm{Jacc}(C_k(i,j),C_\ell(i,j)) = 1$. However, in case (d), the Jaccard distance is $\mathrm{Jacc}(\emptyset,\emptyset) = 0$. We'll use these case labels to abbreviate our summations. Using these facts and the definition of $\mathrm{CASet}_\cap$,

$$\begin{aligned}\mathrm{CASet}_\cup(T_k,T_\ell) &= \frac{1}{\binom{|U_{k,\ell}|}{2}}\Big(\sum_{\text{(a)}} \mathrm{Jacc}(C_k(i,j),C_\ell(i,j)) + \sum_{\text{(b)}} 1 + \sum_{\text{(c)}} 1 + \sum_{\text{(d)}} 0\Big) \\ &= \frac{1}{\binom{|U_{k,\ell}|}{2}}\Big(\binom{|I_{k,\ell}|}{2}\,\mathrm{CASet}_\cap(T_k,T_\ell) + \sum_{\text{(b)}} 1 + \sum_{\text{(c)}} 1\Big)\end{aligned}$$

Finally, notice that the number of pairs $i,j$ in case (b) is $|I_{k,\ell}||U_{k,\ell}\setminus I_{k,\ell}|$ and that the number of pairs in case (c) is $\binom{|M(T_\ell)\setminus M(T_k)|}{2} + \binom{|M(T_k)\setminus M(T_\ell)|}{2}$. We can then explicitly write out the (b) and (c) sums:

$$\begin{aligned}\mathrm{CASet}_\cup(T_k,T_\ell) = \frac{1}{\binom{|U_{k,\ell}|}{2}}\Bigg(\binom{|I_{k,\ell}|}{2}\,\mathrm{CASet}_\cap(T_k,T_\ell) + |I_{k,\ell}||U_{k,\ell}\setminus I_{k,\ell}| \\ + \binom{|M(T_\ell)\setminus M(T_k)|}{2} + \binom{|M(T_k)\setminus M(T_\ell)|}{2}\Bigg)\end{aligned}$$

□

### A.4 Extending Triplet Distance to Tumor Evolutionary Trees

The triplet distance is an algorithm designed to compute the similarity of phylogenetic trees based on subtree relationships [27]. For every triplet of leaf nodes (in a phylogenetic tree, each leaf node has exactly one label), the closer two nodes (determined via the last common ancestor of each pair) are recorded. The count of triplets for which this closeness does not match across trees is normalized by the total number of comparisons made, $\binom{m}{3}$ for an *m*-clonal tree, and recorded as the distance. It is possible to adapt this method to a tumor evolutionary tree by picking all triplets of mutation labels and ranking proximity based on the distance of their last common ancestors to the root.

## A.5 Additional Figures

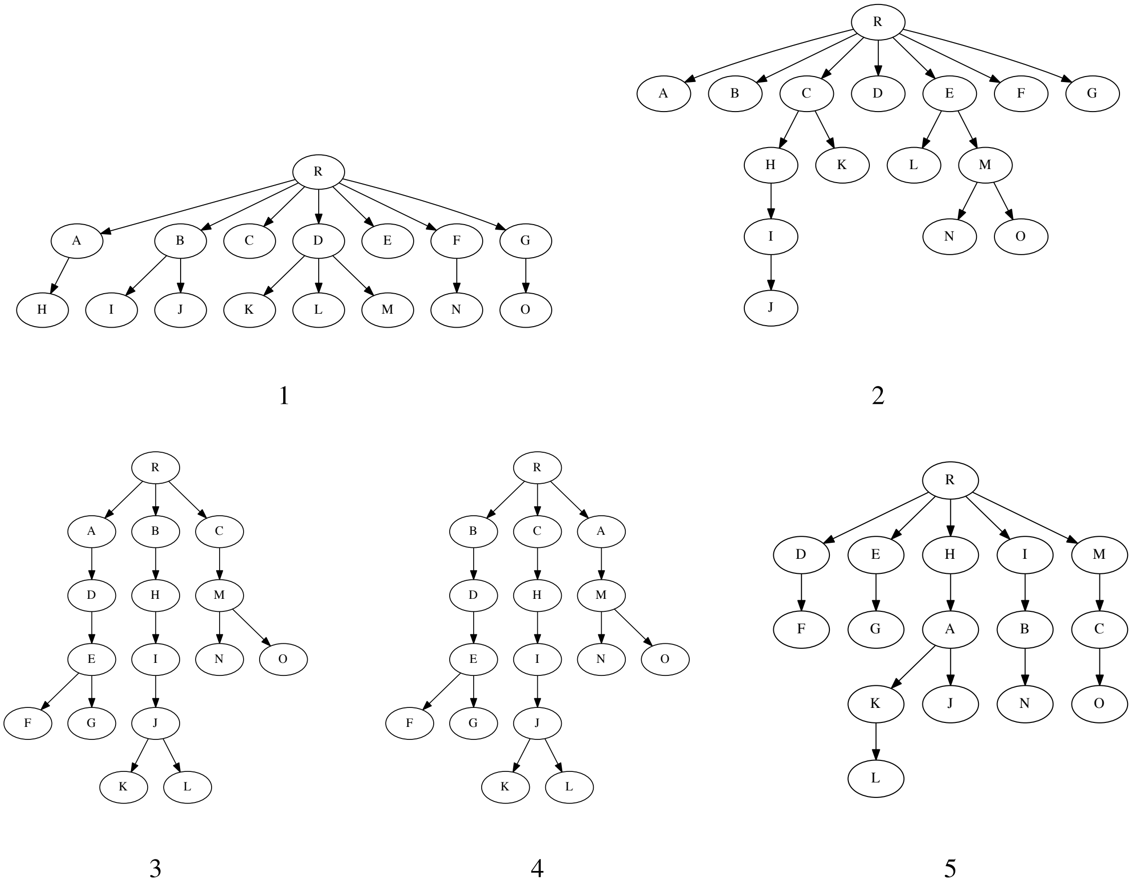
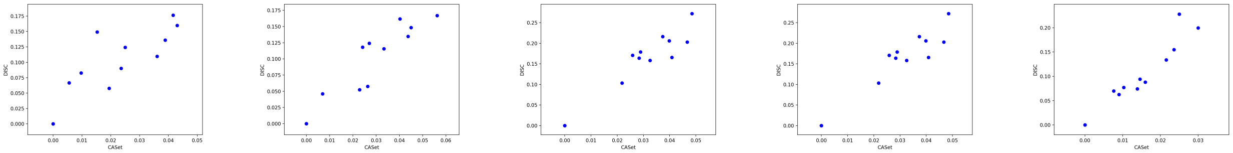


Figure A.1: The base trees for the manual clustering experiment, in which five variants of each of these trees were created.



(a) Tree Family 1 (b) Tree Family 2 (c) Tree Family 3 (d) Tree Family 4 (e) Tree Family 5

Figure A.2: Scatterplots of CASet and DISC correlation on the manual dataset. Again, the shallower trees (families 1 and 2) have the greatest disparity between distance measures (see Section 3.1.2 for an elaboration).

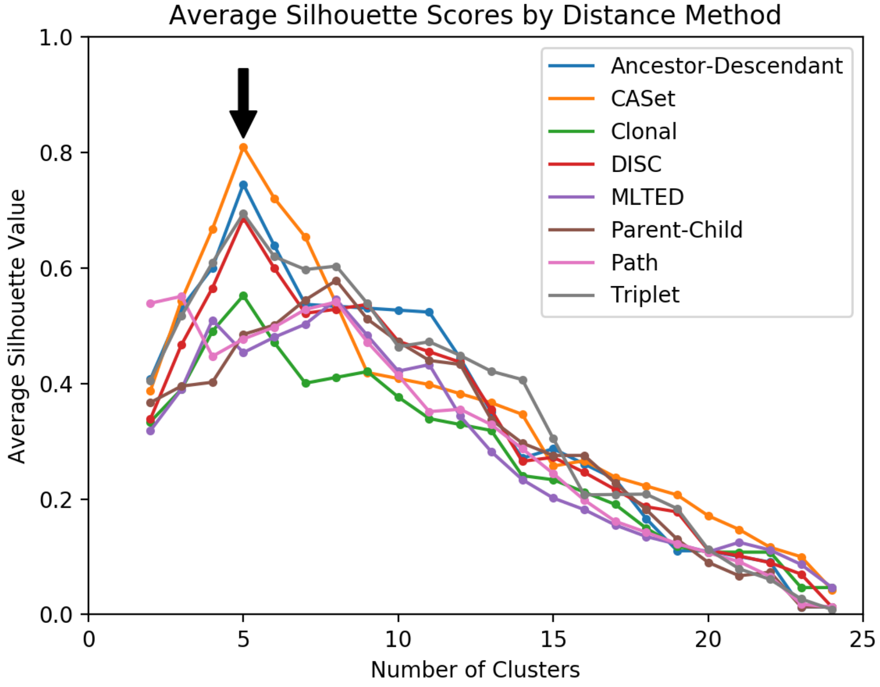
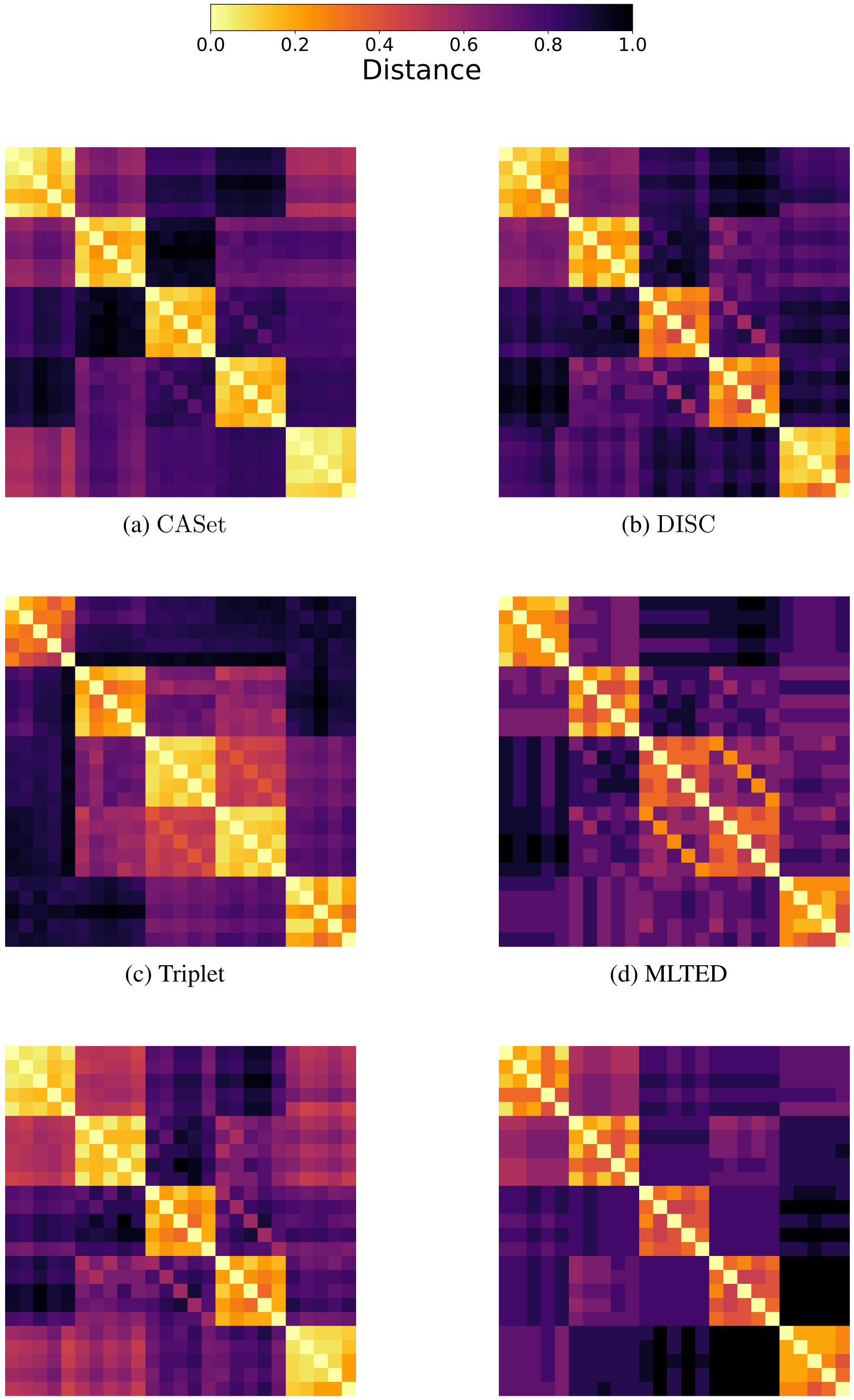


Figure A.3: All silhouette plots for the manual dataset.



(a) CASet

(b) DISC

(c) Triplet

(d) MLTED

(e) Ancestor-Descendent

(f) Clonal

Figure A.4: Heatmaps showing the (unclustered) pairwise distances between trees. In each, lighter shades represent closer trees. Defined 5-by-5 boxes show that the method measures trees within a cluster as being very similar to each other and less similar to other groupings.

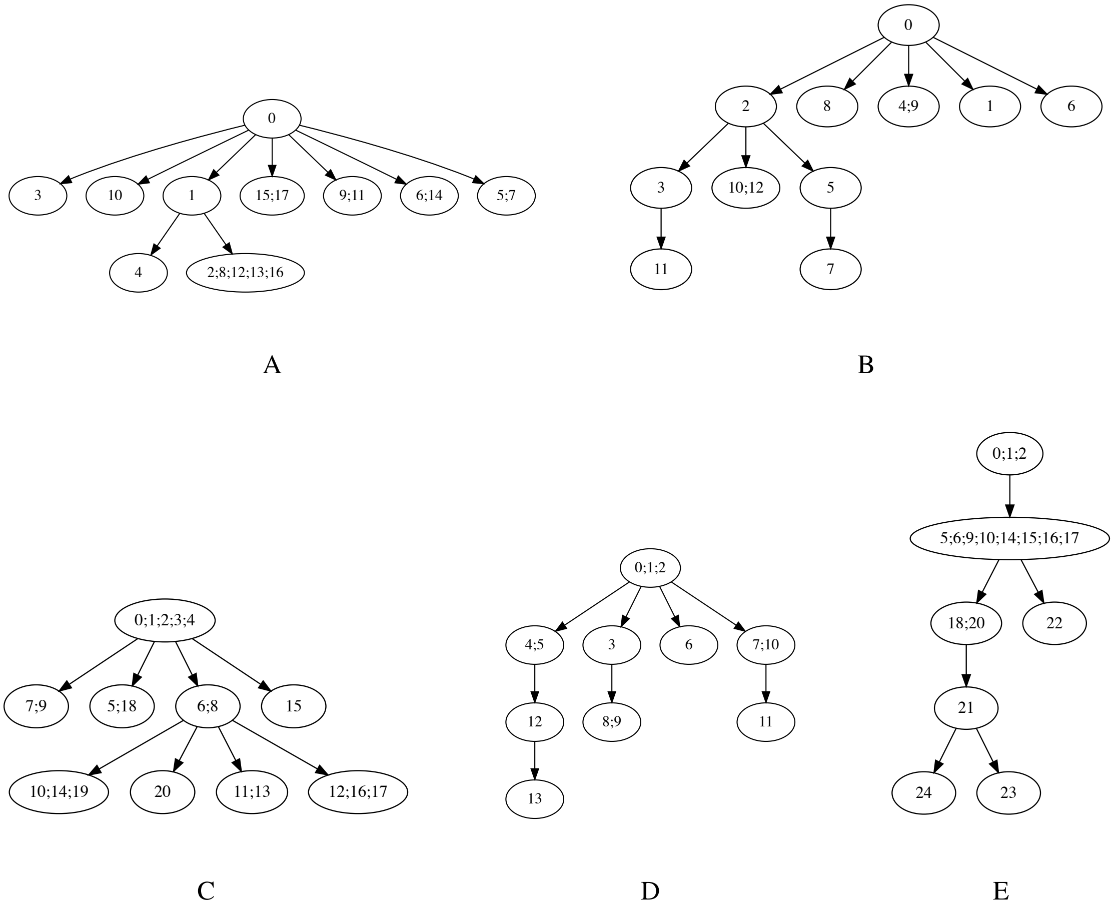
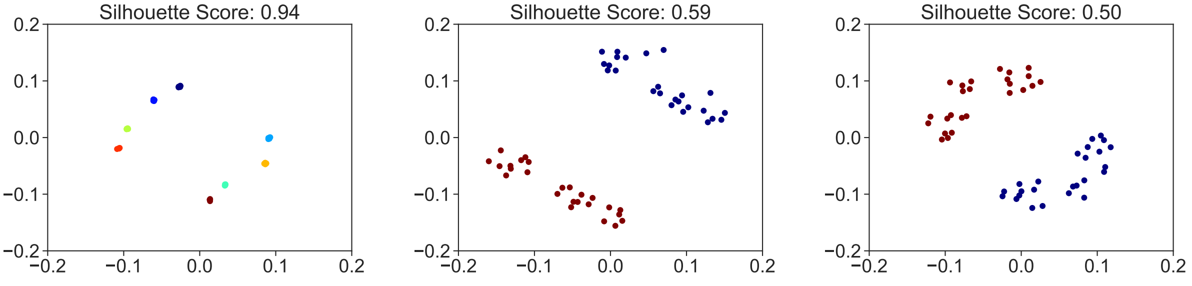


Figure A.5: Mutation trees from OncoLib. OncoLib simulations are structured such that mutations are represented by ordered integers, so all five tree families had at least 10 common mutations, often near the root of the tree.

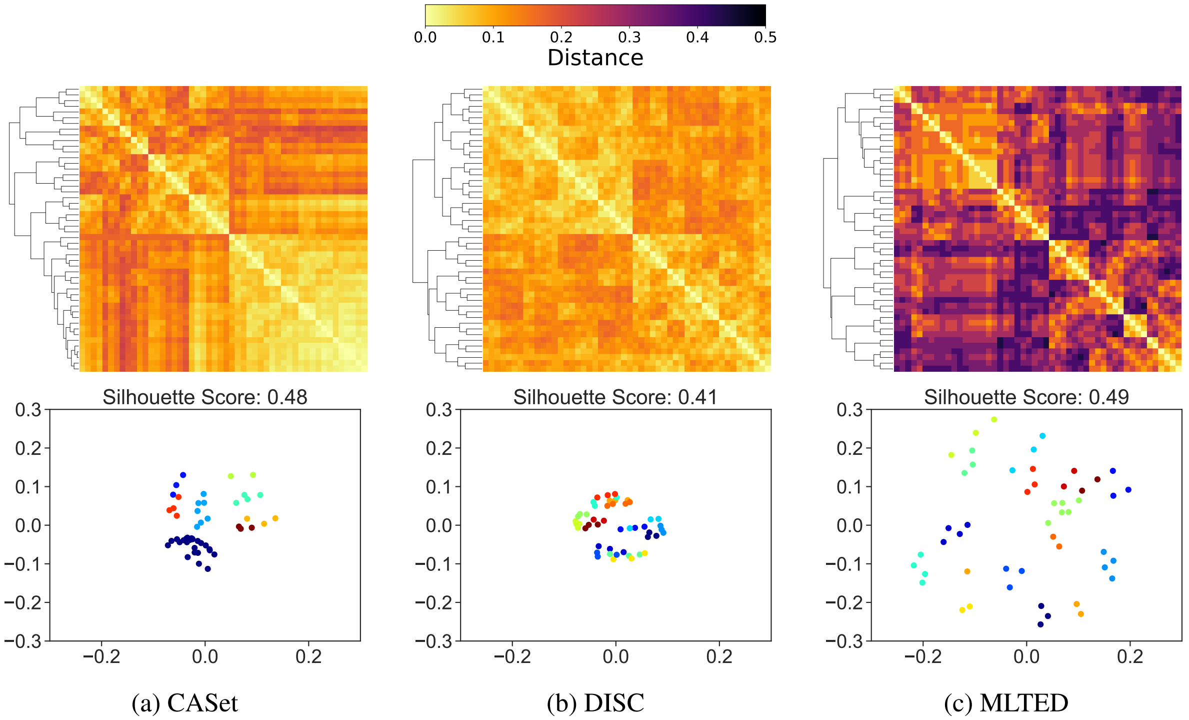


(a) CASet

(b) DISC

(c) MLTED

Figure A.6: Multidimensional scaling (MDS) plots of OncoLib family E. These provide another way of visualizing the clustering in Figure 7. Colors correspond to clusters in the optimal silhouette cut, the score for which is shown above each plot.

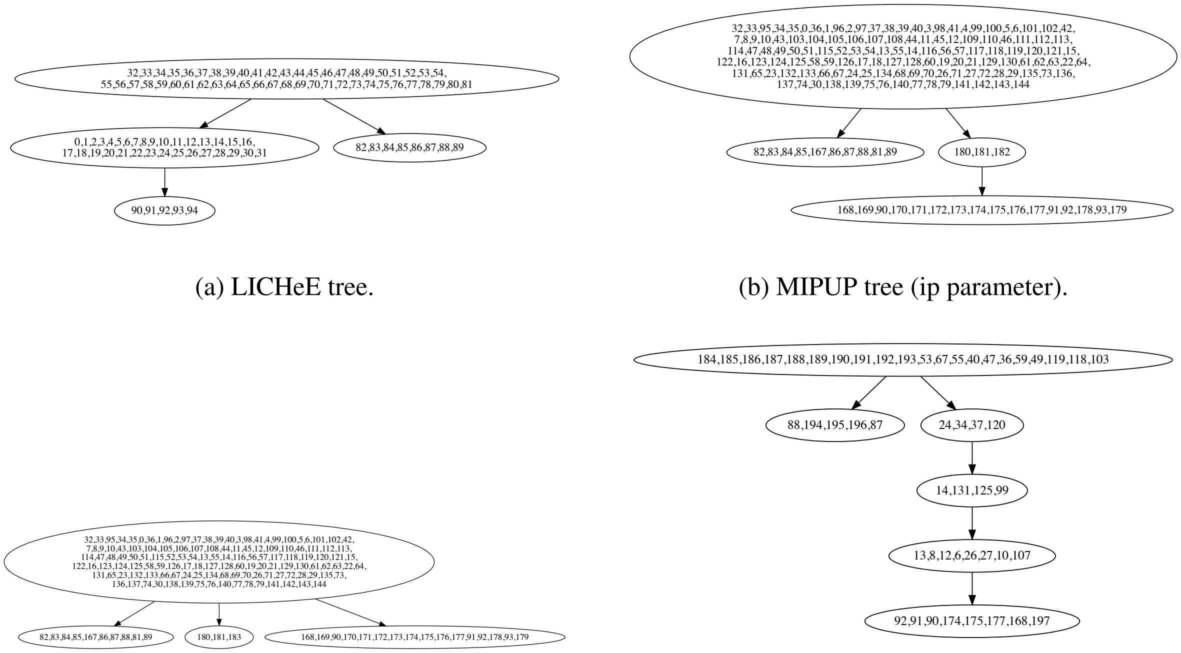


(a) CASet

(b) DISC

(c) MLTED

Figure A.7: Clustering of dataset A. The top row shows hierarchically clustered distance heatmaps and the bottom row shows multidimensional scaling plots.



(a) LICHeE tree.

(b) MIPUP tree (ip parameter).

(c) MIPUP tree (ipd parameter).

(d) Original tree from [37].

Figure A.8: Trees inferred from the SA501 dataset presented in [8]. We used tree (b) when referring to the MIPUP tree. The distances to tree (c) differed by an insignificant amount from the distances we reported (less than 0.01).