Purpose:

Tumors are the result of an evolutionary process (Nowell 1976). Clonal trees describe the evolution of a tumor by defining ancestor and descendant relationships between mutations. The comparison of clonal trees plays an important role in benchmarking software that generates trees from bulk sequencing data and in generating consensus trees from pools of tree candidates. While several techniques have been developed to measure the distance between trees, these techniques provide only a single number. As such, computational biologists lack an effective tool for assessing which aspects of a pair of clonal trees contribute the most to their distance. To fill this gap, we have implemented a visualization tool that, given two input trees, outputs a visualization comparing them according to a distance measure selected by the user. We currently support four such distance measures, each described below.

Parent-Child

High-level: Parent-child is a naive distance measure that counts the number of parent-child pairs that appear in one tree and not the other. It is the only of our four distance measures that we visually encode using

Details: A mutation *a* is a parent of the mutation *b* if the node containing *a* is the parent of the node containing *b*. The parent-child distance between two trees is the number of parent-child relationships that appear in one tree and not the other.

We visually encode parent-child distance using edge colorings. The contribution of an edge is equal to the number of parent-child pairs lying along that edge that do not appear in the other tree. We color edges with greater contribution blue, and edges with a smaller contribution yellow.

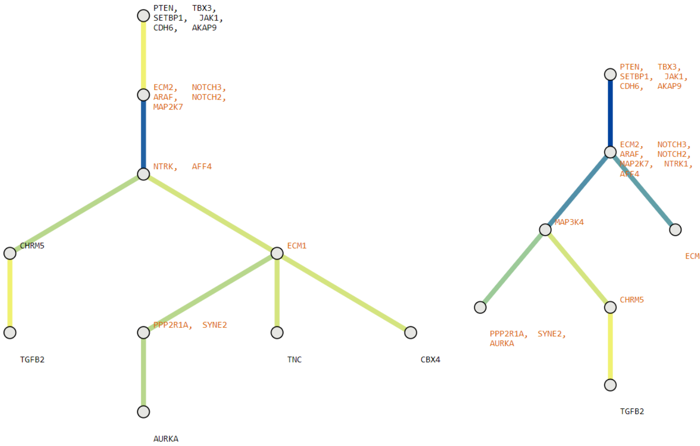
PTEN, TBX3, SETBP1, JAK1, CDH6, AKAP9
ECM2, NOTCH3, ARAF, NOTCH2, MAP2K7
NTRK, AFF4
CHRM5
ECM1
TGFB2
PPP2R1A, SYNE2
TNC
CBX4
AURKA
PTEN, TBX3, SETBP1, JAK1, CDH6, AKAP9
ECM2, NOTCH3, ARAF, NOTCH2, MAP2K7, NTRK1, AFF4
MAP3K4
ECM1
PPP2R1A, SYNE2, AURKA
CHRM5
TGFB2

Fig 1. The visualization of the parent-child distance of two sample trees, taken from triple-negative breast cancer data.

Ancestor-Descendant

The ancestor-descendant measure generalizes the parent-child measure by allowing a contributing pair of nodes to be connected by a directed path rather than an edge. In our implementation, each ancestor-descendant pair contributes 1 if it does not appear in the other tree, and 0 otherwise. The contribution of each node is determined by the number of contributing ancestor-descendant pairs it appears in.

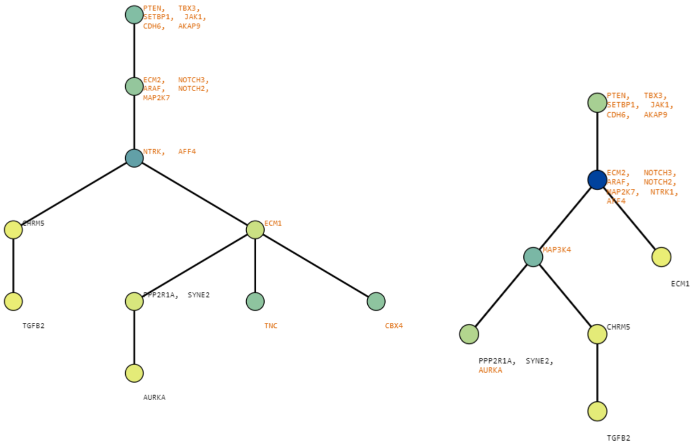
PTEN, TBX3, SETBP1, JAK1, CDH6, AKAP9
ECM2, NOTCH3, ARAF, NOTCH2, MAP2K7
NTRK, AFF4
CHRM5
ECM1
TGFB2
PPP2R1A, SYNE2
TNC
CBX4
AURKA
PTEN, TBX3, SETBP1, JAK1, CDH6, AKAP9
ECM2, NOTCH3, ARAF, NOTCH2, MAP2K7, NTRK1, AFF4
MAP3K4
ECM1
PPP2R1A, SYNE2, AURKA
CHRM5
TGFB2

Fig 2. The visualization of the ancestor-descendant distance of two sample trees, taken from triple-negative breast cancer data.

Common Ancestor Set (CASet)

Overview: CASet emphasizes mutation differences closer to the root of the tree.

Details: If mutations *i* and *j* appear in a tree, the *common ancestor set* of *i* and *j* (denoted $C(i, j)$) consists of the set of mutations that are ancestors of both *i* and *j*. The *Jaccard distance* between two sets *A* and *B* is given by

$$Jacc\ (A,\ B)\ =\ \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

Note that $Jacc(A,B)$ yields a number between 0 and 1 giving the proportion of elements not shared between *A* and *B*. If $Jacc(A,B) = 0$, then A=B, and if $Jacc(A,B) = 1$, then *A* and *B* do not share any elements.

The *CASet distance* between trees T_k and T_l with a full set of mutations *m* is computed as follows

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

That is, the CASet distance measure looks at each pair of mutations present in the trees and computes the common ancestor sets of these mutations in each tree. It takes the Jaccard distance between the common ancestor sets in each tree, and then averages this across all mutation pairs.

From a visualization perspective, the difficulty lies in visually encoding information about the distance between two trees. The key to achieving this task lies in the observation that, much like the ancestor-descendant distance measure, CASet functionally assigns a contribution to each ancestor-descendant pair of mutations, then takes the sum of those contributions. We simply need to extract the contributions of each ancestor-descendant pair, then depict them visually in the same way we depicted the ancestor-descendant distance.

How does CASet assign contributions to ancestor-descendant pairs? If a mutation appears near the root in the left tree, and near a leaf in the right, it will appear in many common-ancestor sets of mutations in the left tree, and few on the right. Each time such a mutation appears in a common ancestor set of two mutations lower down on a tree, its ancestor-descendant relationship with each of the lower mutations contributes towards the distance measure.

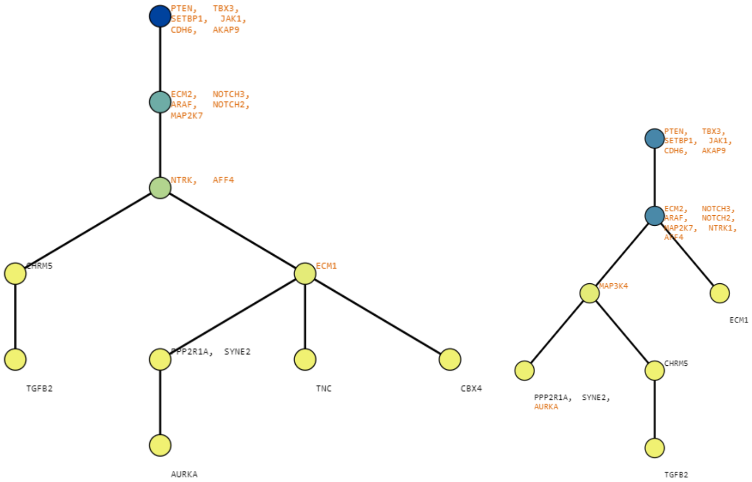

Fig 3. The visualization of the CASet distance of two sample trees, taken from triple-negative breast cancer data.

Distinctly Inherited Set Comparison (DISC)

If mutations A and B appear in the same tree, the distinctly inherited ancestor set of A and B consists of the set of mutations that are distinct ancestors of only either A or B. The DISC distance measure is the average Jaccard distance between all corresponding inherited ancestor sets between trees T_k and T_ℓ:

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

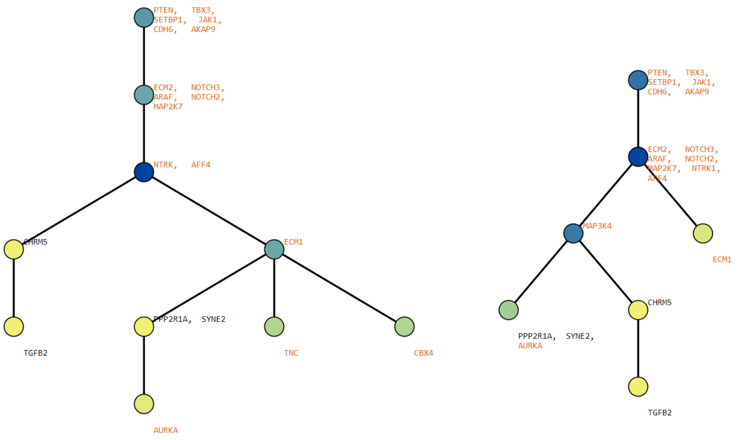

Fig 4. The visualization of the DISC distance of two sample trees, taken from triple-negative breast cancer data.