Tumor Phylogeny Consensus

Allie Warren

8/12/16

Abstract

Accurately modeling the evolutionary process of cancer and the composition of tumors is an uncertain process and current methods often provide contradicting results. I present two methods to aggregate information from contrasting phylogenies of the same tumor to create a single consensus tree. These methods identify similarities in substructures and clustering across input trees. In tests with simulated data, the consensus methods were able to reasonably approximate the true tree when the data was not highly variable. The distance based consensus method was able to more closely approximate the true tree, in terms of topology and clustering, than the input trees.

Background

Tumors are formed through an evolutionary process and often consist of a heterogeneous mixture of subpopulations. Each subpopulation is made up of a collection of cells with the same mutations. Identifying the evolutionary history of the tumor, mutations within the tumor and the proportions of each population within the tumor could significantly assist cancer research [9]. This information could make it possible to predict the progression of the tumor and create therapies that target the different mutation populations.

Tumor phylogenies, showing the evolutionary relationships and subpopulations of the tumor, are often inferred from bulk sequencing of mixed tumor samples [11]. The tumor sequence shows a statistical sample of the variety of genomes within the tumor, making it difficult to identify the separate subpopulations in the tumor. This problem is further complicated by the fact that data is often noisy. Some current methods to infer tumor phylogenies output multiple possible trees. Different methods for constructing tumor phylogenies may also produce different trees. The problem is how to combine the information from all of these trees into one representative tree. Ideally, this "consensus" tree is able to best reflect the variety of results.

Tumor phylogenies differ from classical phylogenetic trees is multiple respects, making the application of classical phylogenetic consensus methods limited. Unlike classical phylogenies, subpopulations still present in the current tumor may appear at internal nodes of the tumor phylogeny. Furthermore, nodes of the tumor phylogeny are composed of a variety of mutations, which often differ between different constructions of the phylogeny. Therefore new methods are required to create consensus trees for tumor phylogenies.

Model Assumptions

In order to simplify and define the problem, I make some assumptions about cancer development and phylogenies. I assume that tumor phylogenies develop according to clonal evolutionary theory. This asserts that tumor cells are derived from ancestral cells that have acquired growth advantages over normal cells. New mutation subpopulations emerge as cells acquire further survival advantages over their parent populations [9, 8, 1].

All input trees must meet the definition of a tumor phylogeny.

Definition: A tree s_i is a *tumor phylogeny* provided:

1. s_i is a tree rooted at the normal cell, where all subsequent clusters are descendants of the normal cell.

2. Each mutation in tree s_i occurs only once in the tree (assuming clusters depict only newly acquired mutations) and once a mutation occurs it does not revert back to its previous state. This is the *infinite sites assumption*. This assumption is nearly always valid as mutations are relatively rare [6].

3. Each mutation in s_i has a *mutational frequency*, which is the percentage of cells in the current tumor that contain that mutation. Each node contains all of the mutations of its ancestral nodes, therefore the mutational frequency of a mutation is always greater than or equal to the mutational frequency of any of its descendants.

4. Each vertex in s_i has a *population frequency*, which is the average of the mutational frequencies of the mutations clustered at that vertex.

5. s_i adheres to the constraints of the *sum rule*, meaning that the sum of the population frequencies of the direct children of a vertex must be less than or equal to the population frequency of that parent vertex [6].

6. s_i can be defined by a m x m distance matrix. A m x m distance matrix d defines a tree provided that d_{ij} is the number of edges on the shortest path from $mutation_i$ to $mutation_j$ and $d_{ij} = d_{ji}$.

Definition: The distance between two trees t_i and t_j is defined as the sum of the absolute values of all differences between matrix entries in the distance matrix of t_i and the distance matrix of t_j [12].

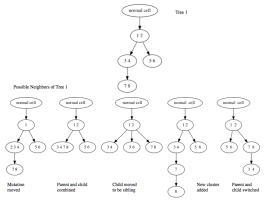
Definition: The *complete-data distance* is the sum of the distances between a tree and each tree in the dataset.

Definition: Trees t_i and t_j are neighboring trees if no cell d_{xy} in the distance matrix that defines t_i differs by more than one from the cell d_{xy} in the distance matrix that defines t_j.

Definition: k-distance neighbor

Trees t_i and t_j are k-distance neighbors if no cell d_{xy} in the distance matrix that defines t_i differs by more than k from the cell d_{xy} in the distance matrix that defines t_j. A normal neighboring tree could also be described as a 1-distance neighbor.

Figure 1: Example of neighboring trees (1-distance neighbors)



The consensus methods I describe do not create consensus mutational frequencies. Therefore, these trees will not meet conditions 3-5 of a tumor phylogeny, but the consensus trees do meet the other conditions.

Tumor Phylogeny Consensus Problem

Given a set S of rooted trees, output a single consensus tree T that contains the m mutations found every tumor phylogeny in S and for which each vertex in T contains at least one mutation and no mutation occurs more than once in T.

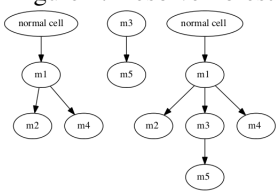
Methods

The **Majority Vote Method** finds the consensus tree T by creating a tree in which the parent-child relationships and clusterings of mutations are those that occurs most frequently in S. In this method, initially all mutations that occur in the same cluster and have the same children and parent in the majority of input trees are clustered together. Edges between mutations that appear in the majority of trees are also included in the consensus tree. This can result in T being a forest, having cycles, mutations occurring more than once or clusters in T having more than one parent. These conflicts are resolved, in no specified order, by adding edges that are more likely in order to connect T, by combining clusters that share mutations or children and by removing edges that are less likely until T meets all the constraints of a tree. This process follows these rules:

(the graphs on the left show an unfinished graph, while the tree on the right is a resolved tree)

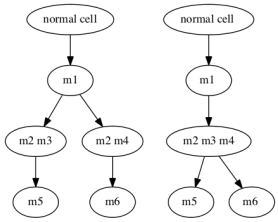
– If any cluster c in T does not have a parent and c is not the normal cell (the root of the tree), then an edge is added to c from the cluster that is most frequently the parent of c in the dataset S.

Figure 2: Resolve Forest



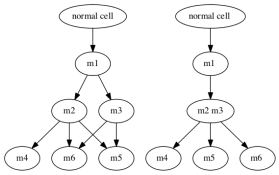
– If two clusters c_1 and c_2 both contain the same mutation m_i then c_1 and c_2 are combined.

Figure 3: Resolve Duplicate Mutation



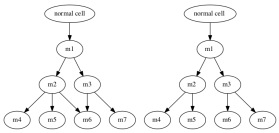
– If c_1 and c_2 both have cluster c_i as a child then they are in conflict. If there are no more than two clusters which are a child of c_1 but not c_2 or a child of c_2 but not c_1, then c_1 and c_2 are combined.

Figure 4: Resolve Shared Children



Otherwise evaluate how frequently c_1 and c_2 adhere to the sum rule given the current topology of the tree and the mutational frequencies from the dataset. If one cluster meets that condition less frequently than the other, remove the edge from it to c_i. If those values are the same, then randomly remove one edge.

Figure 5: Resolve Shared Children



– If there is a cycle in the graph, then collapse the cycle, so that all clusters in the cycle are combined into a single cluster.

Figure 6: Resolve Cycle (example 1)

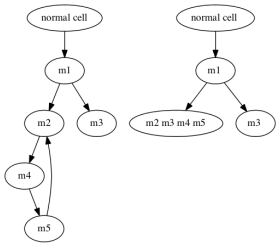
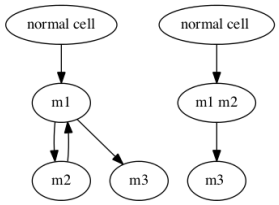


Figure 7: Resolve Cycle (example 2)



This method most closely resembles the majority rule method for creating classical phylogenetic consensus trees. I adapted the basic principles of majority rule to account for the unique properties of tumor phylogenies.

Example

Figure 8: Input Trees

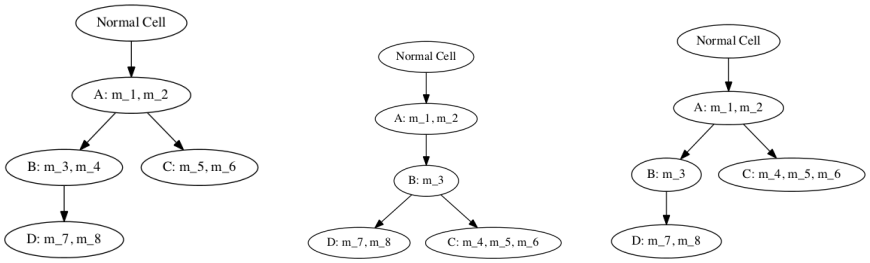
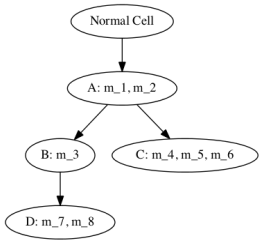


Figure 9: Majority Vote Consensus Tree



The **Distance Based Markov Chain** (DBMC) method uses Markov Chain Monte Carlo (MCMC) sampling to infer a distribution over phylogenies, where the consensus tree that is output is the MCMC sample with the lowest complete-data distance. At each iteration of the Markov chain two neighboring trees are compared, with the tree with the lower complete-data distance being selected.

The Markov chain can be initialized with a random tree or a given input tree. Initializing the Markov chain with the tree created by the majority vote method ensures that the Markov china is sampling trees that are similar to those in the dataset, but it also biases the results from the DBMC method to be more similar to the tree created by the Majority Vote method. Subsequent trees are created by randomly generating trees that are neighbors of the current tree.

The complete-data distance is calculated for both the current tree and its neighboring tree. If the complete-data distance of the neighboring tree is less than that of the current tree, then the neighboring tree becomes the current tree, if not, the current tree remains as the current tree. At each iteration of the Markov Chain the complete-data distance either decreases or remains the same.

The problem with this method is that it can become stuck at a locally optimal complete-data distance. In this case all trees that neighbor the current tree may have a greater complete-data distance, but there still exists a possible tree that has a lower complete-data distance. To solve this problem, I increase the number of Markov chains that explore the space of tumor phylogenies. These secondary chains explore the space in larger steps than the original chain. While the original chain only compares neighboring trees, the secondary chains compare trees that are k-distance neighbors.

In my simulations, I initialized 2-distance, 3-distance and 4-distance neighbors for the secondary chains. It is easier for these chains to move through the space of trees without getting stuck at local optima. At each iteration the trees from the secondary chains are compared to the tree from the original tree. If a tree from a secondary chain has a lower complete-data distance than the tree from the original chain, then that tree becomes the tree for the original chain and the previous tree from the original chain becomes the tree for the secondary chain.

The DBMC method takes into account both the topologies and clustering from the dataset, but also allows for the creation of trees, substructures and clustering, which may not exist in the dataset. This method does not take into account the mutational frequencies from the dataset. The disadvantage of this is that the DBMC method may output a tree which does not fit the mutational frequencies. Future work could also incorporate the mutational frequencies into evaluation of the tree at each iteration, so as to discourage selecting trees that may not reasonably fit all of the data.

Results and Discussion

Simulated Data

The set of trees S is created from simulated data. To create this, a true tumor phylogeny is randomly generated, by randomly partitioning m mutations into c clusters. A cluster containing the normal cell is assigned as the root of the tree, and then subsequent clusters are added to the tree by randomly assigning them a parent cluster from clusters already in the tree. The population frequencies are also randomly selected from within a range of values. The values are restricted to ensure that the population frequencies adhere to the sum rule (and therefore trees in the dataset are tumor phylogenies).

The normal cell has a population frequency f of 1, while the population frequencies of the subsequent clusters are created so that the population frequencies of each parent $cluster_p$ is greater than the sum of the population frequencies of all k of its children.

The population frequencies of $child_i$ of $cluster_p$ is restricted to the bounds, f_{child_i} must be $\leq f_p - \sum_{j=0}^{i-1} f_{child_j}$ and $\geq max(\frac{f_p - \sum_{j=0}^{i-1} f_{child_j}}{5}, .001)$.

$child_i$ refers to the current child of $cluster_p$, while $child_j$, where $0 \leq j \leq i-1$, refers to each child of $cluster_p$ which has already been given a population frequency. Therefore the possible range of population frequencies shrinks as each child of $cluster_p$ is assigned a population frequency.

The mutational frequencies for trees in S are generated by randomly drawing a coverage value from a Poisson distribution centered at the coverage of the true tree. The variant allele read count is drawn from Binomial(n,p), where n is the coverage value and p is the mutational frequency from the true tree. The mutational frequency is then calculated as variant allele read count/coverage. The true tree is then adjusted so that it still adheres to the sum rule given the new mutational frequencies. All clusters are also adjusted so that all mutations in the cluster have a mutational frequency that is no more than ten percent away from the population frequency (average mutational frequency) of the cluster.

Adjustments are made to the tree according to these rules:

– If the mutational frequency of a mutation is greater than 110% or less than 90% of the population frequency of the cluster that it is in, then it is moved out of that cluster. If there is another cluster that it is within 10% of the population frequency of, then it is added to that cluster. Otherwise it is moved to a newly created cluster. If the mutational frequency is less than the population frequency of the original cluster then it becomes the child of the original cluster. If the mutational frequency is more than the population frequency of the original cluster then it becomes the parent of that cluster.

– If the population frequency of a cluster is greater than the population frequency of its parents then those clusters are moved. If the smaller population frequency is less than 80% of the larger then the child cluster becomes the parent of its original parent. Otherwise the child cluster and the parent cluster switch.

– If at any point the sum of the population frequencies of the child clusters is greater than the population frequency of their parent, then the child cluster with the lowest population frequency is moved to be the child of the child cluster with the highest population frequency.

– Lastly, if there is a parent and child cluster that have population frequencies that are within 10% of each other then combine those clusters.

These adjustments ensure that all trees in S are tumor phylogenies.

Evaluating Consensus Trees

I evaluated the consensus methods by comparing the consensus trees to the true tree, which the input trees are based on. By examining how the consensus trees and input trees compare to the true tree, I can evaluate whether the consensus tree better approximates the true tree than the input trees. There are two main aspects of the tree to compare - the clustering of mutations and the topology/ancestral relationships in the tree. The rand index measures the similarity between two data clusterings. The topology of the consensus tree can be evaluated by identifying the ancestral similarity, which is the fraction of ancestral relationships from the true tree that are present in the consensus tree. The distance between the consensus tree and the true tree evaluates both the similarity in clustering and the similarity in topology between the two trees.

rand index: Given the set of mutations $M = \{m_1, ..., m_m\}$ and two clusterings of M to compare, $X = \{X_1, ..., X_n\}$, a partitioning of M into n_x clusters and $Y = \{Y_1, ..., Y_n\}$, a partitioning of M into n_y clusters, define the following:

a, the number of pairs of mutations in M that are in the same cluster in X and in the same cluster in Y

b, the number of pairs of mutations in M that are in different clusters in X and in different clusters in Y

c, the number of pairs of mutations in M that are in the same cluster in X and in different clusters in Y

d, the number of pairs of mutations in M that are in different clusters in X and in the same cluster in Y

rand index = $\frac{a+b}{a+b+c+d}$ [10]

ancestral similarity: The ancestral pairings of a tree are the pairs of mutations m_1 and m_2 where m_2 is ancestral to m_1 in the tree. The ancestral similarity is the fraction of ancestral pairings from the true tree that are also present in the consensus or input tree.

These evaluation methods assume knowledge of the true tree. I also experimented with evaluating the consensus trees assuming that the true tree is unknown. I did this by calculating the beta probability of the evolutionary relationships in the consensus trees using the mutational frequencies from the dataset.

beta probability: We can calculate the probability of ancestral relationships in T by calculating the probability that T adheres to the sum rule, specifically the probability that the population frequency of any parent p (F_p) in T is greater than the sum of the population frequencies ($F_c = \sum_{i=0}^{k} F_i$) of its k children. We can approximate the population frequencies of a cluster i as Beta($c_i + 1, d_i + 1$), where c_i is the number of reads that cover the mutations in i and that contain the variant alleles and d_i is the number of reads that cover the mutations in i and that contain the reference alleles [4]. We can compute Pr($F_p \geq F_c$) using the method described by Cook [2]. The beta probability for a tree is the product of the beta probability for each parent and its children in the tree given a set of mutational frequencies. To calculate the beta probability of a consensus tree I used the average mutational frequency for each mutation from the input trees.

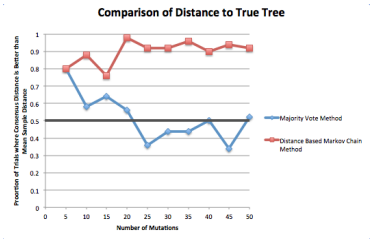
Results on Simulated Data

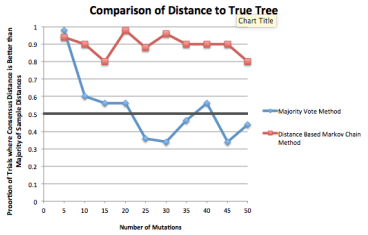
In each trial, the simulated true trees had 5 clusters and coverage of 100, and the dataset contained 5 input trees. The input trees varied from 5 to 50 mutations, and for each mutation value I ran 50 simulations. This simulation mimics the case of creating a consensus tree from the output of multiple tumor phylogeny reconstruction methods.

In each simulation I calculated the distance to the true tree, ancestral similarity with the true tree, and rand index with the true tree for both consensus trees and every input tree, as well as the beta probability of both consensus trees.

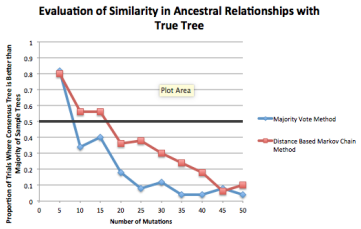
As the number of mutations increased, so did the variability in the data. With more mutations there were more possible errors and the input trees were more different from each other. I would expect that as the number of mutations, and therefore the variability, increases, the consensus methods will be less able to approximate the true tree. For each mutation number and evaluation method, I looked at whether that value for the consensus tree was better than the average value for the input trees and whether that value for the consensus tree was better than the majority of those values from the input trees. As the data became more variable the consensus trees performed worse compared to the input trees in most of the metrics.

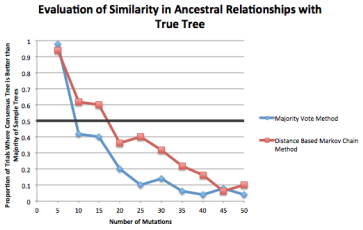
I found that the distance based markov chain method had lower distance to the true tree (in terms of both average and majority of input trees) in 80-98% of all trials. The majority vote method did not perform as well in this metric. It only consistently had a lower distance in the majority of tests when there were 20 or fewer mutations.



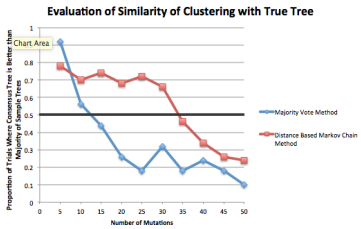


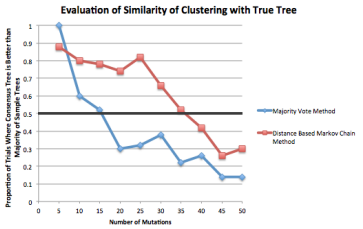
The consensus trees were least similar to the true tree (compared to the input trees) in terms of similarity in ancestral relationships. The DBMC method only performed better than the input trees when there were 15 or fewer mutations, while the Majority Vote method only performed better than the input trees when there were 5 mutations.





The consensus trees also became worse (in comparison to the input trees) in similarity of clustering with the true trees as the number of mutations increased. The DBMC method performed better than the input trees when there were 30 or fewer mutations, while the Majority Vote method only performed better than the input trees when there were 10 or fewer mutations.





These results indicate that, when the input trees do not vary extremely and the number of mutations is fewer, the consensus trees are able to identify substructures that are present in the true tree and create consensus trees that are similar to the true tree in terms of topology and clustering. These results also show that the DBMC method created a consensus tree that was more similar to the true tree in all three respects than the consensus tree created by the Majority Vote method. This difference was especially pronounced as the number of mutations (and the variability of the input trees) increased.

I was not able to make meaningful inferences from the beta probabilities. I compared the beta probabilities of the two consensus methods, but which beta probabil-

ity was higher alternated between the two methods with no distinguishable pattern. The probability did not correlate with which method had a higher ancestral similarity, higher clustering similarity or lower distance to the true tree. This suggests this method for calculating the beta probability, as it is being done currently, is not an informative measure for evaluating the consensus trees when the true tree is unknown.

Case study results

I applied the tumor phylogeny consensus methods to trees output by PhyloWGS [3], PhyloSub [6], AncesTree [4], CITUP [7] and trees created by semi-manual reconstruction by Schuh *et al.*. These trees used data chronic lymphocytic leukemia (CLL) [11] and renal cell carcinoma tumors (RMH, EV) [5].

I was unable to thoroughly evaluate the consensus methods with this data, as most of the tumor phylogenies were the same. Some of the methods output a tumor phylogeny that only included a smaller subset of the mutations. In this case, I reduced all trees in the dataset so that they only included mutations found in all of the trees. This was necessary, as the consensus tree contains the largest subset of mutations that is present in all input trees. Due to the low variability of the data, the consensus trees output by both methods often exactly matched trees in the input dataset. Future tests will use data that produces more uncertain and/or variable tumor phylogenies.

Conclusion

Consensus trees can provide beneficial information, as they can resolve differences between tumor phylogenies and may be more similar to the true tumor phylogeny, in terms of clustering and topology, than input trees. The results of my tests show that when data is not too variable, the consensus methods were able to closely approximate the true tree. The poorer performance of the consensus methods when there were more mutations is likely because the input trees had too much error. The discrepancies in the consensus trees performance in terms of clustering similarity, ancestral similarity and distance evaluations suggest that the input trees were too variable to be a good representation of the data from the true tree. This would explain why the DBMC method has a lower distance to the true tree, yet was not similar in terms of ancestral relationships and/or clustering.

With the real data the problem was the opposite. For almost all of the real datasets, the majority of the input trees were the same, and therefore the consensus tree also matched those trees.

I have made some modifications to the creation of my simulated dataset to reduce the variability. Instead of changing all of the mutational frequencies from the true tree for each input tree, only a certain proportion of frequencies are changed. This means that there are fewer changes between the true tree and each input tree. In the future I plan to run more trials with this modified dataset.

I also plan to continue working on creating a better method to evaluate the consensus trees when the true tree is not known. Currently the beta probability, with average mutational frequencies from the dataset, is used to evaluate the consensus tree. Modifications to this calculation, particularly to how the mutational frequencies from the

dataset are used/averaged, could make it a more meaningful metric. An improved evaluation method will also be important for future tests on real data. I do not have a method for comparing trees with different numbers of mutations. A new metric for this comparison could change how I define a consensus tumor phylogeny.

The Majority Vote method and Distance-Based Markov Chain method provide two new approaches to creating consensus tumor phylogenies. These approaches address the unique properties of tumor phylogenies, and are more suited to creating consensus tumor phylogenies than classical methods for creating consensus phylogenies. By further incorporating the mutational frequencies the methods could create better consensus trees.

References

[1] Brosnan, J. A. and Iacobuzio-Donahue, C. A. (2012). A new branch on the tree: Next-generation sequencing in the study of cancer evolution. *Seminars in Cell and Developmental Biology*, 23(2):237–242.

[2] Cook, J. D. (2005). Exact calculation of beta inequalities. Technical report, UT MD Anderson Cancer Center Department of Biostatistics.

[3] Deshwar, A. G., Vembu, S., Yung, C. K., Ho, J. G., Stein, L., and Morris, Q. (2015). Phylowgs: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):35.

[4] El-Kebir, M., Oesper, L., Acheson-Field, H., and Raphael, B. J. (2015). Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):62–70.

[5] Gerlinger, M., Horswell, S., Larkin, J., Rowan, A. J., Salm, M. P., Varela, I., Fisher, R., McGranahan, N., Santos, C. R., Martinez, P., Phillimore, B., Begum, S., Rabinowitz, A., Spencer-Dene, B., Gulati, S., Bates, P. A., Stamp, G., Pickering, L., Gore, M., Nicol, D. L., Hazell, S., Futreal, A. P., Stewart, A., and Swanton, C. (2014). Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature Genetics*, 46:225–233.

[6] Jiao, W., Vembu, S., Deshwar, A. G., Stein, L., and Morris, Q. (2014). Inferring clonal evolution of tumors from single nucleotide somatic mutations. *Bioinformatics*, 15(35).

[7] Malikic, S., McPherson, A. W., Donmez, N., and Sahinalp, C. S. (2014). Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31(9):1349–1356.

[8] Navin, N. E. and Hicks, J. (2010). Tracing the tumor lineage. *Molecular Oncology*, 4:267–283.

[9] Nowell, P. (1976). The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28.

[10] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

[11] Schuh, A., Becq, J., Humphray, S., Alexa, A., Burns, A., Clifford, R., Feller, S. M., Grocock, R., Henderson, S., Khrebtukova, I., Kingsbury, Z., Luo, S., McBride, D., Murray, L., Menju, T., Timbs, A., Ross, M., Taylor, J., and Bentley, D. (2012). Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood*, 120(20):4191–4196.

[12] Yuan, K., Sakoparnig, T., Markowetz, F., and Beerenwinkel, N. (2015). Bitphylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biology*, 6(36).