Establishing subpopulation structure in tumors by clustering phylogenies

Jeff Wintersinger
Graduation date: Feb. 2016
CSC 2515 Project Proposal
March 5, 2015

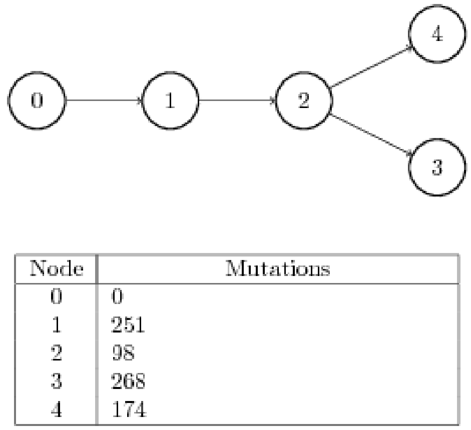
Tumours consist of multiple genetically distinct subpopulations of cells that have evolved from common ancestors. Each subpopulation contains mutations promoting growth and metastasis (1,2). Some cancer-associated mutations are present in all cancerous cells, but many occur in only a subset (3). By drawing on whole-genome sequencing data from tumours, one can infer "tumour phylogenies" that reconstruct the sequence of mutations transforming normal cells into cancerous ones. These phylogenies support classification of tumour subtypes and characterization of key steps in tumour evolution, such as the development of drug resistance and progression from nonmetastatic to metastatic cells (4). By examining the evolutionary progression of tumours from multiple patients, one can understand which variants are “driver” mutations critical for cancer's development, and which are “passenger” mutations with little functional significance. Beyond improving our understanding of cancer's progression, this knowledge can inform development of better treatments.

The Morris lab recently developed PhyloWGS (5), a method for inferring tumour phylogenies from both single nucleotide variants (SNVs) and copy number variations (CNVs) (Fig. 1). PhyloWGS does not, however, produce single consensus trees. Instead, it yields as output multiple trees sampled via Markov-chain Monte Carlo from the distribution of trees consistent with the observed frequencies of SNVs and CNVs (5), using a tree-structured stick-breaking (TSSB) process prior (6). Though each resulting tree has an associated likelihood, the highest-likelihood tree may poorly reflect the model's consensus as to tree structure. For example, multiple trees may be equally consistent with the observed variant allele frequencies (Fig. 2), with the highest-likelihood tree reflecting only one such possibility. In the case of two equally consistent structures, each should be reflected in approximately half of the trees sampled from the distribution of possible trees.

My project will endeavor to establish “consensus trees,” indicating which portions of phylogenetic reconstructions are well-supported by the model (i.e., which remain constant across sampled trees), and which exhibit uncertainty (i.e., which vary across sampled trees). This problem is shared by other tree-generating methods using TSSB process priors (6). Estabilshing consensus trees necessitates two aims. In the first, I will convert each tree to a matrix representation suitable for clustering. Each tree consists of nodes containing sets of tumor mutations clustered by frequency, with each node thus representing a subpopulation of tumor cells. Given mutations A and B, there exist four possible relationships between them: A and B may lie in the same subpopulation; A may be in an ancestral subpopulation of B's subpopulation; B may be in an ancestral subpopulation of A's; or A and B may exist in subpopulations lying on entirely distinct branches on the phylogenetic tree. Thus, a tree can be uniquely represented by binary vectors denoting the relationship between every pair of constituent mutations.

After producing tree-representing matrices, I will cluster them to determine the dominant tree structures extant in our samples. This differs from normal tree clustering because I cannnot simply cluster trees by structure—to fully represent tree diversity, I must also cluster based on assignment of mutations to subpopulations. To cluster, I will develop a non-parametric method using variational Bayes techniques, avoiding the need to specify the desired number of clusters *a priori* in parametric methods such as multinomial clustering. For comparison, I may also explore Markov-chain Monte Carlo methods using Gibbs sampling.

In my second aim, once I have clustered the matrices, I will create consensus trees visualizing the structure of each cluster. The clusters will provide insight into how much divergence exists between the sampled trees, with respect to both subpopulation structure and mutation assignment to subpopulations. My goal, then, will be to create tree visualizations depicting the structure of each cluster. This will be challenging—though each matrix uniquely represents a tree, its associated cluster will contain a collection of matrices that must be rendered as a single tree. One possible solution lies in the majority-rules consensus method established in phylogenetics (7), in which tree branches are drawn only if they exist in at least half the trees composing each cluster.



Node	Mutations
0	0
1	251
2	98
3	268
4	174

Figure 1: Example phylogenetic tree indicating subpopulation structure in a tumour, as well as the number of mutations associated with each subpopulation. Node 0 represents the somatic cells from which the tumour descended, meaning no tumour-specific mutations are associated with it. Node 1 represents the originating tumour cell population, meaning that all tumour cells share its mutations. From this subpopulation, three additional subpopulations descended.

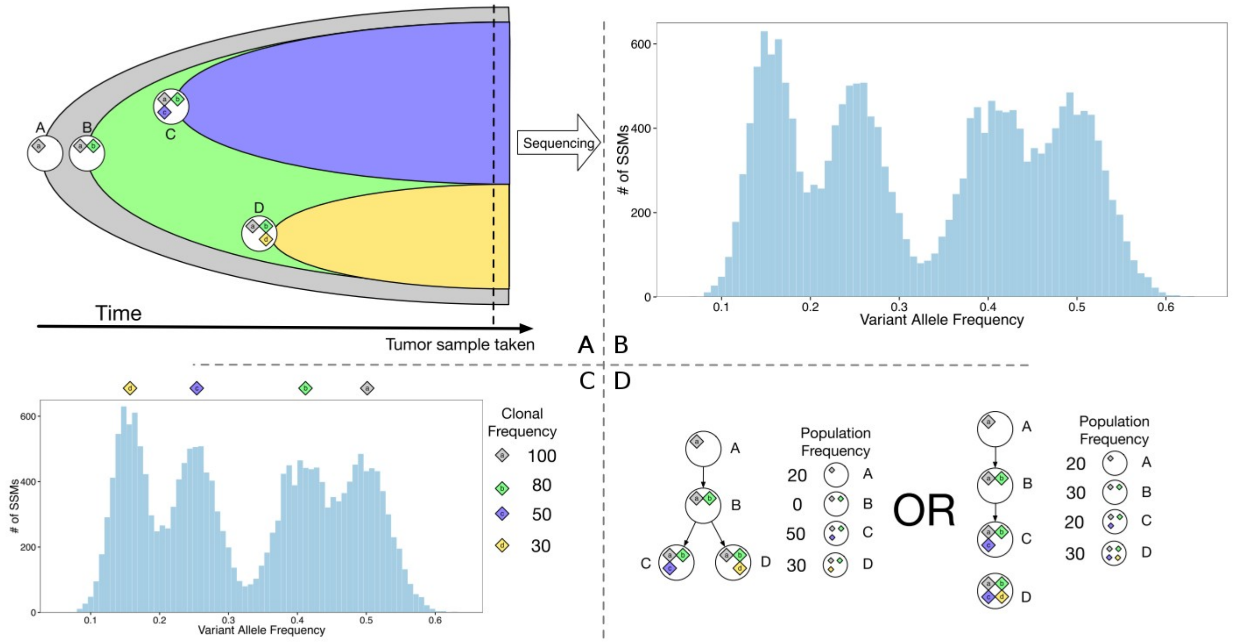


Figure 2: Adapted from (5). **A.** In a given tumour, subpopulations A, B, C, and D exist. The diamonds within each subpopulation's circle indicate sets of mutations associated with that subpopulation. Though only subpopulations C and D are extant when the sample is taken, ancestral subpopulations A and B can be inferred from variant allele frequencies. **B.** The observed distribution of allele frequencies for tumour-specific variants. **C.** Four distinct clusters of allele frequencies exist, representing alleles assumed to be present in varying combinations within the four subpopulations. Clonal frequencies are twice the observed variant allele frequencies, as each mutation is assumed to be heterozygous, with only one copy thus present in every cell. **D.** Two subpopulation structures are equally consistent with the observed variant allele frequencies. Though panel A reveals the branched structure to be correct, the linear structure is no less valid for the given allele frequencies. Each tree structure should correspond to approximately half the trees sampled from the distribution of possible trees.

References

1. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100(1):57–70.

2. Hanahan D, Weinberg R a. Hallmarks of cancer: the next generation. Cell. 2011 Mar;144(5):646–74.

3. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012;366(10):883–92.

4. Subramanian A, Shackney S, Schwartz R. Tumor Phylogenetics in the NGS Era: Strategies, Challenges, and Future Prospects. Wu W, Choudhry H, editors. Gener Seq Cancer Res. 2013;335–57.

5. Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q. PhyloWGS: Reconstructing subclonal composition and evolution from whole genome sequencing of tumors. Genome Biol. 2015;16(1):35.

6. Ghahramani Z, Jordan MI, Adams RP. Tree-structured stick breaking for hierarchical data. Advances in Neural Information Processing Systems. 2010. p. 19–27.

7. Felsenstein J. Inferring phylogenies. Sunderland, Mass: Sinauer Associates; 2004. 664 p.