SubClonal Hierarchy Inference from Somatic Mutations: automatic reconstruction of cancer evolutionary trees from multi-region next generation sequencing

Noushin Niknafs[1], Violeta Beleva-Guthrie[1], Daniel Q. Naiman[2], Rachel Karchin[1,3,*]

1 Department of Biomedical Engineering and Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD, USA

2 Department of Applied Math and Statistics, Johns Hopkins University, Baltimore, MD, USA

3 Department of Oncology, Johns Hopkins Medical Institutions, Baltimore, MD, USA

∗ **E-mail: karchin@jhu.edu**

1 Abstract

Recent improvements in deep next-generation sequencing of tumor samples and the ability to identify somatic mutations at low allelic fractions have opened the way for new approaches to model the evolution of individual cancers. The power and utility of these models is increased when tumor samples from multiple sites are sequenced. Temporal ordering of the samples may provide insight into the etiology of both primary and metastatic lesions and rationalizations for tumor recurrence and therapeutic failures. Additional insights may be provided by temporal ordering of evolving subclones – cellular subpopulations with unique mutational profiles. Current methods for subclone hierarchy inference tightly couple the problem of temporal ordering with that of estimating the fraction of cancer cells harboring each mutation. Here, we present a new framework that includes a rigorous statistical hypothesis test and a collection of tools that make it possible to decouple these problems, which we believe will enable substantial progress in the field of subclone hierarchy inference. The methods presented here can be flexibly combined with methods developed by others addressing either of these problems. We provide tools to interpret hypothesis test results, which inform phylogenetic tree construction, and we introduce the first genetic algorithm designed for this purpose. The utility of our framework is systematically demonstrated in simulations. For most tested combinations of tumor purity, sequencing coverage, and tree complexity, good power (≥ 0.8) can be achieved and Type 1 error is well controlled when at least three tumor samples are available from a patient. Using data from three published multi-region tumor sequencing studies of (murine) small cell lung cancer, acute myeloid leukemia, and chronic lymphocytic leukemia, in which the authors reconstructed subclonal phylogenetic trees by manual expert curation, we show how different configurations of our tools can identify either a single tree in agreement with the authors, or a small set of trees, which include the authors' preferred tree. Our results have implications for improved modeling of tumor evolution and the importance of multi-region tumor sequencing.

2 Introduction

The clonal evolution hypothesis in cancer states that cancer genomes are shaped by numerous rounds of cellular diversification, selection and clonal expansion [Nowell, 1976, Greaves and Maley, 2012]. Recent methods to characterize tumor clonal evolution can be divided into two broad classes – sample tree reconstruction and subclone tree reconstruction. The first class of methods models the history of clonal evolution in an individual as a phylogenetic tree with leaves being the individual's tumor samples, yielding a relative temporal ordering and estimate of divergence between the samples [Kostadinov et al., 2013, Gerlinger et al., 2014, Johnson et al., 2014]. The second class aims at reconstructing the history of clonal evolution as a tree, which summarizes lineage relationships between cellular subpopulations [Jiao et al., 2014, Hajirasouliha et al., 2014, Strino et al., 2013].

Until single-cell sequencing data is widely available, accurate high resolution modeling of tumor

evolution will likely remain exceedingly difficult, if not impossible. On the other hand, it is well accepted that tumor samples are comprised of a few cellular subpopulations or *subclones*. Each of these subclones emerge from a parental population of cells by acquiring additional somatic mutations, and cells within each subclone can be assumed to to be homogeneous. Modeling of subclone evolution often involves estimating the fraction of cancer cells harboring each somatic mutation *i.e.*, somatic mutation *cellularity*, which can be inferred from next generation sequencing read count data. For example, PyClone [Roth et al., 2014] employs a Markov Chain Monte Carlo method to identify groups of mutations with similar cellularities, and SciClone [Miller et al., 2014] uses variational Bayes mixture models to cluster somatic mutations by their read count frequencies, which can be a proxy for cellularities.

Most recently, methods that couple the problems of somatic mutation clustering and phylogenetic reconstruction have emerged. PhyloSub applies a tree-structured stick breaking process that introduces tree-compatible cellularity values for mutation clusters [Jiao et al., 2014]. A combinatoric approach based on an approximation algorithm for binary tree paritions [Hajirasouliha et al., 2014] and a mixture deconvolution algorithm [Strino et al., 2013] have also been developed. However to our knowledge, most recently published studies of multi-region tumor sequencing continue to employ manual curation to construct a subclone phylogeny, after mutation cellularity has been estimated computationally [Ding et al., 2012, McFadden et al., 2014, Schuh et al., 2012].

We propose that progress in methods to reconstruct subclonal phylogenies will be substantially enabled by decoupling the problems of temporal ordering of subclones from that of mutation cellularity estimation. The SubClonal Hierarchy Inference from Somatic Mutations (SCHISM) framework described here can incorporate a variety of methods to estimate the cellularity of individual mutations, the cellularity of mutation clusters, and to build phylogenetic trees. First, we derive a novel mathematical formulation of assumptions about lineage precedence and lineage divergence in tumor evolution that have been fundamental to other subclone tree reconstruction methods [Strino et al., 2013, Jiao et al., 2014, Hajirasouliha et al., 2014]. Lineage precedence is modeled in terms of a statistical hypothesis test, based on a generalized likelihood ratio. Hypothesis test results are combined with lineage divergence assumptions and formulated as a scoring function that can be used to rank tree topologies, generated by a phylogenetic algorithm. In this work, we designed an implementation of genetic algorithms to build phylogenetic trees. However, the scoring function can also be combined with other approaches to phylogenetic tree reconstruction. The hypothesis test can be combined with any method to estimate mutation or cluster cellularities to infer their temporal orderings.

We use simulations to evaluate the power of the hypothesis test and show that for many combinations of tumor purity, sequencing coverage, and phylogenetic tree complexity, the hypothesis test has good power (≥ 0.8) and Type 1 error is well controlled, when at least three samples from a patient are available. The simulations also confirm that the problem of subclonal phylogenetic tree reconstruction is underdetermined in many settings when the tumor sample count per individual is smaller than the number of subclones *i.e*, nodes in the phylogenetic tree. In these cases, we may see that the genetic algorithm identifies multiple equally plausible phylogenetic trees. However, when the problem is sufficiently determined, in general when the number of samples equals or exceeds the tumor sample count, the genetic algorithm reliably reconstructs the true tree, in most combinations listed above.

Using data from three published multi-region tumor sequencing studies of murine small cell lung cancer [McFadden et al., 2014], acute myeloid leukemia [Ding et al., 2012] and chronic lymphocytic leukemia [Schuh et al., 2012], we show how SCHISM can be configured with a variety of inputs. For all samples in these three studies, SCHISM identified either a single tree in agreement with the tree reconstructed manually by the authors, or a small set of trees, which include the authors' published tree.

3 Results

3.1 Simulations

3.1.1 Generalized likelihood ratio hypothesis test

The hypothesis test yielded good power on average and Type 1 error was well controlled (Figure 1). Power improved as the number of samples per individual increased. As the number of nodes in the subclone tree increased, yielding a more complex tree, more samples were required to achieve the same level of power. Even at the lowest purity level (50%) included in our experiments, good power (≥ 0.8) was achieved with 1000X coverage and three or more samples.

3.1.2 Automated subclone tree reconstruction

The performance of the genetic algorithm (GA) used for tree reconstruction varied substantially, depending on the simulation inputs: number of tumor samples, node count and topology of the true tree, mutation cluster cellularity values, tumor purity, and sequencing coverage (Figure 2). Given sample count exceeding or equal to node count, the GA most frequently (with probability ≥ 0.5) identified the true tree or a pair of high scoring trees that included the true tree. This probability increased to ≥ 0.75 at high purity (0.9) and coverage (1000X). As expected, simpler trees, *e.g.*, 3- or 4-node trees, were frequently identified even when the sample count was small. As trees grew more complex, a larger sample count was required, and even the most complex trees in the simulation, which had 8 nodes, were identified frequently when 10 samples were available. However, we also identified combinations of inputs for which the GA had limited success in finding the true tree. We decomposed the performance of the GA into two stages. In Stage 1, we assessed whether the tree reconstruction problem was sufficiently determined by our inputs, meaning that a single high-scoring tree or a pair of two high-scoring trees was identified. The GA was more likely to fail in Stage 1 when sample count was smaller than node count (Figure 2 A1,B1). Furthermore, the settings of purity and sequencing coverage used in our simulations had less of an effect on Stage 1 success than sample count and tree node count. In Stage 2, we assessed whether the single high-scoring or pair of high-scoring trees included the true tree. Samples counts ≥ 5 had the most stable Stage 2 success rates, and the correct tree was recovered with increasing frequency, given higher sample counts, coverage, and purity. As expected, probability of Stage 2 success was higher for trees with smaller node counts. Our estimates of Stage 2 success were noisy when sample count was small and node count high. This behavior was a result of higher failure rates at Stage 1 under these conditions (Figure 2 A2,B2).

3.2 Multi-sample sequencing studies

Recent studies of small cell lung cancer (SCLC) in mice [McFadden et al., 2014], acute myeloid leukemia (AML) [Ding et al., 2012], and chronic lymphocytic leukemia (CLL) [Schuh et al., 2012] attempted to infer the subclonal phylogeny underlying tumor progression, based on sequencing of multiple tumor samples. All of the studies applied computational methods to cluster somatic mutations. In some cases mutation cluster cellularities were provided, while in others read counts or cluster mean variant allele fraction was provided. The authors did not use computational methods to reconstruct the subclonal phylogenies.

We applied SCHISM to these datasets, using a variety of configurations. In cases where mutation cluster cellularities were available [McFadden et al., 2014], we used the hypothesis test (Section 5.2.4) on pairs of mutation clusters. If mean variant allele fraction for clusters was available [Ding et al., 2012], we inferred cellularity as described in (Section 5.3.1 Equation 28) and again used the hypothesis test on pairs of mutation clusters. When only read counts and mutation cluster assignments were available [Ding et al., 2012, Schuh et al., 2012], we used our own naive estimator to derive cellularity values (Section 5.3.1), and applied the hypothesis test to pairs of mutations. For all configurations, we constructed a precedence order violation (POV) matrix for all pairs of mutations (Section 5.2.3) or all pairs of mutation clusters

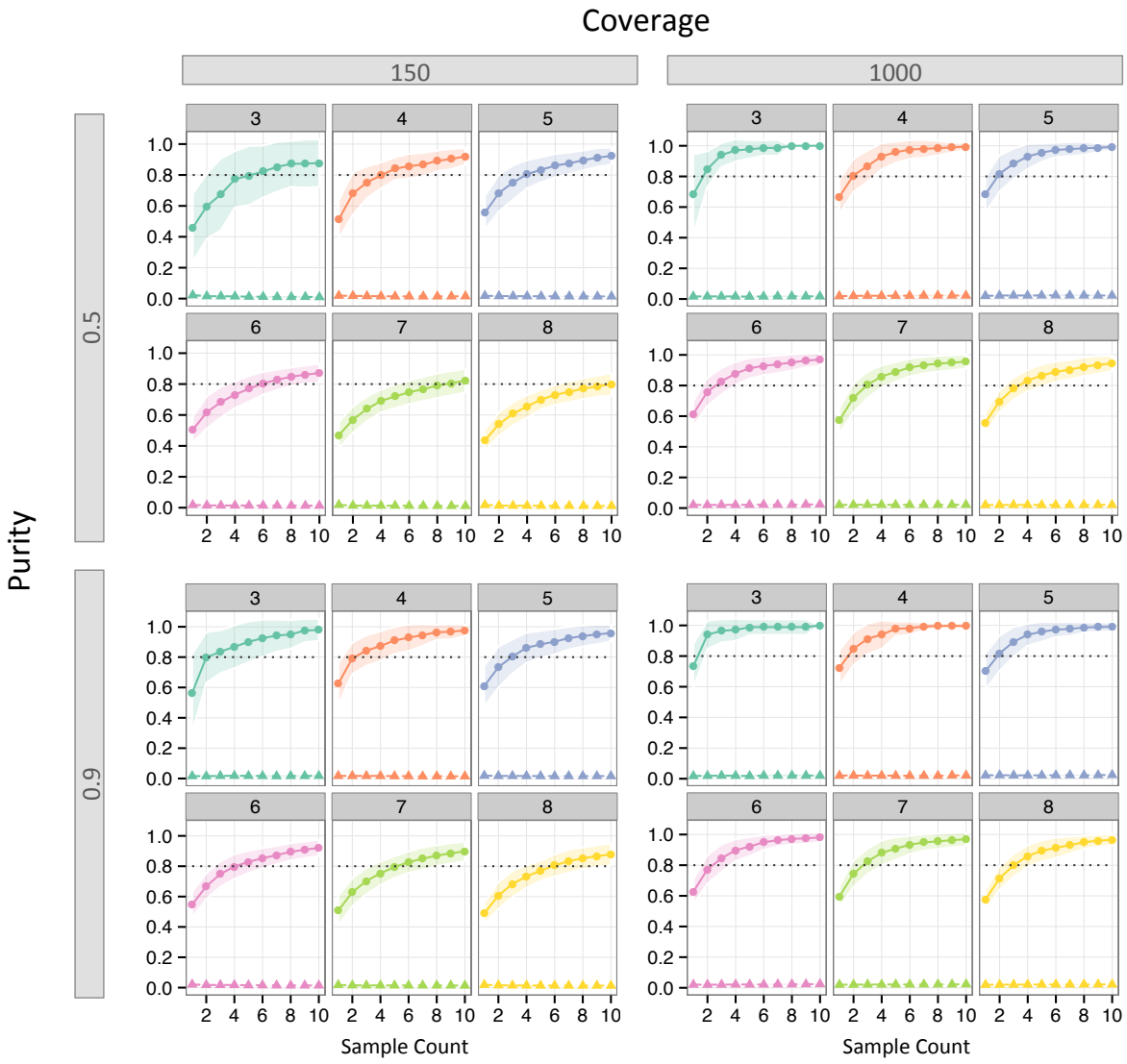

Figure 1: **Power and Type 1 error of hypothesis test on simulated data.** For each combination of coverage and purity, results are shown for trees with node counts from three to eight. Each curve was computed by taking the mean over all *instances* and all replicates for each node count. Curves with circular marks show power and curves with triangular marks show Type 1 error. Transparent coloring indicates $\pm 1SE$. Dotted line indicates power=0.8.

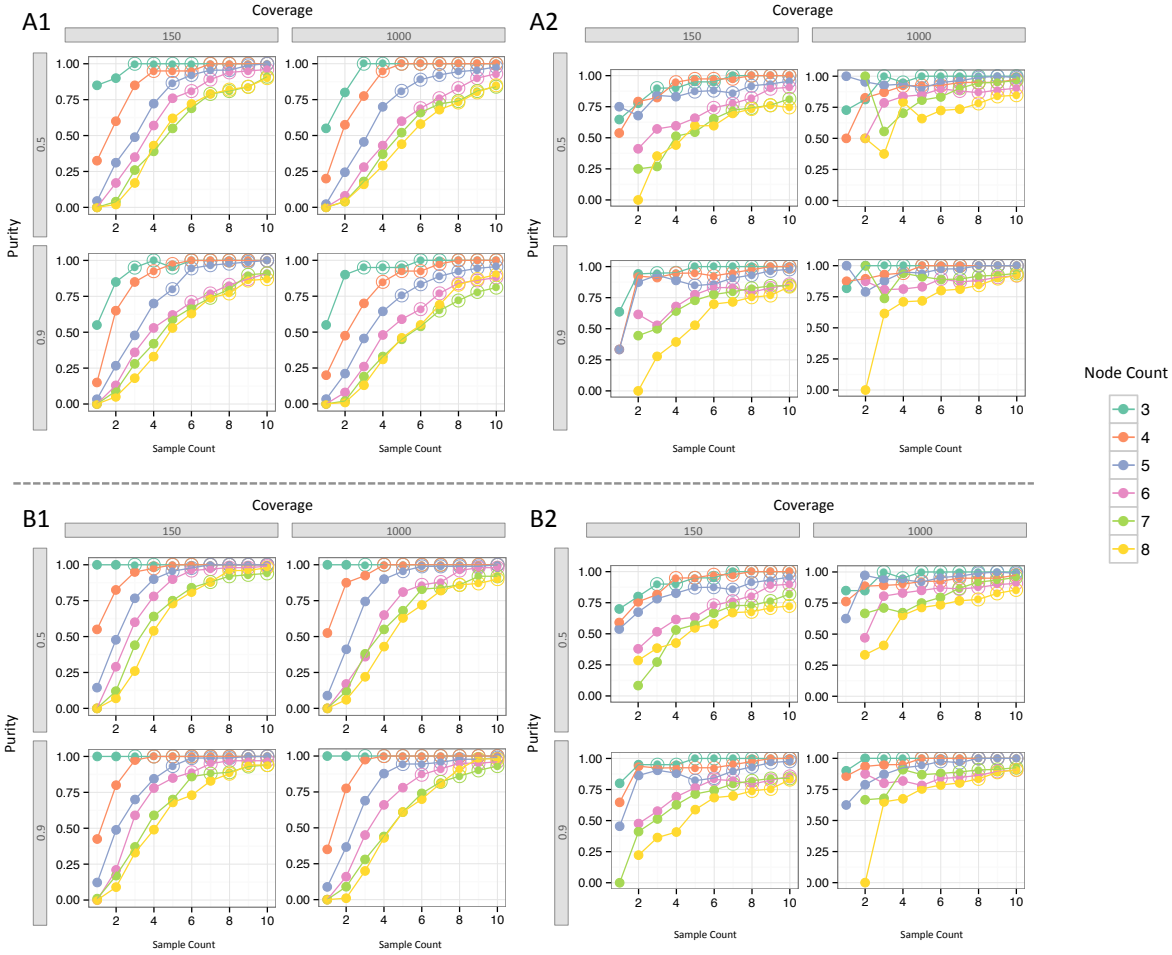

Figure 2: **Performance of the genetic algorithm evaluated in two stages.** Stage 1: Fraction of simulation runs where the genetic algorithm's fitness function identified either a single best tree (A1) or two best trees (B1). Stage 2: Given success in stage 1, fraction of simulation runs where the correct tree was either the single highest scoring tree (A2) or one of the top two highest scoring trees (B2). For each combination of coverage and purity, results are shown for trees with node counts from three to eight. Simulations where sample count $\geq$ node count are marked by a double circle.

(Sections 5.2.4 and 5.4) and ran the genetic algorithm. This approach consistently identified phylogenies identical to those manually constructed by the authors as either the single highest scoring tree or among a small set trees tied for the highest score, in underdetermined cases.

3.2.1 Murine small cell lung carcinoma

This study sequenced small cell lung cancer (SCLC) tumor samples from a cohort of transgenic mice with lung-specific Trp53 and Rb1 compound deletion. The full study included whole exome sequencing (WES) (150X coverage) of 27 primary tumors and metastases from six individual animals [McFadden et al., 2014]. For three of these animals (with identifiers 3588, 3151, and 984), a subclonal phylogenetic tree was manually built by the authors, after using ABSOLUTE for mutation clustering and cellularity estimation.

Using the hypothesis test on the ABOSLUTE cluster mean cellularities (Figure 5 [McFadden et al., 2014]), we generated a cluster-level precedence order violation matrix (CPOV) (Section 5.2.4) and a cluster cellularity by sample matrix for each mouse (Section 5.3.2).

Animals 3588 and 3151 had data available for one primary and two metatastic tumors. For animal 3588, SCHISM identified an 8-node single highest scoring tree and that tree (Figure 3A) was identical to the authors' manually built tree. For animal 3151, SCHISM identified six 9-node highest scoring trees, and one of these trees was identical to the authors' tree. While we consider this problem to be insufficiently determined, interestingly the trees shared a significant number of lineage relationships, and the main discrepancies among trees were the parental lineage for Clone3 and Clone2b (using notation from [McFadden et al., 2014]) (Figure 3B). For animal 984, data was available for one primary one metastatic tumor. SCHISM identified six 7-node highest scoring trees (*i.e.,* underdetermined problem), and one of these trees was identical to the authors' tree (Figure 3C).

3.2.2 Chronic lymphocytic leukemia

This longitudinal study of subclonal evolution in B-cell CLL tracked three patients (CLL 003, CLL006, CLL077) over a period of up to seven years [Schuh et al., 2012]. For each patient, five longitudinal peripheral blood samples were collected, and each sample was whole-genome sequenced. For selected somatically mutated sites, they further applied targeted deep sequencing (at reported 100,000X coverage). K-means clustering and expert curation were used to infer mutation clusters and subclonal phylogenetic trees were manually built.

We used mutation cluster assignments and read counts from targeted deep sequencing, for each mutation in each sample (Tables S6, S7, or S8 [Schuh et al., 2012]). Purity was estimated by identifying the mutation cluster with the maximum mean variant allele frequency in each sample. Next, based on purity and read count, we estimated cellularity (and standard error) for each mutation in diploid or copy number=1 regions (Section 5.3.1). The hypothesis test was performed for each pair of mutations (Section 5.2) and a cluster precedence order violation (CPOV) matrix was constructed, using the voting scheme described in Section 5.4.

For patients CLL003 and CLL077, SCHISM identified a single 4-node highest scoring tree that was identical to the authors' manually built tree (Figure 4A,C). For patient CLL006, two 5-node highest scoring trees were identified, and one was identical to the authors' tree (Figure 4B).

3.2.3 Acute myeloid leukemia

This study of acute myeloid leukemia (AML) relapse consisted of whole genome sequencing for primary and relapse samples from eight patients (AML1, AML15, AML27, AML28, AML31, AML35, AML40, AML43) [Ding et al., 2012]. For AML1, the authors identified mutation clusters with MClust [Fraley and Raftery, 2006] and manually constructed a subclonal phylogeny. For the other patients, mutation cluster

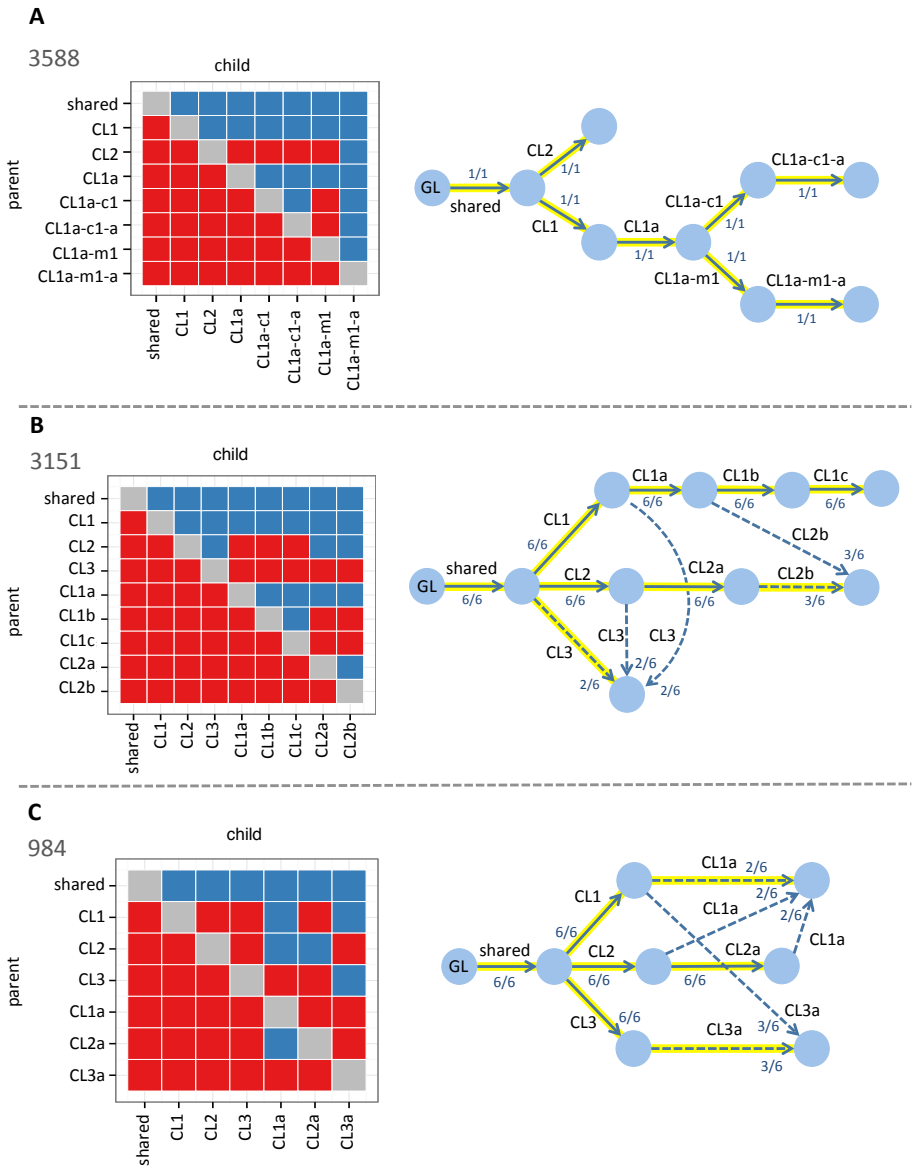

Figure 3: **Reconstruction of subclonal phylogenies in murine models of SCLC. A. Animal 3588.** SCHISM identified a single highest scoring 8-node tree using one primary and two metastatic tumors. **B. Animal 3151.** Six highest scoring 9-node trees were identified using one primary and two metastatic tumors. **C. Animal 984.** Six highest scoring 7-node trees were identified using one primary and one metastatic tumor. Solid arrows represent lineage relationships shared by all six trees and dashed arrows represent lineage relationships shared by only a subset of the trees. Each arrow is labeled with the fraction of highest scoring trees that include the lineage relationship. Highlighted arrows indicate the tree manually constructed by the study authors. GL = germline state. Cluster precedence order violation (CPOV) matrices are shown to the left of each tree. Columns and rows represent subclones (or Clones in the terminology of [McFadden et al., 2014]). Each red square represents a pair of subclones (I,J) for which the null hypothesis that I could be the parent of J was rejected. Each blue square represents a pair for which the null hypothesis could not be rejected.

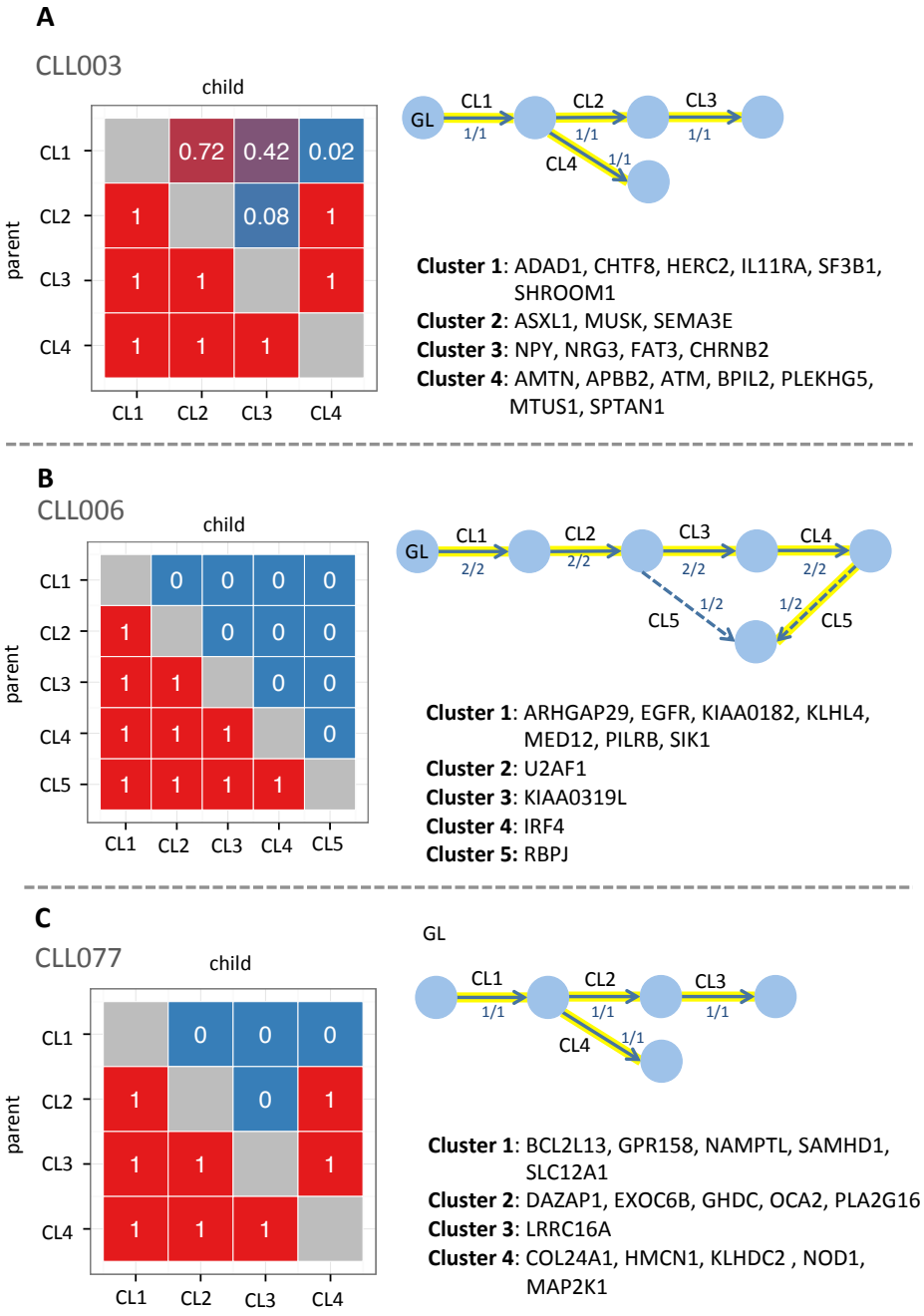

Figure 4: **Reconstruction of subclonal phylogenies in CLL. A. CLL003.** SCHISM identified a single highest scoring 4-node tree using 5 samples. **B. CLL006.** Two highest scoring 5-node trees were identified using 5 samples. Solid arrows represent lineage relationships shared by both trees and dashed arrows represent lineage relationships specific to one of the trees. **C. CLL077.** A single highest scoring 4-node tree was identified using 5 samples. Each arrow is labeled with the fraction of highest scoring trees that include the lineage relationship. Highlighted arrows indicate the tree manually constructed by the study authors. GL = germline state. Cluster precedence order violation (CPOV) matrices are shown to the left of each tree. Columns and rows represent mutation clusters. Each square represents a pair of mutation clusters (I,J) and the numeric value in the square shows the fraction of mutation pairs (i,j) for which the null hypothesis was rejected (Section 5.4). The mutated genes assigned to each cluster in [Schuh et al., 2012, Jiao et al., 2014] are listed.

means were inferred using kernel density estimation. Phylogenetic trees were constructed as in AML1 for four of the patients (AML40, AML27, AML35, AML43).

The authors described two distinct models of clonal evolution to explain relapse. In the first model, the dominant subclone present in the primary leukemia is not eliminated by therapy, but it acquires new mutations and thrives in the relapse. The patient may not have received a sufficiently aggressive treatment or may have harbored resistance mutations. In the second model, the dominant subclone is eliminated by therapy and a minor subclone in the primary acquires new mutations and thrives in the relapse, while some mutations in the primary are absent in the relapse. The mutations that allow the minor subclone to survive may have been present early on or have been acquired during or after chemotherapy, or both [Ding et al., 2012].

For patients where the authors had constructed a tree, we compared it to the best tree(s) identified by SCHISM. For the other patients, we considered whether the SCHISM trees were consistent with their suggested clonal evolution models.

For AML1, we used the published variant allele fractions and mutation cluster assignments (Table S5a [Ding et al., 2012]). The naive estimator was used to derive cellularity values (Section 5.3.1). Hypothesis tests were performed for pairs of mutations and the CPOV matrix was constructed by voting (Section 5.4). SCHISM identified a single highest scoring tree, which was identical to the authors' manually generated tree (Figure 5A). For the remaining seven patients, we used the published cluster mean variant allele frequencies (Table S10 [Ding et al., 2012]) and combined them with the authors' purity estimates to infer cluster mean cellularities (Section 5.3.1).

Patients AML27, AML35, and AML40 were reported to harbor only two mutation clusters each, and SCHISM identified a single highest scoring tree for each, which was identical to the authors' tree (Figure 5B). Patient AML15 was reported to harbor three mutation clusters, and SCHISM identified two highest scoring trees. Each tree supported one of the authors' two alternative models of AML relapse (Figure 5C). In one tree, the relapse-specific mutation cluster 3 descended from the dominant subclone in the primary. This subclone consisted of the 92% of cells in the primary that carried both cluster 1 and cluster 2 mutations (Section 5.7). In the other tree, it descended from the minor subclone in the primary (the 8% of cells in the primary carrying only cluster 1 mutations). Patient AML28 had five mutation clusters, and SCHISM identified a single highest scoring tree. While no subclone tree for AML28 was constructed by the authors, they proposed that this patient fit the second clonal evolution model of relapse driven by a minor subclone in the primary. The SCHISM tree was consistent with this model, as the relapse-specific mutation cluster 5 descended from a minor subclone (1% of cells), which harbored only mutation cluster 1, and mutation clusters 3 and 4, which were present in the primary were absent in the relapse cluster (Figure 5D). Patient AML31 had four mutation clusters, and SCHISM identified a single highest scoring tree (Figure 5E). Although no tree was provided by the authors, they proposed that this patient fit the second clonal evolution model, which was consistent with this tree. In the tree, the relapse-specific mutation cluster 4 descended from a minor subclone (21% of cells) in the primary. Cells in this subclone carried mutations in cluster 1, but not clusters 2 and 3. AML43 was reported to have four mutation clusters and SCHISM identified a single highest scoring tree that was identical to the authors' tree (Figure 5F).

4 Discussion

Representing tumor evolution as a phylogenetic tree of cell subpopulations can inform critical questions regarding the temporal order of mutations driving tumor progression and the mechanisms of recurrence and metastasis. As the cost of next generation sequencing with high coverage depth decreases, many labs are employing multi-region tumor sequencing strategies to study tumor evolution. However, going from multi-region sequencing data to a subclonal phylogeny is a computationally challenging task and methods are still in their early days. Here we derived a novel framework to approach the problem. We described a

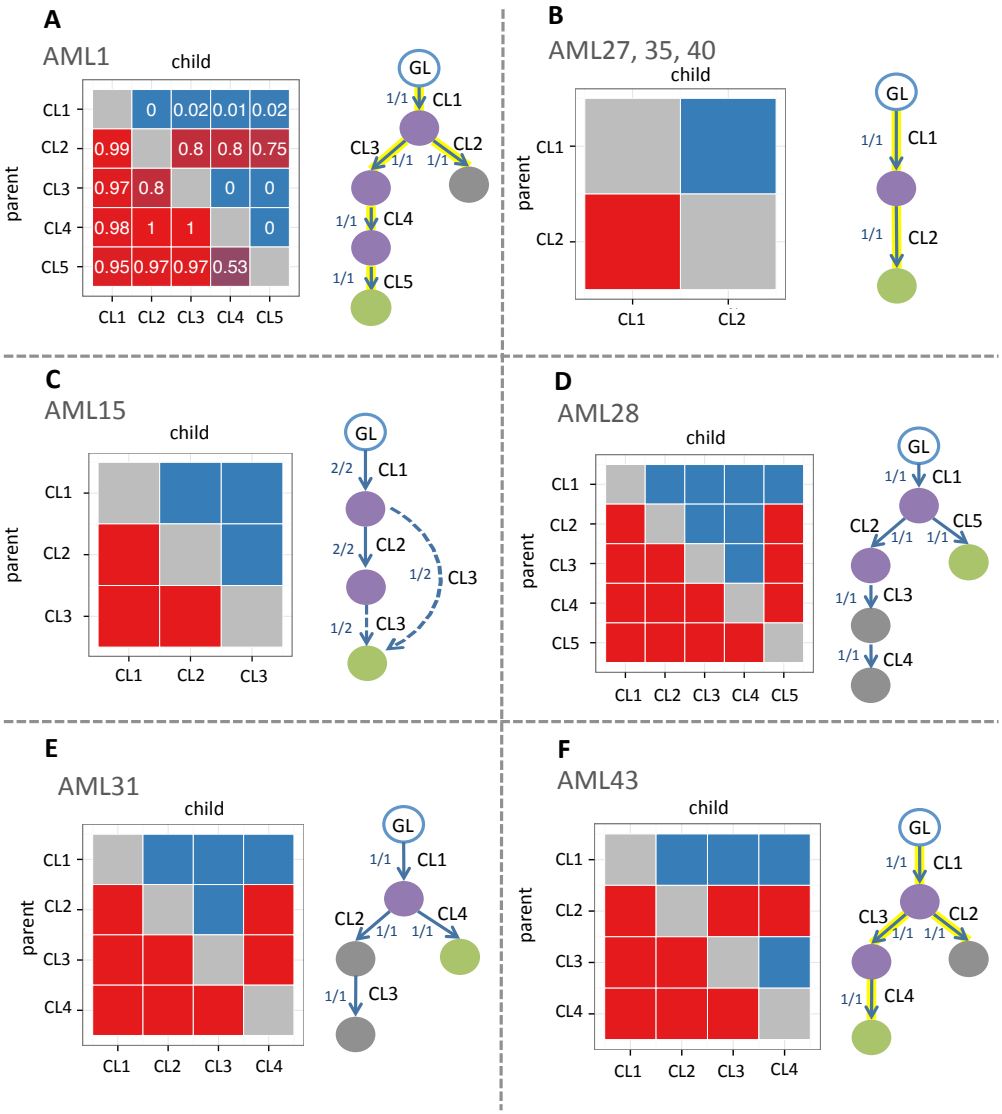

Figure 5: **Reconstruction of subclonal phylogenies in AML.** For each patient, two samples were available from primary and relapsed cancers. Purple nodes represent mutation clusters present in both primary and relapse samples. Gray nodes are present in primary but not relapse and green nodes are present in the relapse but not primary. **A. AML1.** SCHISM identified a single highest scoring 5-node tree. CPOV matrix columns and rows represent mutation clusters. Each square represents a pair of mutation clusters (I,J) and the numeric value in the square shows the fraction of mutation pairs (i,j) for which the null hypothesis was rejected (Section 5.4). **B. AML27,35,40.** For each patient, SCHISM identified a single highest scoring 2-node tree. **C. AML15.** Two highest scoring 3-node trees were identified. Solid arrows represent lineage relationships shared by both trees and dashed arrows represent lineage relationships specific to one of the trees. **D. AML28.** A single highest scoring 5-node tree. **E. AML31.** A single highest scoring 4-node tree was identified. **F. AML43.** A single highest scoring 4-node tree. Each arrow is labeled with the fraction of highest scoring trees that include the lineage relationship. Highlighted arrows indicate the tree manually constructed by the study authors, if it was available. GL = germline state. CL = cluster. Cluster precedence order violation (CPOV) matrices are shown to the left of each tree. Columns and rows represent mutation clusters. Each red square represents a pair of mutation clusters (I,J) for which the null hypothesis that I could be the parent of J was rejected. Each blue square represents a pair for which the null hypothesis could not be rejected.

statistical hypothesis test and mathematical representation of constraints on subclone phylogenies, based on rules of lineage precedence and divergence that have informed previous works in the field. We designed a new scoring function that can be used to constrain the process of subclone tree reconstruction. These tools comprise a flexible framework called SCHISM, which can be integrated with many existing methods for mutation cellularity estimation and phylogenetic reconstruction. Combined with a new implementation of genetic algorithms, we demonstrated the utility of SCHISM with simulations and by application to published multi-region sequencing studies. We were able to reconstruct the subclonal phylogenies derived by manual curation in these studies with high fidelity.

Today's multi-region sequencing studies may often have a limited number of tumor samples, due to restrictions on the number of biopsies likely to be performed for living patients. Our results suggest that even when only a few samples are available, more accurate estimates of mutation cellularity at higher purity and coverage increase the power of the SCHISM hypothesis test. A more subtle result is that the power and Type 1 error of the test also depend on the accuracy of the standard error estimates for cellularity values. The dependency can be seen directly in the derivation of the test statistic itself and indirectly in the ability of SCHISM to reconstruct complex subclone phylogenies in murine models of SCLC [McFadden et al., 2014]. Although only two or three samples were sequenced from each mouse, the authors provided robust statistical estimates of mutation cluster cellularity and standard deviations.

To our knowledge, our study is the first to apply genetic algorithms to the problem of subclone tree reconstruction. The scoring function ranks candidate phylogenetic trees, according to the extent to which they violated the rules of lineage precedence and divergence. We expect that scoring could be further improved by incorporating measures of mutation or mutation cluster importance, using knowledge about ordering of specific driver mutations based on tumor biology, synthetic lethality, or results from single-cell sequencing. The genetic algorithm used in this work could itself be improved by the addition of online termination criteria and adaptive modulation of its key parameters, such as crossover and mutation probabilities. It is also possible that better results could be achieved with alternative heuristic search methods such as tabu search [Glover, 1989][Glover, 1990], simulating annealing [Kirkpatrick et al., 1983][Cerny, 1985], or iterated local search [Lourenco et al., 2010].

Finally, it is clear that under many circumstances, particularly when sample count is low and tree complexity is high, the problem of subclone tree reconstruction is underdetermined. It is likely that for at least some tumor types, the true subclone trees may be very complex. In the future, sequencing studies with a large number of samples per patient will be essential to accurately characterize these trees.

5 Materials and Methods

5.1 Modeling assumptions

A rooted phylogenetic tree represents the history of tumor clonal evolution in an individual. Each tree node represents cells harboring a unique compartment of mutations, defining a subclone. Each edge represents a set of mutations, acquired by the cells in the child node and differentiating them from the cells in its parental node. The somatic mutations of each tumor cell then uniquely map it to one of the nodes in the tree. Given multiple tumor samples from an individual, a mutation may be present in varying proportions of the tumor cells in each sample, referred to as varying *cellularity* of the mutation across samples.

Two fundamental assumptions that make the modeling problem tractable [Strino et al., 2013, Jiao et al., 2014] are:

- **Maximum Parsimony** Each mutation arises only once throughout the history of the disease.
- **Lack of Reversions** Once a mutation emerges in a cell, it is inherited by all of its descendants.

The maximum parsimony assumption means that a mutation is uniquely assigned to an edge. The lack of reversions assumption means that all cells in a node harbor all mutations present in their parental node. Thus, a mutation present at a node cannot have greater cellularity than mutations present at its parental node. Furthermore, mutations present at a node with multiple descendant nodes cannot have lesser cellularity than the sum of cellularities of mutations present in the descendants. These assumptions have been incorporated into other methods for building subclone trees [Strino et al., 2013, Jiao et al., 2014, Hajirasouliha et al., 2014], and can be generally characterized as a *lineage precedence rule* and a *lineage division rule*.

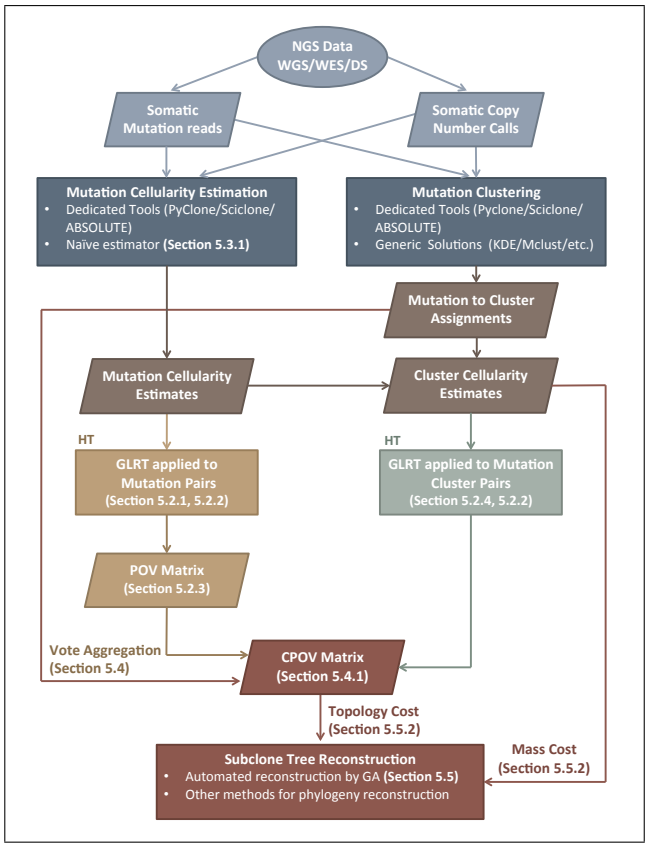

Figure 6: **Overview of SCHISM framework.** The framework decouples estimation of somatic mutation cellularities and reconstruction of subclone phylogenies. Given somatic mutation read counts from next generation sequencing data and somatic copy number calls if available, any tools for mutation cellularity estimation and mutation clustering can be applied. Their output is used to estimate the statistical support for temporal ordering of mutation or mutation cluster pairs, using a generalized likelihood ratio test (GLRT). Other approaches to tree reconstruction can be applied, by using the topology cost and mass cost scoring function as the objective for optimization. GA=genetic algorithm, WGS=whole genome sequencing, WES=whole exome sequencing, DS=(targeted) deep sequencing. KDE=kernel density estimation. POV=precedence order violation.

5.2 Hypothesis test

5.2.1 Generalized Likelihood Ratio Test

The lineage precedence rule for a pair of mutations i and j, where i precedes j in the same lineage implies that

$$C_i^s >= C_j^s \tag{1}$$

for s in $\{1, ..., S\}$, where C_i^s and C_j^s represent the cellularity of mutations i and j in sample s and S denotes the total number of samples from an individual. Then, for each ordered pair of mutations i and j, the null hypothesis ($H_0^{i \to j}$) to be tested is whether mutation i *can be* an ancestor of mutation j, and the alternative hypothesis ($H_A^{i \to j}$) is that it is not possible for i to be ancestral to j. Let the estimated cellularity for mutation i in sample s be ($\hat{C}_i^s$). It can be represented as a draw from a normal distribution centered at the true cellularity of mutation i in sample s (C_i^s) and with standard deviation σ_i^s.

$$\hat{C}_i^s \sim N(C_i^s, {\sigma_i^s}^2) \tag{2}$$

Assuming independence between the cellularity estimates for mutations i and j

$$d_{ij}^s = C_i^s - C_j^s \tag{3}$$

$$\hat{d}_{ij}^s \sim N(d_{ij}^s, {\sigma_{ij}^s}^2) \tag{4}$$

$${\sigma_{ij}^s}^2 = {\sigma_i^s}^2 + {\sigma_j^s}^2 \tag{5}$$

where d_{ij}^s represents the true difference in cellularity of mutations i, and j in sample s, and σ_{ij}^s is the standard deviation of $\hat{d}_{ij}^s$, the observed difference in cellularity of mutations i and j in sample s. Under $H_0^{i \to j}$, the cellularity of mutation i should exceed or be equal to that of mutation j in all tumor samples, based on the lineage precedence rule. Thus,

$$d_{ij}^s \geq 0, \ \forall s \in \{1, ..., S\} \tag{6}$$

Under the alternative hypothesis ($H_A^{i \to j}$), mutation i cannot be an ancestor to mutation j, and it is supported by the existence of samples in which the lineage precedence rule does not hold, and for which

$$d_{ij}^s < 0, \ \exists s \in \{1, ..., S\} \tag{7}$$

For a pair of mutations i and j and observations $\hat{d}_{ij}^s$ across S tumor samples, ($H_0^{i \to j}$) can be tested with a generalized likelihood ratio test ($GLRT$).

$$\Lambda = \frac{\max\left[lik\left(\hat{d}_{ij}^1, ..., \hat{d}_{ij}^S | d_{ij}^1, ..., d_{ij}^S \in \omega_0\right)\right]}{\max\left[lik\left(\hat{d}_{ij}^1, ..., \hat{d}_{ij}^S | d_{ij}^1, ..., d_{ij}^S \in \omega_0 \cup \omega_A\right)\right]} \tag{8}$$

The numerator represents the maximum likelihood estimate of observations $\hat{d}_{ij}^s$ for d_{ij}^s in ω_0, the parameter space corresponding to ($H_0^{i \to j}$) (Equation 6). The denominator represents the maximum likelihood estimate of observations $\hat{d}_{ij}^s$ where d_{ij} is in the union of the parameter spaces for ($H_0^{i \to j}$) and ($H_A^{i \to j}$), which implies there is no restriction on the values of d_{ij}^s.

Given the assumption of independence between d_{ij}^s across samples and considering parameter spaces defined by $H_0^{i \to j}$ (Equation 6) and $H_A^{i \to j}$ (Equation 7), Equation 8 can be simplified as

$$\Lambda = \prod_{s=1}^{s=S} \frac{max\left[lik\left(\hat{d}_{ij}^s | d_{ij}^s \in [0, +\infty)\right)\right]}{max\left[lik\left(\hat{d}_{ij}^s | d_{ij}^s \in (-\infty, +\infty)\right)\right]} \tag{9}$$

Using the normality assumption (Equation 4), Equation 9 is rewritten as

$$\Lambda = \prod_{s=1}^{s=S} \frac{\underset{d_{ij}^s \in [0,+\infty)}{max} \left[\frac{1}{\sigma_{ij}^s \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\hat{d}_{ij}^s - d_{ij}^s}{\sigma_{ij}^s})^2} \right]}{\underset{d_{ij}^s \in (-\infty,+\infty)}{max} \left[\frac{1}{\sigma_{ij}^s \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\hat{d}_{ij}^s - d_{ij}^s}{\sigma_{ij}^s})^2} \right]} \tag{10}$$

To derive the Λ, each d_{ij}^s is replaced by its single sample maximum likelihood estimator under $H_0^{i \to j}$ in the numerator. If $\hat{d}_{ij}^s \geq 0$, the value of d_{ij}^s that maximizes the numerator is equal to $\hat{d}_{ij}^s$. On the other hand, for $\hat{d}_{ij}^s < 0$, the value of d_{ij}^s that maximizes the numerator is 0, since negative values of d_{ij}^s are not allowed under $H_0^{i \to j}$. In the denominator, since the value of d_{ij}^s is a non-restricted parameter, its maximum likelihood estimator is always equal to $\hat{d}_{ij}^s$. Thus Equation 10 can be rewritten as

$$\Lambda = \prod_{\substack{s=1 \\ \hat{d}_{ij}^s < 0}}^{s=S} \left(\frac{\frac{1}{\sigma_{ij}^s \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\hat{d}_{ij}^s - 0}{\sigma_{ij}^s})^2}}{\frac{1}{\sigma_{ij}^s \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\hat{d}_{ij}^s - \hat{d}_{ij}^s}{\sigma_{ij}^s})^2}} \right) \cdot \prod_{\substack{s=1 \\ \hat{d}_{ij}^s \geq 0}}^{s=S} \left(\frac{\frac{1}{\sigma_{ij}^s \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\hat{d}_{ij}^s - \hat{d}_{ij}^s}{\sigma_{ij}^s})^2}}{\frac{1}{\sigma_{ij}^s \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\hat{d}_{ij}^s - \hat{d}_{ij}^s}{\sigma_{ij}^s})^2}} \right) \tag{11}$$

which simplifies to

$$\Lambda = \prod_{\substack{s=1 \\ \hat{d}_{ij}^s < 0}}^{s=S} e^{-\frac{1}{2}(\frac{\hat{d}_{ij}^s}{\sigma_{ij}^s})^2} \tag{12}$$

Therefore, we reject the null hypothesis that mutation i could be an ancestor of mutation j ($H_0^{i \to j}$) if the test statistic T is significantly large.

$$T = -2 \log(\Lambda) = \sum_{s=1}^{s=S} (\frac{\hat{d}_{ij}^s}{\sigma_{ij}^s})^2 \cdot I_{\hat{d}_{ij}^s < 0} \tag{13}$$

where I is a binary indicator variable. Equation 13 explicitly shows how T depends on the accuracy of cellularity values and their standard deviation. When standard deviation is overestimated T is underestimated, yielding power loss. When standard deviation is underestimated, T is overestimated, yielding Type 1 error inflation.

Since the standard deviation is unknown, standard error was used to calculate the test statistic.

5.2.2 Significance Evaluation

To assess the significance of an observed value of the *GLRT* test statistic (T) (Equation 13), we consider the distribution of T under $H_0^{i \to j}$ and derive the Type 1 error probability of the test. Similar results have previously been derived for a more general class of GLRTs [Barlow et al., 1972]. The distribution of each summation term in Equation 13 ($(\frac{\hat{d}_{ij}^s}{\sigma_{ij}^s})^2 \cdot I_{\hat{d}_{ij}^s < 0}$) depends on the true value of d_{ij}^s (Equation 4). Given a fixed value of σ_{ij}^s, large positive values of d_{ij}^s make observation of negative $\hat{d}_{ij}^s$ and thus a corresponding non-zero term in the summation less likely. Therefore,

$$\underset{d_{ij}^s \geq 0}{P} \left((\frac{\hat{d}_{ij}^s}{\sigma_{ij}^s})^2 \cdot I_{\hat{d}_{ij}^s < 0} > C \right) \leq \underset{d_{ij}^s = 0}{P} \left((\frac{\hat{d}_{ij}^s}{\sigma_{ij}^s})^2 \cdot I_{\hat{d}_{ij}^s < 0} > C \right) \tag{14}$$

By extending the above argument to every term in the summation, we derive an upper bound for the probability of test statistic T exceeding a critical value C under the null hypothesis $H_0^{i\to j}$.

$$\underset{d_{ij}^1,\ldots,d_{ij}^S\in\omega_0}{P}[T\geq C]=\underset{d_{ij}^1,\ldots,d_{ij}^S\geq 0}{P}[T\geq C]\leq\underset{d_{ij}^1,\ldots,d_{ij}^S=0}{P}[T\geq C] \tag{15}$$

Therefore, to control the Type 1 error probability of the test, it is sufficient to control Type 1 error probability of a test where the null hypothesis is reduced to $d_{ij}^1,\ldots,d_{ij}^S=0$ (*reduced null hypothesis*). Under the reduced null hypothesis we can derive the exact distribution of the test statistic T as follows. Let z_{ij}^s denote $\frac{\hat{d}_{ij}^s}{\sigma_{ij}^s}$.

$$P\left[\sum_{s=1}^{S}(\frac{\hat{d}_{ij}^s}{\sigma_{ij}^s})^2\cdot I_{\hat{d}_{ij}^s<0}\geq C\right]=P\left[\sum_{s=1}^{S}{z_{ij}^s}^2\cdot I_{z_{ij}^s<0}\geq C\right] \tag{16}$$

Next, let δ_V be a binary random variable representing the event when, for a particular subset V of $\{1,\ldots,S\}$, we have $z_{ij}^s<0$ for $s\in V$ and $z_{ij}^s\geq 0$ for $s\notin V$. By the law of total probability,

$$P\left[\sum_{s=1}^{S}{z_{ij}^s}^2\cdot I_{z_{ij}^s<0}\geq C\right]=\sum_{V\subset\{1,\ldots,S\}}P\left[\sum_{s=1}^{S}{z_{ij}^s}^2\cdot I_{z_{ij}^s<0}\geq C,\delta_V\right] \tag{17}$$

or equivalently,

$$P\left[\sum_{s=1}^{S}{z_{ij}^s}^2\cdot I_{z_{ij}^s<0}\geq C\right]=\sum_{V\subset\{1,\ldots,S\}}P\left[\sum_{s\in V}{z_{ij}^s}^2\geq C,\delta_V\right] \tag{18}$$

But the reduced null hypothesis (Equation 15) states that all d_{ij}^s in Equation 4 are set to zero, so

$$\hat{d}_{ij}^s\sim N\left(0,{\sigma_{ij}^s}^2\right) \tag{19}$$

or equivalently,

$$z_{ij}^s\sim N\left(0,1\right) \tag{20}$$

Also, for a set of independent identically distributed random draws from $N(0,1)$, the value of the summation in Equation 18 is independent of the signs of $\{z_{ij}^1,\ldots,z_{ij}^S\}$, and is a χ^2 random variable with $|V|$ degrees of freedom.

$$\sum_{V\subset\{1,\ldots,S\}}P\left[\sum_{s\in V}{z_{ij}^s}^2\geq C,\delta_V\right]=\sum_{V\subset\{1,\ldots,S\}}P\left[\sum_{s\in V}{z_{ij}^s}^2\geq C\right]P\left[\delta_V\right] \tag{21}$$

$$=\sum_{V\subset\{1,\ldots,S\}}P\left[\chi^2_{|V|}\geq C\right]P\left[\delta_V\right] \tag{22}$$

Equation 20 implies that each random variable z_{ij}^s assumes positive and negative signs with equal probability. Thus, the probability of observing a particular sequence of signs for random variables $z_{ij}^1,\ldots,z_{ij}^S$, i.e. the probability of each δ_V being true, is $\left(\frac{1}{2}\right)^S$.

$$\sum_{V\subset\{1,\ldots,S\}}P\left[\chi^2_{|V|}\geq C\right]P\left[\delta_V\right]=\sum_{V\subset\{1,\ldots,S\}}P\left[\chi^2_{|V|}\geq C\right]\left(\frac{1}{2}\right)^S \tag{23}$$

Finally, summarizing the sum above over all possible values $|V|$ can take,

$$\sum_{V\subset\{1,\ldots,S\}}P\left[\chi^2_{|V|}\geq C\right]\left(\frac{1}{2}\right)^S=\sum_{k=0}^{S}\frac{\binom{S}{k}}{2^S}P\left[\chi^2_k\geq C\right] \tag{24}$$

Thus it can be concluded that the distribution of $GLRT$ test statistic T under the reduced null hypothesis is that of a random variable drawn from a mixture of χ^2 distributions, with degrees of freedom varying in $k \in \{0, ..., S\}$, and the weight of each mixture component equal to $\frac{\binom{S}{k}}{2^S}$. Here, a χ^2_0 random variable is defined as one that is fixed at zero. We use this derivation to assign a conservative estimate of significance level to an observed value of the test statistic T under $(H_0^{i \to j})$.

5.2.3 Precedence Order Violation Matrix

For each possible ordered pair of mutations (i, j) characterized in a set of tumor samples from the same individual, we test the hypothesis that mutation i is a potential ancestor of mutation j. A fixed common significance level of $\alpha = 0.05$ is assigned to decide the outcome of each pairwise test. These results can then be organized as a binary *Precedence Order Violation* (POV) matrix, where non-zero entries mark mutation pairs (i, j) for which the null hypothesis $H_0^{i \to j}$ was rejected.

5.2.4 Application to mutation clusters

Given a mapping of individual mutations onto clusters, the hypothesis test can be applied to pairs of clusters rather than to pairs of mutations, and in this case is used to generate a straightforward extension of the POV matrix *Cluster Precedence Order Violation* ($CPOV$ matrix) where non-zero entries mark mutation cluster pairs (I, J) for which the null hypothesis $H_0^{I \to J}$ was rejected. An alternate approach for generating a CPOV matrix is described in Section 5.4.

5.3 Cellularity estimation

The hypothesis test requires an estimate of the mean and standard deviation of cellularity for each mutation or mutation cluster as an input, but it is agnostic to how the estimate is derived.

5.3.1 Naive estimate for diploid regions

In diploid regions, a maximum likelihood estimate based on the observed variant allele fraction of a mutation can be used to infer its cellularity. We use the following simple derivation to estimate the cellularity and its standard error from reference and variant read counts for a mutation i in a diploid region of the genome. We further assume that the genotype of normal cells and non-variant tumor cells (tumor cells not carrying the mutation) is AA where A is the reference allele, and there is no loss of heterozygosity. Under these conditions, the probability of sampling a variant allele from tumor sample s with purity α^s, for mutation i with cellularity C_i^s is

$$V_{exp} = \frac{\alpha^s \cdot C_i^s}{2} \tag{25}$$

where V_{exp} is the expected variant allele fraction. By the binomial read count assumption, $\frac{r_B}{r_T}$ is a single sample unbiased estimator of V_{exp}; here r_B and r_T represent the observed variant and total read count of a mutation, respectively.

$$r_B \sim binomial(r_T, V_{exp}) \tag{26}$$

$$V_{obs} = \frac{r_B}{r_T} \tag{27}$$

Therefore, mutation cellularity can be estimated as

$$\hat{C}_i^s = \frac{2\ V_{obs}}{\alpha^s} \tag{28}$$

Finally, the estimated variance of the above estimator is

$$\sigma_i^{s\ 2} = \frac{4}{\alpha^{s\ 2}}\sigma^2_{V_{obs}} \tag{29}$$

$$= \frac{4}{\alpha^{s\ 2}}\frac{V_{exp}\ (1 - V_{exp})}{r_T} \tag{30}$$

$$\sigma_i^{s\ 2} \approx \frac{4}{\alpha^{s\ 2}}\frac{V_{obs}\ (1 - V_{obs})}{r_T} \tag{31}$$

Equation 31 shows that the variance of the cellularity estimates decreases as purity and coverage increase.

Note that a simple modification allows us to extend this approach to regions with copy number = 1. Then mutation cellularity is

$$\hat{C}_i^s = \frac{(2 - \alpha^s)\ V_{obs}}{\alpha^s} \tag{32}$$

The estimated variance is

$$\sigma_i^{s\ 2} \approx \frac{(2 - \alpha^s)^2}{\alpha^{s\ 2}}\frac{V_{obs}\ (1 - V_{obs})}{r_T} \tag{33}$$

More sophisticated estimates of cellularity incorporate variable ploidy states [Roth et al., 2014, Miller et al., 2014], and these can also be used within the SCHISM framework.

5.3.2 Cluster cellularity

By definition, all mutations in a cluster are assumed to have the same cellularity in each sample. If the cellularity of individual mutations is available, the cellularity of a cluster is estimated to be the mean cellularity of its members

$$CC[I, s] = \frac{\sum_{i \in I} C_i^s}{|M(I)|}. \tag{34}$$

Otherwise cluster cellularity values from other sources can be used. The cellularity of each cluster across multiple samples is represented as a a matrix CC whose elements report the cellularity of each cluster I, for each sample s.

5.4 Vote aggregation

The Cluster Precedence Order Violation (CPOV) matrix can be generated by the following vote aggregation approach. Let the set of mutations assigned to cluster I be $M(I)$. Rows and columns of the POV matrix can be reordered so that mutations belonging to the same cluster are adjacent. Then the ordered interaction of any pair of clusters (I, J) is represented by a block of matrix entries with addresses

$$\underset{\forall i \in M(I), \forall j \in M(J)}{POV[i, j]} \tag{35}$$

The support for potential lineage precedence of cluster I to cluster J can be summarized by a vote of the matrix elements within the block, represented as an element (I, J) in a cluster-level POV matrix $CPOV$

$$CPOV[I, J] = \frac{\sum_{i \in M(I), j \in M(J)} POV[i, j]}{|M(I)| \cdot |M(J)|} \tag{36}$$

where $|M(X)|$ denotes the number of mutations in cluster X.

5.5 Genetic algorithm

A genetic algorithm (GA) is a heuristic search inspired by the process of natural selection. In an initial generation, a set of random objects is created and their fitness with respect to a scoring criteria is evaluated. Next, objects from the initial generation are selected according to their fitness to be parents of the following generation, with a preference for the fittest parents. The parental objects reproduce themselves, and their progeny may harbor new variation. The process is repeated for either a fixed number of generations or until a pre-defined convergence criteria is reached.

In our implementation, the GA searches through a space of phylogenetic tree topologies, ranking their fitness with a scoring function that we derived based on our model assumptions. In the initial generation, we generate $G_0 = 1000$ random tree topologies and evaluate the fitness of each tree (Section 5.5.2). A sample of size $0.8 * G_0$ from the highest scoring trees are selected for reproduction by a fitness proportional selection method [Miller and Goldberg, 1996] and their progeny are generated, using crossover and mutation operations (Figures 7 and 8). To increase diversity and avoid too fast convergence to a local optimum, $0.2 * G_0$ random tree topologies are also generated. The following generation then consists of a mixture of the progeny of the previous generation(s) and new random trees, and the total number of trees is the same as in the previous generation, so that $G_1 = G_0$. The process is repeated for a fixed number $\gamma = 20$ generations. For each generation, the trees selected to be parents are not limited to the previous generation only, but can be selected from any preceding generations. To avoid getting trapped in local optima, four independent runs of the GA are performed, each with 20 generations (1000 trees per generation), and the entire ensemble of trees sampled in the four runs is ranked by tree fitness.

In this work, the number of mutation clusters and the cellularity of each mutation cluster in each sample is assumed to be known, and the GA is applied to explore the space of tree topologies with a given node count, including both linear and branched topologies.

5.5.1 Random Topology Generation

To generate a random tree topology, a mutation cluster is randomly selected and assigned to the incident edge downstream of the root node. An incident node downstream of the edge is appended. Next, one of the remaining mutation clusters and a non-root node are randomly selected and the cluster is assigned to the incident edge downstream of this node, again appending a new incident node downstream of the edge. The process continues until all mutation clusters in the data have been assigned.

5.5.2 Scoring function

Trees are scored as

$$Z(T) = TC(T) + MC(T) \tag{37}$$

where $TC(T)$ is a *topology cost* that summarizes violations of the lineage precedence rule, and $MC(T)$ is a *mass cost* that summarizes violations of the lineage division rule. The *topology cost* (TC) of tree T is

$$TC(T) = \sum_{I,J \in E(T), I \to J} tc(I, J) \tag{38}$$

where $tc(I, J)$ is the topology cost of each ($ancestor \to descendant$) edge pair, equivalent to $CPOV(I, J)$ (Equation 36) and $E(T)$ represents the set of edges in tree T. The *mass cost* (MC) of tree T is

$$MC(T) = \sum_{n \in N(T)} mc(n) \tag{39}$$

Algorithm **Crossover Operation**

INPUT: *A pair of parental trees T_1, T_2 selected for sexual reproduction*
OUTPUT: *A pair of progeny intermediate trees T_1^I, T_2^I*

1. **for** each tree T_i, i in {1,2}:
   1.1 select a random edge e_i;
   1.2 let the incident node downstream of the edge be the root of subtree T_i' with node count N_i';
   1.3 let the incident node upstream of the edge be $u(e_i)$;

2. **if** $N_1' \neq N_2'$ **then**
   while the subtrees have unequal node count:
   3.1 select a random edge e_r incident to a leaf in the larger subtree;
   3.2 detach the edge and leaf node from the subtree;

3. **for** each tree T_i, i in {1,2}:
   cut the tree at edge e_i to detach subtree T_i'

4. T_1^I is the tree generated by attaching subtree T_2' to node $u(e_1)$ and T_2^I by attaching T_1 to $u(e_2)$

5. T_1^I and T_2^I are a pair of progeny intermediate trees.

Figure 7: Crossover operation used by the genetic algorithm.

where $mc(n)$, the total mass cost for node n, is the Euclidean norm of the vector of node mass costs across samples

$$mc(n) = \sqrt{\sum_{s=1}^{s=S} \left(mc^s\left(n\right)\right)^2} \tag{40}$$

and $mc^s(n)$ is the mass cost for node n in sample s

$$mc^s(n) = \begin{cases} 0, & C^s_{p(n)} \geq \sum_{q \in D(n)} C^s_q \\ \sum_{q \in D(n)} C^s_q - C^s_{p(n)}, & C^s_{p(n)} < \sum_{q \in D(n)} C^s_q \end{cases} \tag{41}$$

where $C^s_{p(n)}$ is the cellularity of the mutation cluster associated with the upstream edge incident to node n, *i.e.*, $p(n)$ in sample s, and $\sum_{q \in D(n)} C^s_q$ is the sum of cellularities of mutation clusters associated with its set of *descendant edges* $D(n)$. The fitness F of the tree T is then a monotonically decreasing function of the tree cost $Z(T)$.

$$F(T) = e^{-f_c \cdot Z(T)} \tag{42}$$

where f_c is a positive-valued scaling coefficient (default $f_c = 5$), yielding fitness reduction by a factor of $\sim 150\texttt{X}$ for each unit increase in total cost.

5.5.3 Mutation and Crossover Operations

The tree topologies selected for sexual reproduction are randomly paired, and each pair undergoes a crossover operation with probability P_c to yield two progeny intermediate trees (Figure 7). Next, a mutation operation is applied to each progeny intermediate tree with probability P_m. There are two possible mutation operations (Figure 8), and for each tree one is selected with probability P_s. The result is a collection of new progeny trees that will appear in the next generation. (Default $P_c = 0.25$ and $P_m = 0.9$ and $P_s = 0.6$.)

Algorithm **Mutation Operations**

INPUT: *A progeny intermediate tree* T^I
OUTPUT: *A progeny tree* T^O

1. Draw $x \sim uniform(0,1)$;

2. **if** $x < P_s$ **then**
   1.1 select a random pair of edges e_i, e_j from T^I, representing mutation clusters I and J, respectively;
   1.2 assign cluster I to e_j and cluster J to e_i;

3. **else**
   2.1 select two random edges e_i and e_j from T^I;
   2.11 **if** e_i is upstream of e_j **then**
   let $e_u = e_i$ and $e_d = e_j$;
   2.12 **else if** e_j is upstream of e_i **then**;
   let $e_u = e_j$ and $e_d = e_i$;
   2.13 **else** ;
   randomly assign e_u and e_d to e_i and e_j;
   2.3 let the incident node downstream of e_d be the root of subtree T_d^I;
   2.4 cut subtree T_d^I from T^I at e_d;
   2.5 T^O is the tree generated by attaching subtree T_d^I to the incident node downstream of e_u;

4. T^O is a progeny tree.

Figure 8: Mutation operations used by the genetic algorithm.

5.6 Simulation

The simulations were designed to generate data compatible with a set of likely tree topologies and assess how well SCHISM could recover these topologies from this data. Given a tree topology, a simulation produces a set of tumor samples consistent with lineage relationships summarized in the tree. We assume that while the samples share these lineage relationships, each represents an independent instantiation of cellularity distributions on the edges of the tree. The variability among these simulated samples captures the stochastic process of preferential sampling of tumor cells in an individual's multiple tumor samples. In each simulated sample, we model variant and reference read counts for mutations belonging to each edge in the tree, taking into account sequencing coverage depth, sample purity level and mutation cluster cellularity.

5.6.1 Generating subclonal phylogenies

Simulated trees range in size from three to eight nodes, with no restrictions on the number of child nodes. For trees with three to five nodes, an exhaustive set of topologies is generated. Otherwise, ten topologies are selected. Each unique topology at a given node count is considered an *instance*. A Poisson process with rate parameter $\lambda = 10$ is used to simulate the number of mutations that occurred along each each edge in the tree. Node count does not include the root node, which represents the germline state, prior to any somatic mutations (Figure 9).

5.6.2 Generating mutation cellularities

In each simulated sample s, a breadth-first-search (BFS) of the tree begins at the incident edge downstream of the root node. The mutation cluster corresponding to this edge has cellularity of 1, and it represents clonal mutations occurring in the most recent common progenitor cell of all the patient's tumor cells. For subsequent edges, cellularity values are distributed with a modified version of the tree-structured stick-breaking process model [Ghahramani et al., 2010, Jiao et al., 2014].

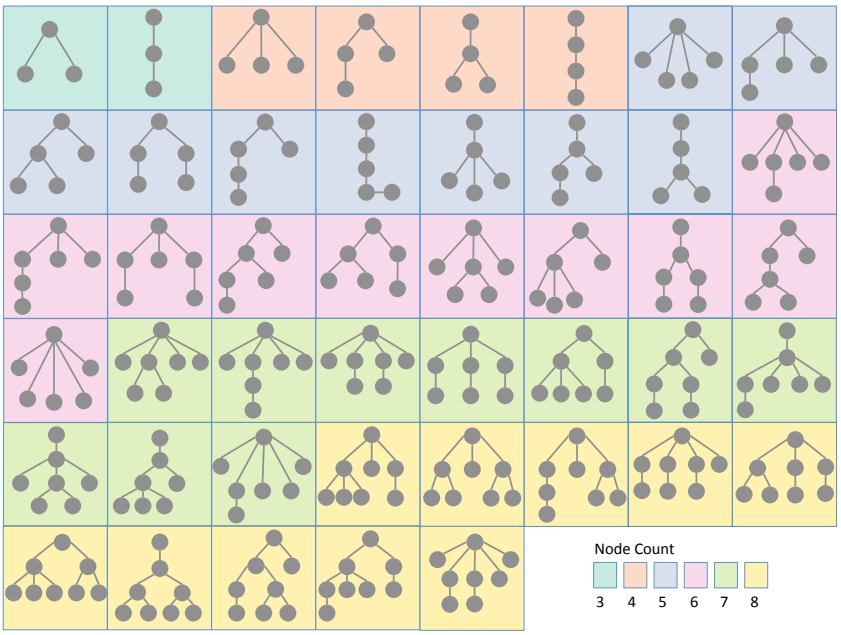

Figure 9: **Tree topologies used in the simulations.** For trees with 3-5 nodes, the exhaustive set of possible topologies were used. Otherwise, we manually selected ten topologies. Each box depicts a tree instance.

For each tree topology *instance* at each node count level, ten sets of mutation cluster cellularity values are generated, representing 10 samples from an individual. Each tree node then represents a unique set of cells or subclones harboring mutations, which have accumulated along the path from the root to that node. As in (Section 5.5.2 Equation 41), $p(n)$ is the mutation cluster associated with the edge immediately upstream of node n, and $D(n)$ is the set of mutation clusters associated with its downstream edges. Letting $n = 0$ correspond to the node immediately downstream of the root, the edge corresponding to clonal mutations is assigned cellularity $C^s_{p(0)} = 1$. For each subsequent node n, the fraction of tumor cells that harbor $p(n)$ but none of the mutation clusters downstream of n is $C^s_{p(n)}\omega(n)^s$ where

$$\omega(n)^s \sim B\left(\alpha_b = 1, \beta_b = \alpha_0\ \lambda_0^{(d_n-1)}\right) \tag{43}$$

and $\alpha_0 = 5$, $\lambda_0 = 0.5$ and d_n is the depth of node n in the rooted tree. Then $C^s_{p(n)}(1-\omega(n)^s)$ is the fraction of cancer cells harboring at least one mutation cluster in $D(n)$ in addition to $p(n)$. These are cells that diverge from their parental population at node n. Therefore,

$$\sum_{q\in D(n)} C^s_q = C^s_{p(n)}(1-\omega(n)^s) \tag{44}$$

Letting V be a vector of size $|D(n)|$, and

$$V \sim Dir\,(\alpha_{dir} = 1) \tag{45}$$

then the cellularity of each downstream mutation cluster is

$$[C^s_q, q \in D(n)] = VC^s_{p(n)}(1-\omega(n)^s) \tag{46}$$

To capture variability among individuals, this procedure is replicated ten times for each tree *instance*.

5.6.3 Mutation Variant Allele Fractions

To obtain read counts that are consistent with the simulated cellularity values, in each sample s, variant and reference read count values are generated for each mutation in mutation cluster $p(n)$ associated with the edge immediately upstream of node n, given simulated cellularity value $C^s_{p(n)}$ as follows. Following the assumptions in Section 5.3.1, the expected variant allele frequency for mutations in $p(n)$ is

$$V^{p(n)}_{exp} = \frac{\alpha^s C^s_{p(n)}}{2} \tag{47}$$

Since read counts may be overdispersed, a noisy expected variant allele frequency for each mutation $i \in p(n)$ is generated as

$$S \sim \Gamma\left(k = 10000, \theta = 1\right) \tag{48}$$

$$V^i_{exp} = B\left(S\ V^{p(n)}_{exp}, S\ \left(1 - V^{p(n)}_{exp}\right)\right) \tag{49}$$

where S is a global precision parameter for each simulated individual. Given coverage r^i_T for variant i, variant read count is

$$r^i_B \sim binom\left(r^i_T, V^i_{exp}\right) \tag{50}$$

5.7 Subclone size estimation

Based on our modeling assumptions and from (Section 5.5.2, Equation 41), it is straightforward to conclude that in each sample s, the fraction of tumor cells belonging to the subclone described by node n in a tree can be calculated as

$$C^s_{p(n)} - \sum_{q \in D(n)} C^s_q \tag{51}$$

where $C^s_{p(n)}$ is the cellularity of the mutation cluster associated with the upstream edge incident to node n, *i.e.*, $p(n)$ in sample s, and $\sum_{q \in D(n)} C^s_q$ is the sum of cellularities of mutation clusters associated with its set of *descendant edges* $D(n)$.

5.8 Assessment of hypothesis testing

Each element of the precedence order violation matrix $POV[i, j]$ (Section 5.2.3) is a binary indicator of whether the null hypothesis that mutation i can be ancestral to mutation j is rejected. Each element in the POV matrix is compared with its true value, given the correct tree topology. Performance is summarized by power and Type 1 error.

5.9 Assessment of automated subclone tree reconstruction

The ability of the genetic algorithm to identify the correctly reconstructed subclone tree is assessed across multiple settings of key variables: tumor purity, sequencing coverage depth, tree node count, and tumor sample count. For each node count, multiple tree instances (alternate topologies for a given node count) are considered. Then for each combination of settings, 10 different replicates are run (Figure 10). Each replicate can be viewed as an *in silico* patient having the selected number of samples and a distinct cellularity profile across the samples.

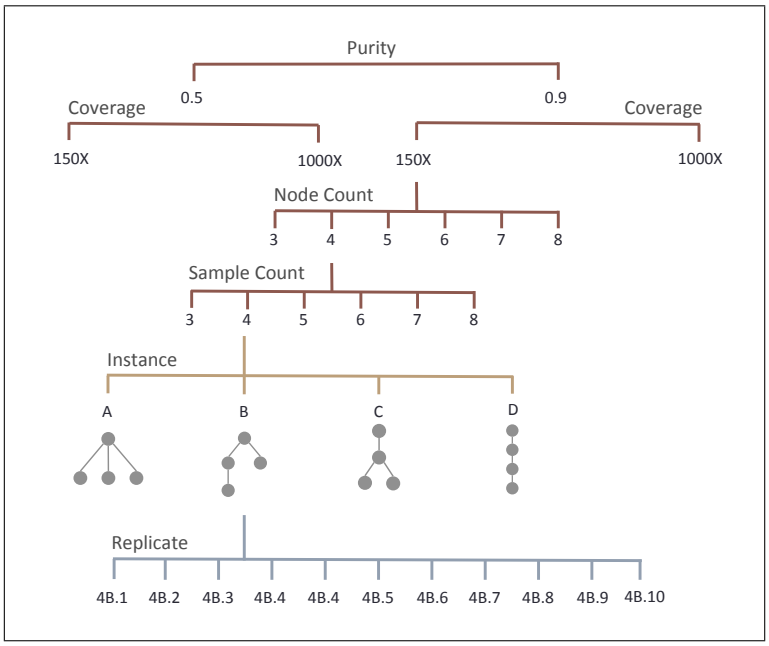

Figure 10: **The genetic algorithm is assessed across multiple settings of tumor purity, sequencing coverage depth, tree node count, and tumor sample count.** The example shows purity of 0.9, coverage of 150X, node count of 4, sample count of 4, and a selected tree instance.

5.9.1 Number of highest scoring trees identified by the genetic algorithm

In underdetermined cases, ranking of trees generated by the GA for a replicate may result in ties. We conservatively define a *Stage 1 success* for a replicate as an outcome where only one or two highest scoring trees have been identified. To estimate the probability of Stage 1 success, the frequency of success of all tree instances and their replicates for each of 240 unique scenarios is computed. In practice, multiple highest scoring trees can be a useful result, but limiting the number of ties in this way makes the assessment of the simulations more tractable.

5.9.2 Agreement of highest scoring tree(s) with the true tree

Even in the absence of ties, the highest scoring tree discovered by the GA may not be the true tree that was used as the basis for the simulation. For this assessment, *Stage 2 success* for a replicate is an outcome where either a single highest scoring or top two highest scoring trees are the true tree. To assess Stage 2 success, we eliminate replicates where Stage 1 failed and of those remaining, calculate the fraction in which the true tree is either the highest scoring or one of two equally highest scoring trees.

6 Acknowledgments

NIH NCI grant R01CA179991 provided support for NN. We thank Christine Iacobuzio-Donahue and Alvin Makohon-Moore for valuable discussions about subclonal evolution.

References

R.E. Barlow, D. J. Bartholomew, J.M. Bremner, and H.D. Brunk. *Statistical Inference under Order Restrictions*. John Wiley, 1972.

Vladimir Cerny. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of optimization theory and applications*, 45(1):41–51, 1985.

Li Ding, Timothy J Ley, David E Larson, Christopher A Miller, Daniel C Koboldt, John S Welch, Julie K Ritchey, Margaret A Young, Tamara Lamprecht, Michael D McLellan, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–510, 2012.

Chris Fraley and Adrian E Raftery. MCLUST version 3: an R package for normal mixture modeling and model-based clustering. Technical report, DTIC Document, 2006.

Marco Gerlinger, Stuart Horswell, James Larkin, Andrew J Rowan, Max P Salm, Ignacio Varela, Rosalie Fisher, Nicholas McGranahan, Nicholas Matthews, Claudio R Santos, et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature Genetics*, 46(3): 225–233, 2014.

Zoubin Ghahramani, Michael I Jordan, and Ryan P Adams. Tree-structured stick breaking for hierarchical data. In *Advances in Neural Information Processing Systems*, pages 19–27, 2010.

Fred Glover. Tabu search-Part I. *ORSA Journal on Computing*, 1(3):190–206, 1989.

Fred Glover. Tabu search-Part II. *ORSA Journal on Computing*, 2(1):4–32, 1990.

Mel Greaves and Carlo C. Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–313, Jan 2012. doi: 10.1038/nature10762.

Iman Hajirasouliha, Ahmad Mahmoody, and Benjamin J Raphael. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, 30(12):i78–i86, 2014.

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(1):35, 2014.

Brett E Johnson, Tali Mazor, Chibo Hong, Michael Barnes, Koki Aihara, Cory Y McLean, Shaun D Fouse, Shogo Yamamoto, Hiroki Ueda, Kenji Tatsuno, et al. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science*, 343(6167):189–193, 2014.

Scott Kirkpatrick, C Daniel Gelatt, Mario P Vecchi, et al. Optimization by simmulated annealing. *Science*, 220(4598):671–680, 1983.

Rumen L Kostadinov, Mary K Kuhner, Xiaohong Li, Carissa A Sanchez, Patricia C Galipeau, Thomas G Paulson, Cassandra L Sather, Amitabh Srivastava, Robert D Odze, Patricia L Blount, et al. NSAIDs modulate clonal evolution in Barrett's esophagus. *PLoS Genetics*, 9(6):e1003553, 2013.

Helena R Lourenco, Olivier C Martin, and Thomas Stutzle. Iterated local search: Framework and applications. In *Handbook of Metaheuristics*, pages 363–397. Springer, 2010.

David G McFadden, Thales Papagiannakopoulos, Amaro Taylor-Weiner, Chip Stewart, Scott L Carter, Kristian Cibulskis, Arjun Bhutkar, Aaron McKenna, Alison Dooley, Amanda Vernon, et al. Genetic and clonal dissection of murine small cell lung carcinoma progression by genome sequencing. *Cell*, 156 (6):1298–1311, 2014.

Brad L Miller and David E Goldberg. Genetic algorithms, selection schemes, and the varying effects of noise. *Evolutionary Computation*, 4(2):113–131, 1996.

Christopher A Miller, Brian S White, Nathan D Dees, Malachi Griffith, John S Welch, Obi L Griffith, Ravi Vij, Michael H Tomasson, Timothy A Graubert, Matthew J Walter, et al. SciClone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Computational Biology*, 10(8):e1003665, 2014.

Peter C Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah. Pyclone: statistical inference of clonal population structure in cancer. *Nature Methods*, 2014.

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

Francesco Strino, Fabio Parisi, Mariann Micsinai, and Yuval Kluger. TrAp: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Research*, 41(17):e165–e165, 2013.