Phylogenetic Inference of Copy Number Alterations and Single Nucleotide Variants from Longitudinal Single-Cell Sequencing

Ethan Kulman[1,2], Quaid Morris[2,⋆], and Rui Kuang[1,⋆]

[1] Department of Computer Science and Engineering, University of Minnesota, Minneapolis MN 55455, USA
[2] Computational and Systems Biology Program, Sloan Kettering Institute, 417 E 68th St, New York, 10065

Abstract. Longitudinal phylogenetic reconstruction reveals how cancers evolve over time and respond to treatments. Advances in targeted single-cell sequencing, combined with longitudinal sampling, now enable detailed longitudinal tracking of single nucleotide variants (SNVs) and copy number alterations (CNAs) at single-cell resolution. While some phylogenetic inference methods can jointly reconstruct SNV and CNA evolution, none are designed for longitudinal single-cell data. Here, we present scaOrchard, a new sequential tree building algorithm to reconstruct longitudinal phylogenies of SNVs and CNAs from single-cell sequencing. scaOrchard maximizes a new factorized longitudinal tree reconstruction objective by incrementally growing the tree, adding SNVs and CNAs in the order they are observed across samples. Applied to a cohort of 15 acute myeloid leukemias (AMLs) and 4 TP53-mutated AMLs, scaOrchard produced phylogenies that are biologically and temporally consistent with clinical observations, with many inferred CNAs validated by orthogonal bulk sequencing from the same patient. These reconstructions further reveal that AML clones selected after therapy are often defined by large-scale CNAs as well as SNVs, highlighting the role of CNAs in disease progression and resistance. More broadly, scaOrchard can help uncover how SNVs and CNAs jointly shape the evolutionary trajectories of individual cancers.

Keywords: tumor phylogeny · longitudinal sequencing · cancer evolution.

1 Introduction

Cancer evolves through the accumulation of somatic mutations, most commonly single nucleotide variants (SNVs) and copy number alterations (CNAs) [4, 12, 18]. This evolutionary process gives rise to genetically distinct populations of cancer cells, known as *clones*, each defined by a unique genome. The clonal composition of a cancer can change over time due to selective pressures such as immune response or therapy. Longitudinal sequencing, which involves sampling and sequencing tumor cells from the same patient at multiple time points, enables researchers to track these evolutionary dynamics and gain insight into how tumors adapt and progress. Recent advances in targeted single-cell DNA sequencing—particularly high-throughput microfluidic protocols known as *single-cell amplicon sequencing*—enable the detection of both SNVs and CNAs at single-cell resolution [10, 13]. Several studies have applied this technology to longitudinal tumor samples [10, 16], sequencing portions of oncogenes at ~40x per amplicon per cell. However, no computational methods currently exist to jointly reconstruct the longitudinal evolution of both SNVs and CNAs from this data—a gap this work addresses.

The evolutionary history of an individual cancer is typically represented as a phylogenetic tree, or *clone tree*, which depicts the ancestral relationships among cancer clones. Existing methods that reconstruct clone trees from longitudinal cancer samples focus exclusively on either SNVs [1, 6, 11] or CNAs [9], but not both. This largely reflects the limitations of previously available sequencing technologies. For example, single-cell whole-genome sequencing (scWGS) provides genome-wide coverage, enabling the detection of large-scale CNAs, but lacks the depth needed to reliably call SNVs. In contrast, older targeted single-cell sequencing technologies have sufficient depth to call SNVs, but too low of coverage

[1] *These authors are co-senior authors

to detect CNAs. Bulk DNA sequencing is useful for identifying SNVs and CNAs at the population level but lacks the resolution to detect changes in smaller subpopulations.

Two algorithms—BiTSC^2 [2] and COMPASS [14]—can jointly reconstruct clone trees containing both SNVs and CNAs from single-cell amplicon sequencing data, but neither is designed for longitudinal analysis. Applying them to longitudinal samples either requires reconstructing clone trees separately for each time point—which can yield disjoint evolutionary histories lacking temporal coherence—or pooling all samples together, which ignores sample-specific technical variation such as dropout rates and region-specific coverage. Failing to model sample-specific technical variation can lead to incorrect clone trees—clones may appear in samples where they were absent, or CNAs may be inferred from variations in coverage between samples, rather than genuine biological changes.

Here, we introduce scaOrchard, an algorithm for reconstructing longitudinal clone trees that depict the joint evolution of SNVs and CNAs from longitudinal single-cell sequencing. The method relies on a factorized longitudinal tree reconstruction objective—a novel formulation derived in this work. To optimize this objective, scaOrchard uses a sequential tree-building strategy in which mutations are added to the tree in the order they are first detected across longitudinal samples, ensuring that mutations detected earlier in time are placed earlier in evolution. Given S longitudinal samples, scaOrchard generates a sequence of trees, $\mathcal{T}^1, \mathcal{T}^2, \ldots, \mathcal{T}^S$, where each $\mathcal{T}^s$ describes the evolutionary history up to time point s and forms a subtree of all subsequent trees. Each subtree is refined through hill climbing, which allows SNVs to be relocated and CNAs to be added or removed to maximize tree likelihood. This design naturally enforces longitudinal constraints while remaining computationally efficient, scaling to tens or even hundreds of thousands of cells. Structuring inference in this way simplifies the estimation of sample-specific technical parameters—such as dropout rates and region-specific coverage—making CNA inference more reliable. Using both simulations and real data, we show that scaOrchard reconstructs temporally consistent and biologically plausible longitudinal clone trees describing the joint evolution of SNVs and CNAs.

We apply scaOrchard to published longitudinal single-cell datasets profiled with the Tapestri® platform (Mission Bio, Inc.), including a cohort of 15 acute myeloid leukemias (AMLs) undergoing diverse treatments, and four TP53-mutated AMLs sampled before and after venetoclax therapy. The inferred clone trees reveal a coherent picture of how these cancers evolved over time, with many CNAs identified by scaOrchard being orthogonally validated with bulk sequencing data from the same cancer. These reconstructed clone trees show that AML clones emerging after therapy or at relapse are frequently defined not only by SNVs but also by CNAs. While CNAs are well established drivers of oncogenesis, our results suggest that modeling them alongside SNVs is essential for fully characterizing key clonal populations.

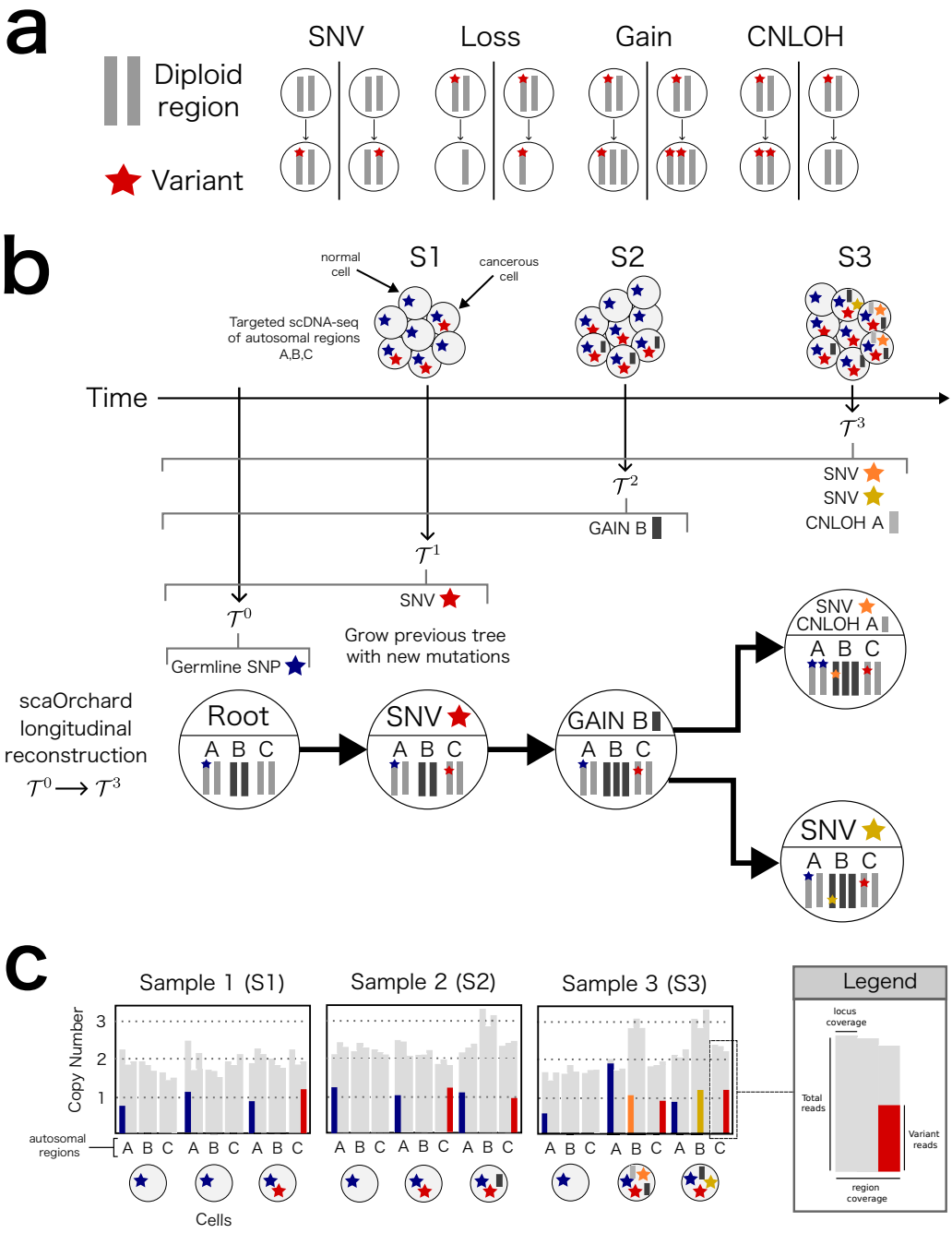
2 Materials and Methods

scaOrchard reconstructs longitudinal clone trees from targeted single-cell sequencing data collected at multiple time points from the same cancer. This section details the algorithm's main components: required sequencing inputs, clone tree representation, evolutionary model, likelihood model, and optimization strategy. Central to scaOrchard is a novel factorized longitudinal likelihood objective, which extends concepts from our previous work [7]. The optimization procedure searches for a sequence of subtrees that maximizes this factorized probability distribution over somatic mutations.

2.1 Inputs from single-cell amplicon sequencing experiment

Single-cell amplicon sequencing focuses on small genomic regions (150-450 base pairs) within disease-related genes. Amplicons—generated by amplifying the reverse transcription products of these regions—are used to measure the frequency of the target regions in the genome and identify variants. For each cell j, the data from a single-cell amplicon sequencing experiment is summarized as follows: D_j is the total number of reads mapped across all targeted regions,

Fig. 1. Overview of scaOrchard. **a.** Four types of somatic events are modeled. **b.** Longitudinal reconstruction: scaOrchard infers somatic events from targeted single-cell sequencing to build subtrees $\mathcal{T}^1, \mathcal{T}^2, \mathcal{T}^3$, representing evolution up to time of sampling. **c.** Inputs from longitudinal single-cell sequencing include total coverage of targeted regions and locus-level variant/reference read counts.



D_{kj} is the number of reads mapped to region k, D_{ij} is the total number of reads mapped to locus i, and A_{ij} is the number of variant (alternate) reads observed at locus i. Each cell is associated with a specific longitudinal sample, allowing the data to be partitioned into sample-specific subsets. We use $\mathbf{D}$ and $\mathbf{A}$ to denote the full dataset, containing read counts for all cells across all longitudinal samples. For a specific longitudinal sample s, we use $\mathbf{D}^s$ and $\mathbf{A}^s$ to represent the corresponding subset of data for cells originating from that sample.

2.2 Clone tree representation

Inferring a clone tree that includes both SNVs and CNAs involves reconstructing a rooted, directed tree $\mathcal{T}$, and clone assignment vector σ. Each node v in $\mathcal{T}$ represents a distinct clone, with the root corresponding to the germline (non-cancerous) clone, which may contain germline SNPs but no somatic SNVs. The clone assignment vector σ specifies the clone membership of each cell, such that if $\sigma_j = v$, cell j is assigned to clone v. The copy-number profile of cell j is determined by its assigned clone v, where $c_k(\sigma_j)$ denotes the total copy number for target region k, and $c_i^{(r)}(\sigma_j)$ and $c_i^{(a)}(\sigma_j)$ denote the number of copies of the reference and alternative alleles, respectively, at locus i.

2.3 Evolutionary Model

The scaOrchard algorithm models four types of somatic events: SNV acquisition, copy-number gains, copy-number losses, and copy-neutral loss of heterozygosity (CNLOH). CNAs may affect either individual target regions or entire chromosomes. Each variant locus lies within a target region, and its copy number can change if a CNA impacts the allele containing that region. Because targeted sequencing data are noisy and offer limited resolution for detecting multiple copy-number changes along the same lineage, we restrict each region to be impacted by at most one CNA per lineage in the tree. As a result, each region may have a total of 1,2, or 3 copies. SNV evolution in $\mathcal{T}$ follows the k-Dollo model [3], in which each SNV is acquired exactly once but may be lost at most once per lineage due to a deletion of the genomic region on which it resides. This combined evolutionary model for SNVs and CNAs is identical to that used by COMPASS [14].

2.4 Longitudinal Likelihood Model

Given S longitudinal samples, scaOrchard generates a sequence of clone trees $\mathcal{T}^1, \ldots, \mathcal{T}^S$ to maximize the following factorized probability distribution over somatic events:

$$\prod_{s=1}^{S} P(\mathbf{D}^s, \mathbf{A}^s \mid \mathcal{T}^s) \tag{1}$$

Each $\mathcal{T}^s$ represents the evolutionary history up to time point s and is a subtree of all subsequent trees, $\mathcal{T}^1 \subseteq \mathcal{T}^2 \subseteq \cdots \subseteq \mathcal{T}^S$, reflecting the accumulation of somatic mutations over time. The sample-specific likelihood is:

$$\begin{aligned} P(\mathbf{D}^s, \mathbf{A}^s \mid \mathcal{T}^s) = & \prod_{\text{cell } j} \sum_{\sigma_j} \pi_{\sigma_j} \\ & \times \prod_{\text{region } k} \underbrace{P\big(D_{kj} \mid c_k(\sigma_j), \rho_k\big)}_{\text{Read depth likelihood}} \\ & \times \prod_{\text{locus } i} \underbrace{P\big(A_{ij} \mid D_{ij}, c_i^{(r)}(\sigma_j), c_i^{(a)}(\sigma_j), \mu_i\big)}_{\text{Allelic read count likelihood}}. \end{aligned} \tag{2}$$

This likelihood marginalizes over all possible clonal assignments for σ_j, with π_{σ_j} denoting the prior probability of assignment to clone σ_j. The read depth likelihood, which evaluates the inferred total copy numbers for each target region in each cell, and the allelic read count likelihood, which evaluates the inferred number of variant and reference allele copies at each locus in each cell, are described in the sections below.

The factorized distribution in Eq. 1 approximates the joint distribution over the data from all S longitudinal samples:

$$\begin{aligned}\prod_{s=1}^{S} P(\mathbf{D}^s, \mathbf{A}^s \mid \mathcal{T}^s) &\approx P(\mathbf{D}^1, \ldots, \mathbf{D}^S, \mathbf{A}^1, \ldots, \mathbf{A}^S \mid \mathcal{T}) \\ &\approx P(\mathbf{D}, \mathbf{A} \mid \mathcal{T}).\end{aligned} \tag{3}$$

The left and right hand sides of Eq 3 are equivalent if and only if

$$\mathbf{D}^s, \mathbf{A}^s \perp\!\!\!\perp \mathbf{D}^{s^>}, \mathbf{A}^{s^>} \mid \mathbf{D}^{s^<}, \mathbf{A}^{s^<}, \tag{4}$$

which states that the read count data for sample s is conditionally independent of all data from future samples, $s^>$, given the data from all prior samples, $s^<$. This conditional independence assumption reflects the idea that longitudinal sequencing captures tumor evolution over time. Genomic changes arising between two time points should be detectable, provided that cells carrying those mutations are sampled.

Read Depth Likelihood Current single-cell amplicon sequencing technologies have non-uniform coverage across amplicons [14]. Since each amplicon covers a portion of a disease-specific gene (i.e., a region), the probability a read falls into region k, denoted with ρ_k, is impacted by a variety of factors including the size of region, its GC content, etc.. We estimate ρ_k using the observed fraction of reads that fall into region k in a population of normal (i.e., absent of CNAs) cells. For a cell j with D_j total reads, assigned to clone σ_j, and copy number $c_l(\sigma_j)$ for region l, the expected read depth in region k is given by

$$\mathbb{E}(D_{kj}) = D_j \frac{c_k(\sigma_j)\rho_k}{\sum_l c_l(\sigma_j)\rho_l} \tag{5}$$

Cells assigned to the same clone may have larger than expected variances in their regional read depths, a phenomenon known as *overdispersion*. To model this, we use a Negative Binomial distribution for each clone's regional read counts, parameterized by the mean $\mu = \mathbb{E}(D_{kj})$ and inverse-dispersion parameter θ:

$$P\left(D_{kj} \middle| c_k(\sigma_j), \rho_k, \theta\right) = \frac{\Gamma(D_{kj}+\theta)}{\Gamma(D_{kj}+1)\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^{\theta} \left(\frac{\mu}{\theta+\mu}\right)^{D_{kj}} \tag{6}$$

This parametrization, often referred to as the *overdispersed Poisson*, differs from the standard Poisson in that its variance is $\mu + \frac{\mu^2}{\theta}$, while the standard Poisson has variance equal to μ. Following [14], we set $\theta = 6.0$ to account for overdispersion; setting $\theta = \infty$ assumes no overdispersion and reduces it to the Poisson. We omit θ from the read depth likelihood in Eq. 2, since it is treated as fixed.

Allelic Read Count Likelihood Although single-cell amplicon sequencing has low coverage, allelic imbalances resulting from CNAs can still be detected from shifts in the observed proportions of variant and reference read counts. However, inferring the relative proportions of mutant and reference alleles from these read counts is complicated by overdispersion, which arises from allelic biases during amplification. To address this, the allelic read counts are modeled

using a beta-binomial distribution. Using this model, the likelihood of observing A variant reads out of D total reads at a position, given $c^{(a)}$ mutant alleles and $c^{(r)}$ reference alleles, is:

$$P(A \mid D, c^{(a)}, c^{(r)}, \omega, \epsilon) = \frac{\text{Beta}\left(A + \omega f, D - A + \omega(1-f)\right)}{\text{Beta}\left(\omega f, \omega(1-f)\right)}, \tag{7}$$

where $\omega > 0$ is the beta-binomial concentration parameter, $\text{Beta}(\cdot)$ denotes the Beta function, and f is the expected fraction of variant reads assuming a fixed sequencing error rate of ϵ:

$$f = (1-\epsilon)\frac{c^{(a)}}{c^{(a)} + c^{(r)}} + \epsilon\frac{c^{(r)}}{c^{(a)} + c^{(r)}}.$$

Allelic read counts are also impacted by allele-specific dropout during sequencing, leading to false negatives. To account for this, we introduce global allele-specific dropout rate parameters, $\mu_i^{(a)}$ and $\mu_i^{(r)}$, representing the dropout probabilities at locus i for the alternate and reference allele, respectively. The number of amplified alleles can be modeled using a binomial distribution. Specifically, the likelihood that l alternate alleles and k reference alleles are amplified at locus i during sequencing when there are $c_i^{(a)}$ alternate alleles and $c_i^{(r)}$ reference alleles is given by:

$$P(l, k \mid c_i^{(a)}, c_i^{(r)}, \mu_i^{(a)}, \mu_i^{(r)}) = \text{Binom}(l \mid c_i^{(a)}, \mu_i^{(a)}) \times \text{Binom}(k \mid c_i^{(r)}, \mu_i^{(r)}),$$

where $\text{Binom}(k \mid n, p) = \binom{n}{k} p^k (1-p)^{n-k}$ is the binomial likelihood of observing k successes out of n trials with a success rate of p. To fully account for dropout uncertainty, the likelihood of the observed read count A is obtained by summing over all possible dropout scenarios:

$$P(A \mid D, c_i^{(a)}, c_i^{(r)}, \mu_i, \omega, \epsilon) = \sum_{\substack{0 \le l \le c_i^{(a)} \\ 0 \le k \le c_i^{(r)} \\ (k,l) \ne (0,0)}} P(l, k \mid c_i^{(a)}, c_i^{(r)}, \mu_i^{(a)}, \mu_i^{(r)}) P(A \mid D, c_i^{(a)}, c_i^{(r)}, \omega, \epsilon). \tag{8}$$

For succinctness, we omit the parameters ω, ϵ, defining the shorthand used in Eq 2:

$$P(A \mid D, c_i^{(a)}, c_i^{(r)}, \mu_i, \omega, \epsilon) \equiv P(A \mid D, c_i^{(a)}, c_i^{(r)}, \mu_i). \tag{9}$$

2.5 Tree Prior

Single-cell sequencing data are noisy, and complex evolutionary models can overfit the data by introducing spurious clones or CNAs. To mitigate this, we impose a prior on the number of clones and CNAs in the inferred tree.

Each CNA event incurs a penalty, with CNLOH assigned a stronger penalty than a gain or loss. CNAs affecting contiguous regions are counted as a single event, as large-scale alterations (e.g., whole-chromosome changes) are common. We also penalize the total number of clones, since incorporating more SNVs and CNAs into the tree increases the chance that multiple mutations originated in the same clone. Both penalties scale linearly with the number of cells, as larger datasets exert a proportionally greater influence on the likelihood. Let p_1 denote the penalty for the number of clones and p_2 the penalty for CNAs. The prior is defined as:

$$\log(P(\mathcal{T})) = -p_1 n_{\text{muts}} n_{\text{nodes}} - p_2(n_{\text{gains}} + n_{\text{losses}} + 2n_{\text{CNLOH}}) + \text{Constant}, \tag{10}$$

where n_{muts}, n_{nodes}, n_{gains}, n_{losses}, and n_{CNLOH} denote, respectively, the number of mutations, clones, non-contiguous gains, non-contiguous losses, and non-contiguous CNLOH events. .

2.6 Algorithm

The scaOrchard algorithm uses a stochastic hill climbing approach to generate a sequence of clone trees $\mathcal{T}^1, \ldots, \mathcal{T}^S$ that maximize Eq 1. The procedure consists of three stages.

Hyperparameters Normal (non-cancerous) cell populations are first identified in each sample to estimate region-specific coverage probabilities ρ_k for each region k. scaOrchard does this by independently reconstructing an SNV-only tree for each longitudinal sample, maximizing the sample-specific likelihood in Eq. 2, and designating cells assigned to the root node as the normal population. Finally, the sample in which each SNV is first detected is determined from the single-cell sequencing inputs.

Tree initialization The tree $\mathcal{T}^0$ is initialized with a root node that represents a normal cell population, which has germline SNPs but no SNVs.

Stochastic hill climbing For each sample s, SNVs first observed in that sample are incorporated into $\mathcal{T}^{s-1}$ by forming new clones that descend from existing ones. After introducing new SNVs, stochastic hill climbing proceeds for η epochs. At each iteration, either a newly added SNV or a genomic region is selected at random, and all valid modifications are evaluated. Three types of moves are permitted: (1) relocate an SNV, (2) remove a CNA, or (3) add a CNA. CNA moves are constrained so that only regions not previously altered by a CNA in $\mathcal{T}^{s-1}$ may be modified. All tree modifications respect longitudinal constraints: clones in $\mathcal{T}^{s-1}$ remain fixed, and new SNVs/CNAs are restricted to descendant clones. Tree modifications are accepted if they increase the likelihood (Eq. 2).

2.7 Inferring Clone Assignment Priors and Dropout Rates

Dropout rates and clone assignment priors are estimated for each sample s using its corresponding tree $\mathcal{T}^s$. We employ the expectation–maximization (EM) algorithm from [14] to jointly estimate the prior probability of assignment to each clone v, π_v, and the allele specific dropout rates μ_i, for each locus i. We place a Beta(α, β) prior on μ_i centered at 0.05, taking advantage of its conjugacy with the binomial likelihood used to model dropout rates. For clone assignments, we use a flat Dirichlet$(1, \ldots, 1)$ prior, which is conjugate to the categorical distribution for clone prior probabilities. This EM algorithm is run after every tree modification during stochastic hill climbing. There are two types of latent variables for each cell j: the clone assignment σ_j, and the number of amplified alternate and reference alleles for each locus i, $C_{ij}^{(a)}$ and $C_{ij}^{(r)}$. Below, we present the update equations for the E- and M-steps, referring the reader to [14] for a complete derivation and description of the algorithm. The update equations for the E-step are defined as:

$$Q(\sigma_j = v) = P(\sigma_j = v \mid \mathbf{D}_j, \mathbf{A}_j, \pi, \mu) = \frac{\pi_v P(\mathbf{D}_j, \mathbf{A}_j \mid \sigma_j = v, \pi, \mu)}{\sum_{v'} \pi_{v'} P(\mathbf{D}_j, \mathbf{A}_j \mid \sigma_j = v', \pi, \mu)}, \tag{11}$$

$$\begin{aligned} Q(C_{ij}^{(r)} = k, C_{ij}^{(a)} = l \mid \sigma_j = v) &= P(C_{ij}^{(r)} = k, C_{ij}^{(a)} = l \mid D_{ij}, A_{ij}, \sigma_j = v, \mu) \\ &= \frac{\mu_i^{c_i^{(r)}(v) + c_i^{(a)}(v) - k - l} (1 - \mu_i)^{k+l} P(A_{ij} \mid D_{ij}, k, l)}{\sum_{k', l'} \mu_i^{c_i^{(r)}(v) + c_i^{(a)}(v) - k' - l'} (1 - \mu_i)^{k'+l'} P(A_{ij} \mid D_{ij}, k, l)}. \end{aligned}$$

The update equations for the M-step are defined as:

$$\pi_v = \frac{1}{n_{\text{cells}}} \sum_{\text{cell } j} Q(\sigma_j = v) \tag{12}$$

$$\mu_i = \frac{\alpha - 1 + \sum_{\text{node } v} \sum_{\text{cell } j} Q(\sigma_j = v) \sum_{k,l} Q(C_{ij}^{(r)} = k, C_{ij}^{(a)} = l \mid \sigma_j = v)(c_i^{(r)}(v) + c_i^{(a)}(v) - k - l)}{\alpha + \beta - 2 + \sum_{\text{node } v} \sum_{\text{cell } j} Q(\sigma_j = v)(c_i^{(r)}(v) + c_i^{(a)}(v))} \tag{13}$$

3 Results

3.1 Evaluation on Simulated Longitudinal Data

We benchmarked scaOrchard on simulated longitudinal cancer data, comparing it to COMPASS [14], LACE [1], and SCITE [5]. COMPASS reconstructs clone trees incorporating both SNVs and CNAs but does not account for longitudinal sampling. LACE is designed for longitudinal data but models only SNVs. SCITE is a widely used method for SNV-based tree reconstruction, though it is not designed for longitudinal data. We also considered BiTSC^2[2], which supports joint reconstruction of SNVs and CNAs, but excluded it due to prohibitively long runtimes, even on small longitudinal datasets.

We simulated longitudinal cancer datasets to assess the benefits of jointly inferring SNVs and CNAs while explicitly modeling longitudinal sampling, compared to methods that either infer only SNVs or are not designed for longitudinal data. We simulated 40 longitudinal cancers in total: 10 for each setting of longitudinal samples (2, 3, 4, or 5). Each dataset spanned 2–5 time points, with 3,000 cells sampled per time point (6,000–15,000 cells per dataset). Sequencing data were generated to resemble output from the Tapestri ® platform (see Appendix C), comprising 20 genomic regions, 20 SNVs, and 3 CNAs. Ground-truth clone trees included 6–9 clones, with the number of clones increasing in datasets with more time points, reflecting the expectation that more longitudinal sampling reveals greater clonal diversity. The longitudinal dynamics for each simulated cancer were generated using a Moran process [8], a well-established model in population genetics.

Performance was evaluated using cell-to-clone assignment accuracy—comparing the mutant and total copy number profiles of each cell's assigned clone to its ground-truth profile. Specifically, we computed the mean absolute error (MAE) between mutant copy numbers (MCN) and total copy numbers (TCN) per cell, normalized by the total number of cells and loci/regions in the dataset (denoted MCN-MAE and TCN-MAE, respectively). We also introduce a new metric, the false emergence rate (FER), that measures the fraction of cells assigned copy numbers found exclusively in clones from future samples and absent from all clones detected up to the cell's sampling time. FER captures a key pitfall of ignoring longitudinal structure: mutations may appear to emerge earlier in evolution than they actually did. Detailed descriptions for these metrics are provided in Appendix B.

Figure 2 shows the performance of scaOrchard and competing methods on the 40 simulated longitudinal cancer datasets. scaOrchard on average achieved the best performance (lowest error) across all evaluation metrics, regardless of the number of longitudinal samples, and showed consistently strong results—particularly for total copy number estimation. In contrast, COMPASS struggled with total copy number inference, with an average TCN-MAE exceeding 0.35 in all simulation settings. This is expected since it does not model coverage variation between longitudinal samples, which confounds CNA inference. For mutant copy numbers, scaOrchard achieved the best average performance. While COMPASS and SCITE showed slightly lower MCN-MAE on a subset of datasets, the average difference was small—approximately 6×10^{-3}.

SCITE estimates whether loci are heterozygous or homozygous but does not infer total copy numbers. Accordingly, we report mutant copy number results for SCITE but exclude it from TCN-MAE results. LACE outputs a binary mutation matrix, so we include it in mutant copy number comparisons while acknowledging that calls beyond presence/absence are not possible; it is likewise excluded from the TCN-MAE results. However, LACE performed well on

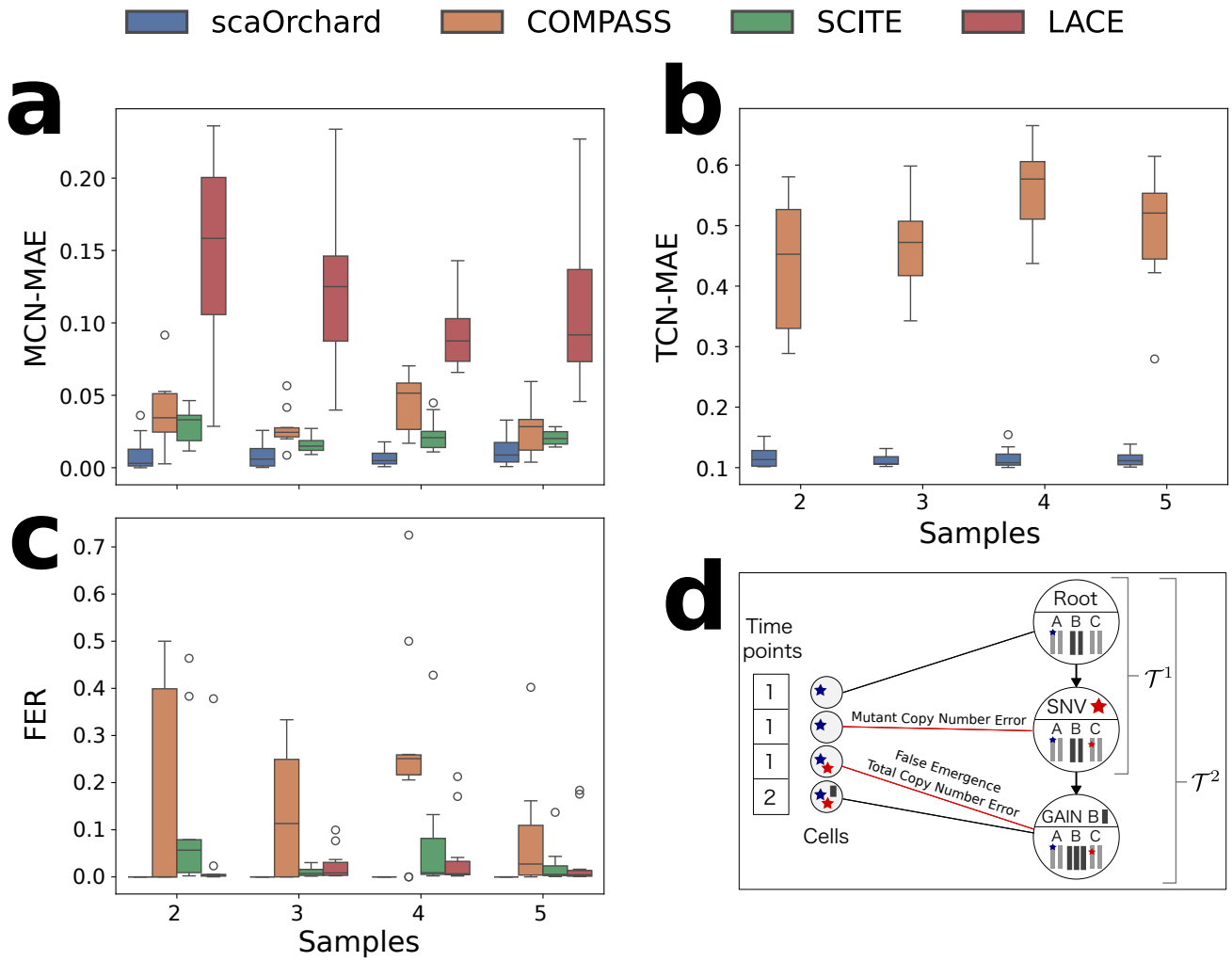


Fig. 2. Results from scaOrchard, COMPASS, LACE, and SCITE on 40 simulated cancers. **a-c.** Box plots for each evaluation metric show the median (center bar), first and third quartiles (boxes above and below center bar), range of data (whiskers) and outliers (unfilled circles). Lower values indicate more accurate reconstructions. **d.** Examples of cell assignments which result in the three different error types.

the FER metric, as expected given its longitudinal modeling. By design, scaOrchard achieved perfect FER, in contrast to COMPASS, which frequently inferred mutant and/or total copy number changes occurred earlier in evolution than they actually did.

Finally, scaOrchard was substantially faster than COMPASS, completing datasets with 5 longitudinal samples in an average of about 2 minutes, compared to roughly 1 hour and 40 minutes for COMPASS. Runtime comparisons for all methods on simulated datasets are shown in Appendix D.

3.2 Longitudinal Phylogenetic Analysis of Acute Myleoid Leukemias

We applied scaOrchard to a cohort of 15 longitudinally observed acute myeloid leukemias (AMLs) profiled with the Tapestri® platform [10], each with 2–5 time points sequenced using either a 50-amplicon panel (19 genes) or a 279-amplicon panel (37 genes). The authors in [10] provided clinical descriptions for all of these AMLs, detailing treatments and changes in the clonal composition of the AMLs over time. Our analyses with scaOrchard are consistent with these clinical descriptions, and reveal that AML clones dominant after treatment or at relapse frequently harbor both CNAs and SNVs, demonstrating the clinical relevance of jointly modeling SNVs and CNAs in a longitudinal framework. Below, we highlight several representative cases, including AML-99 and two AMLs where we directly compare reconstructions from scaOrchard and COMPASS. Analyses of the remaining AMLs, including a second cohort consisting of four TP53-mutated cases collected before and after venetoclax treatment [16], are provided in Appendix E.

Dominant AML Clones Are Characterized By Both CNAs and SNVs AML-99 illustrates how incorporating longitudinal information and modeling CNAs yields more biologically meaningful clone trees. This patient was diagnosed with AML carrying an IDH2 R140Q mutation (AML-99-001) and had two additional samples collected during therapy (AML-99-002, AML-99-003). The patient remained in remission until relapse (AML-99-004), after which therapy was adjusted, though the disease remained refractory (AML-99-005).

Figure 3a shows the tree reconstructed by scaOrchard, revealing that CNA-driven clones dominated as the disease evolved. Across the first three samples (AML-99-001 through AML-99-003), the dominant clone carried a MYC/RAD21 gain and CNLOH of the RUNX1 reference allele, consistent with the bulk ASCAT [17] profile for AML-99-001 (Figure 3b). At relapse, however, a new clone emerged with the same MYC/RAD21 gain but CNLOH of the RUNX1 alternate allele, dominating AML-99-004 (Figure 3c). After therapy was adjusted, the relapse clone contracted, while a FLT3-positive subclone of the original MYC/RAD21 clone expanded to dominate AML-99-005 (Figure 3c).

The CNLOH of the RUNX1 alternate allele and the emergence of the NRAS p.G60E mutation both occur at relapse. Nearly all cells at relapse carrying the NRAS mutation had near-zero VAFs for RUNX1 p.R162K, whereas NRAS-wildtype cells from earlier samples had RUNX1 VAFs near 100%. These CNLOH events are unlikely to result from dropout, as nearly all cells in both NRAS p.G60E-positive (1272/1404) and NRAS-wildtype (11,231/12,381) groups have ≥ 10 reads at the RUNX1 p.R162K locus, with median coverages of 43 and 33, respectively. Such high, consistent coverage supports a genuine biological event rather than a technical artifact. Matched bulk or SNP array data for AML-99-004 could verify these findings.

We provide COMPASS's reconstruction for AML-99 in Appendix E, which agrees poorly with the clinical description for this patient and infers CNAs not supported by bulk sequencing data. Across both cohorts—the 15 AMLs from [10] and the 4 TP53-mutated AMLs from [16]—scaOrchard reconstructed clone trees where CNA-driven clones dominated in 17 of 19 AMLs. Bulk data exists for 5/17 of these CNA-dominated AMLs, supporting the majority of CNAs inferred by scaOrchard.

Comparing and contrasting scaOrchard and COMPASS longitudinal AML reconstructions Figure 4 shows longitudinal reconstructions by scaOrchard and COMPASS for AML-01 and AML-83. These two cases highlight the importance of longitudinal constraints and estimating sample-specific coverage and dropout rates.

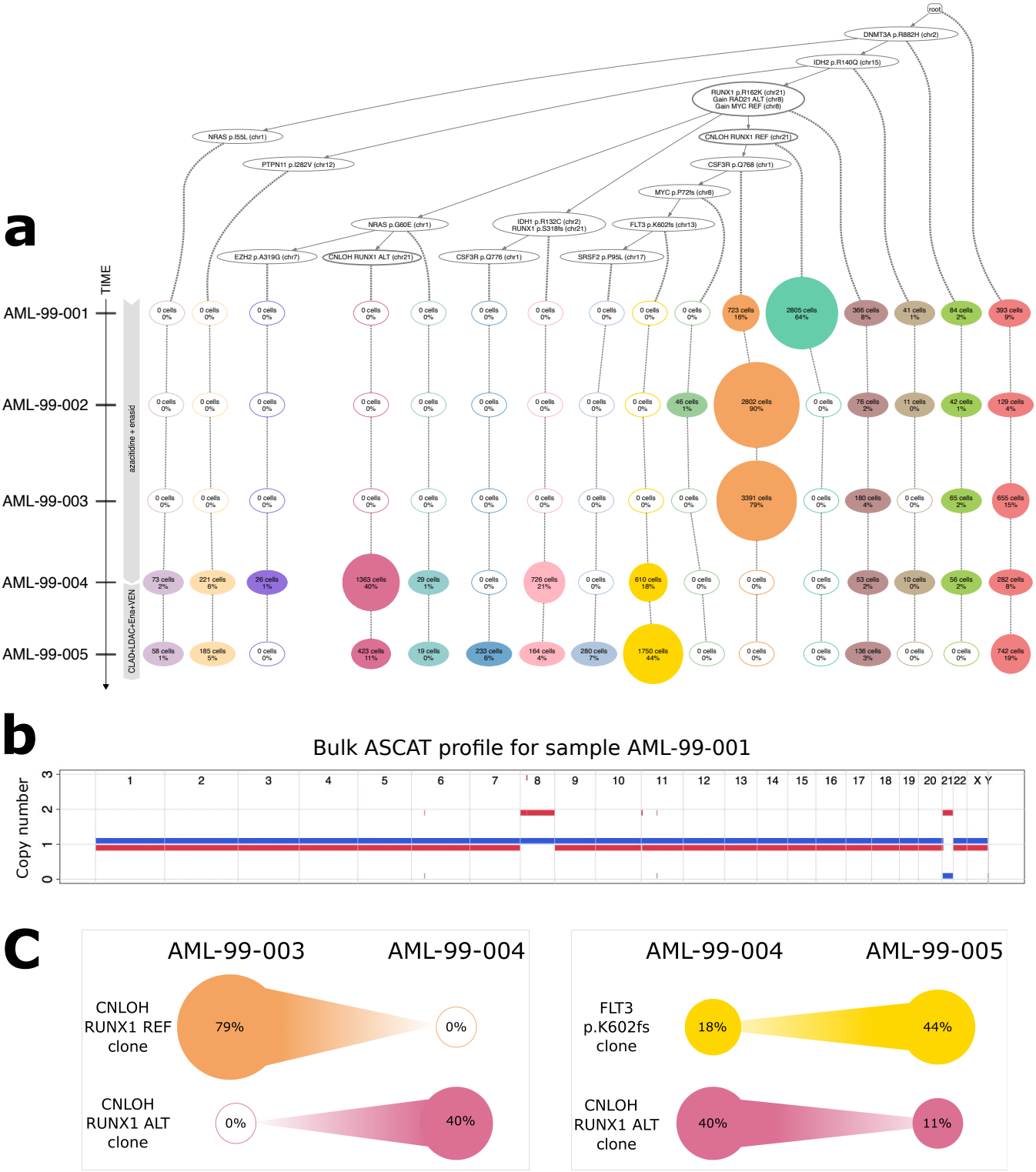


Fig. 3. a. Longitudinal clone tree reconstructed by scaOrchard. **b.** ASCAT profile for sample AML-99-001 from [14]. **c.** Fluctuations in size of key clonal populations in response to therapy.

Figure 4a shows the trees reconstructed by scaOrchard and COMPASS for AML-01. This patient was diagnosed with FLT3-ITD AML, with samples collected at diagnosis (AML-01-001) and at relapse (AML-01-002). The patient remained positive for FLT3-ITD after therapy. Although both scaOrchard and COMPASS infer a CNLOH of the FLT3 reference allele, their reconstructions differ in its placement. scaOrchard places the FLT3 CNLOH before the two WT1 mutations, identifying 85% of cells at diagnosis as positive for this event. COMPASS places the CNLOH event after the first WT1 mutation, assigning 86% of cells at diagnosis both WT1 mutations and the CNLOH event—even though no WT1 mutations were detected at diagnosis. This high rate of false emergences occurs because, without WT1 mutation data at diagnosis or longitudinal constraints, COMPASS relies solely on the strong CNLOH signal when assigning cells to clones. This also results in COMPASS estimating an unrealistically high dropout rates of 50% for the WT1 loci, highlighting the importance of longitudinal constraints and modeling sample-specific dropout rates for reconstructing plausible longitudinal clone trees.

Figure 4b shows the trees reconstructed by scaOrchard and COMPASS for AML-83. This patient is characterized at diagnosis (AML-83-001) by an IDH2 p.R172K mutation. After multiple rounds of therapy, the patient relapsed (AML-83-002) with two NRAS mutations not detected at diagnosis. Bulk CNV analysis using CNVkit [15], as reported in [14], supports a copy number gain on chromosome 8 at relapse (Figure 4b top left). scaOrchard reconstructs an evolutionary history consistent with the case description in [10], correctly identifying the MYC/RAD21 gain on chromosome 8 at relapse. The MYC/RAD21 clone dominates at relapse, giving rise to three subclones distinguished by distinct FLT3/NRAS SNVs. In contrast, while COMPASS generally recovers the correct SNV evolution, it misses the MYC/RAD21 gain. This occurs because it does not model sample-specific coverage, so differences between the two sequencing experiments—particularly among the "normal" cells used to define baseline coverage—distort the signal needed to detect the gain. When run only on the relapse sample (AML-83-002), where sample-specific coverage does not need to be modeled, COMPASS reconstructs a tree nearly identical to scaOrchard's, including the MYC/RAD21 gain. However, analyzing the two AML-83 samples independently with COMPASS produces inconsistent clone structures across the trees, obscuring the temporal ordering of genomic changes and the relationships between cells from the two samples.

In [10], the cohort of 15 AMLs was analyzed with SCITE [5] by pooling all time points and inferring clone trees based solely on SNVs. Of these cases, only AML-99 was later reanalyzed with COMPASS [14], which reconstructed separate trees for each time point incorporating both SNVs and CNAs. In contrast, scaOrchard provides the first coherent view of how CNAs and SNVs jointly shaped evolution in these AMLs over time. Analyses of both AML cohorts (Appendix E) recapitulate findings similar to AML-01 and AML-83 highlighted here, emphasizing the advantages of scaOrchard's modeling.

4 Discussion

Longitudinal single-cell sequencing is increasingly used to study cancer evolution, particularly in leukemias and lymphomas. Methods exist to reconstruct clone trees from this data; however, they typically model the evolution of either SNVs or CNAs, but not both. Recent advances in targeted single-cell sequencing technologies now allow reliable detection of both CNAs and SNVs, enabling detailed longitudinal tracking of these alterations at single-cell resolution. Here, we show that longitudinal clone tree reconstruction from this new single-cell data reveals that dominant leukemia clones are often defined by both SNVs and CNAs, and that these clones frequently expand or contract after therapy. These insights are made possible by our algorithm, scaOrchard, which optimizes a novel factorized objective to reconstruct clone trees from longitudinal single-cell sequencing. We show that scaOrchard reconstructions are consistent with clinical observations and supported by orthogonal bulk sequencing data from the same cancers. The resulting trees provide contiguous joint histories of how SNVs and CNAs emerge and evolve over time within individual cancers. The accuracy of our algorithm is further supported by simulations, where scaOrchard consistently outperformed existing reconstruction methods across multiple evaluation metrics.

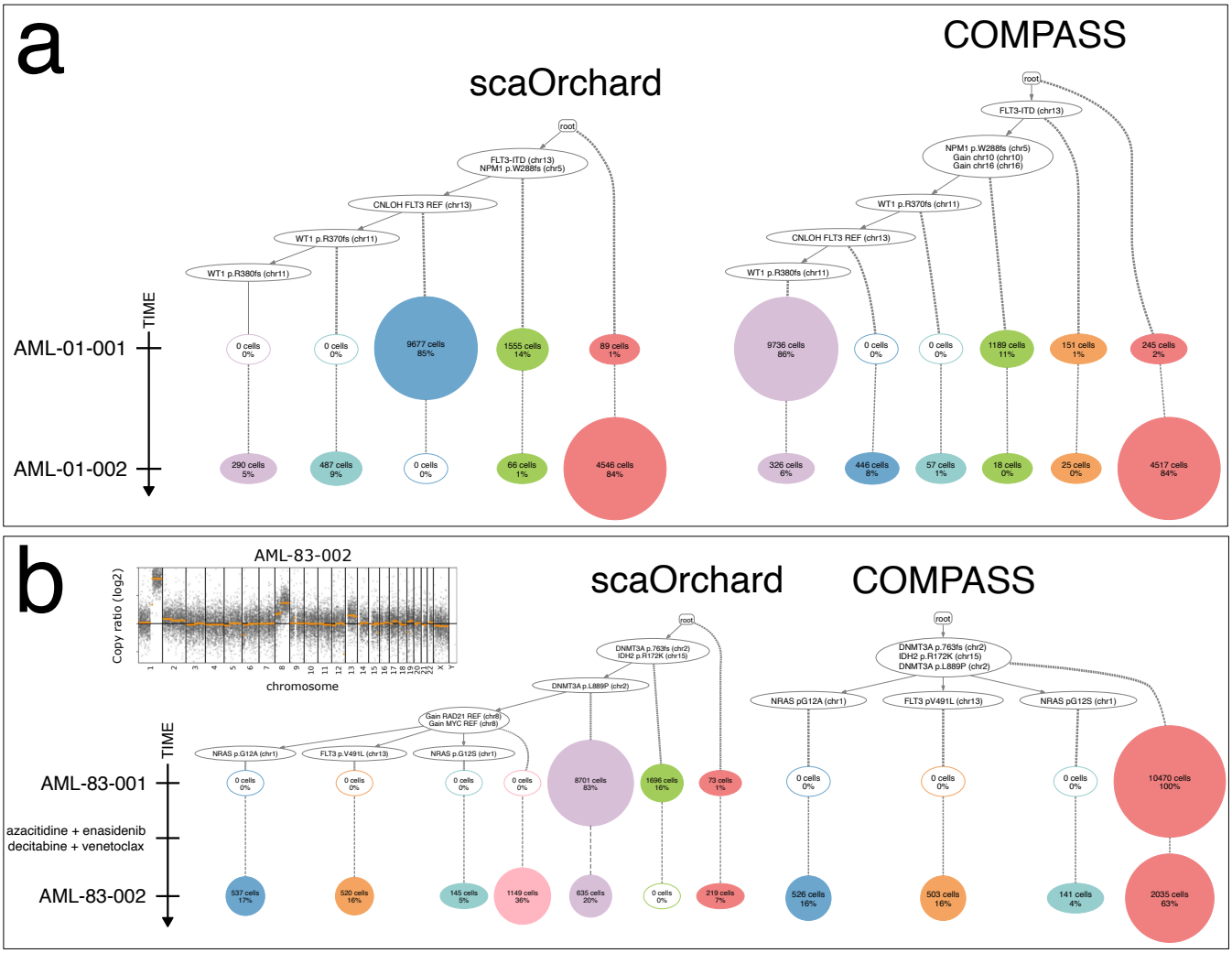


Fig. 4. Longitudinal reconstructions by scaOrchard and COMPASS for AML-01 (**a**), and AML-83 (**b**).

The current scaOrchard algorithm has a few notable limitations. First, it relies on a greedy search strategy, which can become trapped in local maxima. While targeted single-cell sequencing experiments typically profile thousands of cells, making the reconstruction problem highly constrained, sparsely sampled subclones may be excluded from the inferred clone tree under this approach. This issue can be mitigated by increasing the number of cells sampled at each time point. Second, our copy number evolution model is potentially too simplistic: it only allows each region to be impacted by at most one CNA per lineage, models a restricted set of CNA types, and does not account for homozygous loss—a rare but possible event.

Future work could address these limitations. A more comprehensive copy number model that incorporates homozygous loss and multiple CNAs affecting the same region within a lineage is feasible, though these additions may become more relevant as single-cell technologies advance to better capture such events. The search strategy could be improved

by replacing the greedy approach with a Markov Chain Monte Carlo framework, reducing the risk of getting trapped in local maxima. However, this would require a carefully designed proposal distribution and acceptance criteria to avoid introducing spurious CNAs. Finally, while scaOrchard was developed for targeted single-cell sequencing, the framework could be extended to reconstruct longitudinal clone trees from single-cell whole-genome sequencing or even bulk sequencing data.

References

1. Ascolani, G., Angaroni, F., Maspero, D., Craighero, F., Bhavesh, N.L.S., Piazza, R., Damiani, C., Ramazzotti, D., Antoniotti, M., Graudenzi, A.: LACE 2.0: an interactive R tool for the inference and visualization of longitudinal cancer evolution. BMC Bioinformatics **24**(1), 99 (Mar 2023). https://doi.org/10.1186/s12859-023-05221-3, https://doi.org/10.1186/s12859-023-05221-3
2. Chen, Z., Gong, F., Wan, L., Ma, L.: BiTSC2: Bayesian inference of tumor clonal tree by joint analysis of single-cell SNV and CNA data. Briefings in Bioinformatics **23**(3), bbac092 (May 2022). https://doi.org/10.1093/bib/bbac092, https://doi.org/10.1093/bib/bbac092
3. Dollo, L.: THE LAWS OF EVOLUTION
4. Hanahan, D., Weinberg, R.A.: The Hallmarks of Cancer. Cell **100**(1), 57–70 (Jan 2000). https://doi.org/10.1016/S0092-8674(00)81683-9, https://www.cell.com/cell/abstract/S0092-8674(00)81683-9
5. Jahn, K., Kuipers, J., Beerenwinkel, N.: Tree inference for single-cell data. Genome Biology **17**(1), 86 (May 2016). https://doi.org/10.1186/s13059-016-0936-x, https://doi.org/10.1186/s13059-016-0936-x
6. Khan, R., Mallory, X.: scLongTree: an accurate computational tool to infer the longitudinal tree for scDNAseq data (Nov 2023). https://doi.org/10.1101/2023.11.11.566680, https://www.biorxiv.org/content/10.1101/2023.11.11.566680v1, pages: 2023.11.11.566680 Section: New Results
7. Kulman, E., Kuang, R., Morris, Q.: Orchard: building large cancer phylogenies using stochastic combinatorial search (Nov 2023). https://doi.org/10.48550/arXiv.2311.12917, http://arxiv.org/abs/2311.12917
8. Lacy, M.S., Jenner, A.L.: Impact of Resistance on Therapeutic Design: A Moran Model of Cancer Growth. Bulletin of Mathematical Biology **86**(4), 43 (2024). https://doi.org/10.1007/s11538-024-01272-6, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10950993/
9. Lu, B., Curtius, K., Graham, T.A., Yang, Z., Barnes, C.P.: CNETML: maximum likelihood inference of phylogeny from copy number profiles of multiple samples. Genome Biology **24**(1), 144 (Jun 2023). https://doi.org/10.1186/s13059-023-02983-0, https://doi.org/10.1186/s13059-023-02983-0
10. Morita, K., Wang, F., Jahn, K., Hu, T., Tanaka, T., Sasaki, Y., Kuipers, J., Loghavi, S., Wang, S.A., Yan, Y., Furudate, K., Matthews, J., Little, L., Gumbs, C., Zhang, J., Song, X., Thompson, E., Patel, K.P., Bueso-Ramos, C.E., DiNardo, C.D., Ravandi, F., Jabbour, E., Andreeff, M., Cortes, J., Bhalla, K., Garcia-Manero, G., Kantarjian, H., Konopleva, M., Nakada, D., Navin, N., Beerenwinkel, N., Futreal, P.A., Takahashi, K.: Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. Nature Communications **11**(1), 5327 (Oct 2020). https://doi.org/10.1038/s41467-020-19119-8, https://www.nature.com/articles/s41467-020-19119-8, publisher: Nature Publishing Group
11. Myers, M.A., Satas, G., Raphael, B.J.: CALDER: Inferring Phylogenetic Trees from Longitudinal Tumor Samples. Cell Systems **8**(6), 514–522.e5 (Jun 2019). https://doi.org/10.1016/j.cels.2019.05.010, https://www.cell.com/cell-systems/abstract/S2405-4712(19)30191-7
12. Nowell, P.C.: The Clonal Evolution of Tumor Cell Populations. Science **194**(4260), 23–28 (Oct 1976). https://doi.org/10.1126/science.959840, https://www.science.org/doi/abs/10.1126/science.959840
13. Pellegrino, M., Sciambi, A., Treusch, S., Durruthy-Durruthy, R., Gokhale, K., Jacob, J., Chen, T.X., Geis, J.A., Oldham, W., Matthews, J., Kantarjian, H., Futreal, P.A., Patel, K., Jones, K.W., Takahashi, K., Eastburn, D.J.: High-throughput single-cell DNA sequencing of acute myeloid leukemia tumors with droplet microfluidics. Genome Research **28**(9), 1345–1352 (Sep 2018). https://doi.org/10.1101/gr.232272.117, https://genome.cshlp.org/content/28/9/1345, company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab

14. Sollier, E., Kuipers, J., Takahashi, K., Beerenwinkel, N., Jahn, K.: COMPASS: joint copy number and mutation phylogeny reconstruction from amplicon single-cell sequencing data. Nature Communications **14**(1), 4921 (Aug 2023). https://doi.org/10.1038/s41467-023-40378-8, https://www.nature.com/articles/s41467-023-40378-8, publisher: Nature Publishing Group
15. Talevich, E., Shain, A.H., Botton, T., Bastian, B.C.: CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLOS Computational Biology **12**(4), e1004873 (Apr 2016). https://doi.org/10.1371/journal.pcbi.1004873, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004873, publisher: Public Library of Science
16. Thijssen, R., Diepstraten, S.T., Moujalled, D., Chew, E., Flensburg, C., Shi, M.X., Dengler, M.A., Litalien, V., MacRaild, S., Chen, M., Anstee, N.S., Reljić, B., Gabriel, S.S., Djajawi, T.M., Riffkin, C.D., Aubrey, B.J., Chang, C., Tai, L., Xu, Z., Morley, T., Pomilio, G., Bruedigam, C., Kallies, A., Stroud, D.A., Bajel, A., Kluck, R.M., Lane, S.W., Schoumacher, M., Banquet, S., Majewski, I.J., Strasser, A., Roberts, A.W., Huang, D.C.S., Brown, F.C., Kelly, G.L., Wei, A.H.: Intact TP-53 function is essential for sustaining durable responses to BH3-mimetic drugs in leukemias. Blood **137**(20), 2721–2735 (May 2021). https://doi.org/10.1182/blood.2020010167, https://doi.org/10.1182/blood.2020010167
17. Van Loo, P., Nordgard, S.H., Lingjærde, O.C., Russnes, H.G., Rye, I.H., Sun, W., Weigman, V.J., Marynen, P., Zetterberg, A., Naume, B., Perou, C.M., Børresen-Dale, A.L., Kristensen, V.N.: Allele-specific copy number analysis of tumors. Proceedings of the National Academy of Sciences **107**(39), 16910–16915 (Sep 2010). https://doi.org/10.1073/pnas.1009843107, https://www.pnas.org/doi/10.1073/pnas.1009843107, publisher: Proceedings of the National Academy of Sciences
18. Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhang, C.Z., Wala, J., Mermel, C.H., Sougnez, C., Gabriel, S.B., Hernandez, B., Shen, H., Laird, P.W., Getz, G., Meyerson, M., Beroukhim, R.: Pan-cancer patterns of somatic copy number alteration. Nature Genetics **45**(10), 1134–1140 (Oct 2013). https://doi.org/10.1038/ng.2760, https://www.nature.com/articles/ng.2760, publisher: Nature Publishing Group