Bayesian inference of fitness landscapes via tree-structured branching processes

Xiang Ge Luo,[1,2] **Jack Kuipers**,[1,2] **Kevin Rupp**,[1,2] **Koichi Takahashi**[3,4] **and Niko Beerenwinkel**[1,2,*]

[1]Department of Biosystems Science and Engineering, ETH Zurich, Klingelbergstrasse 48, 4056, Basel, Switzerland, [2]SIB Swiss Institute of Bioinformatics, Switerland, [3]Department of Leukemia, The University of Texas MD Anderson Cancer Center, Houston, TX, USA and [4]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

*Corresponding author. niko.beerenwinkel@bsse.ethz.ch

Abstract

Motivation: The complex dynamics of cancer evolution, driven by mutation and selection, underlies the molecular heterogeneity observed in tumors. The evolutionary histories of tumors of different patients can be encoded as mutation trees and reconstructed in high resolution from single-cell sequencing data, offering crucial insights for studying fitness effects of and epistasis among mutations. Existing models, however, either fail to separate mutation and selection or neglect the evolutionary histories encoded by the tumor phylogenetic trees.

Results: We introduce FiTree, a tree-structured multi-type branching process model with epistatic fitness parameterization and a Bayesian inference scheme to learn fitness landscapes from single-cell tumor mutation trees. Through simulations, we demonstrate that FiTree outperforms state-of-the-art methods in inferring the fitness landscape underlying tumor evolution. Applying FiTree to a single-cell acute myeloid leukemia dataset, we identify epistatic fitness effects consistent with known biological findings and quantify uncertainty in predicting future mutational events, offering a unified framework for understanding tumor progression and potentially for guiding therapeutic strategies.

Availability and implementation: The `Python` package `FiTree` and the analysis workflows are available at `https://github.com/cbg-ethz/FiTree`.

Keywords: cancer evolution, fitness landscape, branching process, tumor mutation trees

1. Introduction

Cancer progression involves a somatic evolutionary process, where cells accumulate mutations that confer selective advantages, allowing them to evade immune surveillance and proliferate uncontrollably (Nowell, 1976; Hanahan and Weinberg, 2011; Greaves and Maley, 2012). The extensive genetic and phenotypic heterogeneity within and between tumors enables cancer cells to rapidly adapt to different microenvironments, promoting metastasis and drug resistance (McGranahan and Swanton, 2017; Dagogo-Jack and Shaw, 2018). To better understand the disease and develop effective treatment strategies, it is crucial to disentangle different evolutionary forces that shape the diverse tumor populations and improve evolutionary-aware predictions (Lässig et al., 2017; Watson et al., 2020; Salehi et al., 2021).

With the increasing quality and availability of single-cell DNA sequencing data for cancer research (Morita et al., 2020; Miles et al., 2020; Minussi et al., 2021; Zhang et al., 2023), we are now able to reconstruct the evolutionary histories of tumors at unprecedented resolution using specifically designed phylogenetic methods (e.g., Jahn et al., 2016; Zafar et al., 2019; Kuipers et al., 2020; Kang et al., 2022; Kozlov et al., 2022; Markowska et al., 2022; Sollier et al., 2023; Sashittal et al., 2023) (Fig. 1ab). Analyzing tumor phylogenies across patients has shown promises in identifying reproducible features of cancer evolution, including common mutation orders and partitions (Ciccolella et al., 2021; Jahn et al., 2021), conserved evolutionary trajectories (Caravagna et al., 2018; Khakabimamaghani et al., 2019; Christensen et al., 2020; Hodzic et al., 2020; Pellegrina and Vandin, 2022), and patterns of co-occurrence and exclusivity among mutations (Kuipers et al., 2021; Luo et al., 2023; Ivanovic and El-Kebir, 2023). These features facilitate predictions of tumor evolution with probabilistic measures and improve patient clustering for precise clinical diagnosis and prognosis (Baciu-Drăgan and Beerenwinkel, 2024). However, most existing approaches for modeling intra-tumor phylogenies focus exclusively on the tree topologies, overlooking the abundances of different subclonal populations. Given that the frequencies of genetic variants are crucial markers for clinical disease classification and treatment planning (Bastian, 2014; Lee et al., 2019; Döhner et al., 2022), integrating these additional information into the modeling framework is essential for uncovering the complex evolutionary dynamics of tumors and translating them into actionable insights for clinical applications.

When tree structures are not considered, population genetics models provide well-established tools to study how mutation,

selection, and drift interact to shape the genetic composition and evolutionary trajectories of tumor cell populations over time (Ewens, 2004; Durrett, 2008; Beerenwinkel et al., 2015). State-of-the-art methods, built upon the coalescent (Dinh et al., 2020; Werner et al., 2020), the Wright-Fisher process (Salehi et al., 2021), and branching processes (Williams et al., 2018; Dinh et al., 2020; Lang et al., 2020; Salichos et al., 2020; Lee and Bozic, 2022), have proven effective in inferring key parameters of cancer evolution (*e.g.* mutation rates and fitness advantages) from next-generation sequencing data. However, none of them can leverage the evolutionary trajectories encoded in intra-tumor phylogenetic trees to inform evolutionary parameter estimates. Some methods are tailored to bulk sequencing data (Williams et al., 2018; Dinh et al., 2020; Watson et al., 2020; Salichos et al., 2020), while others require time series measurements (Salehi et al., 2021; Lee and Bozic, 2022), which are not typically available from observational human data. Furthermore, these methods typically assume simple fitness landscapes (Wright, 1932), where fitness advantages are either uniform across mutations or independent of one another, neglecting the well-established effects of epistasis (Phillips, 2008; De Visser and Krug, 2014). SCIFIL (Skums et al., 2019) is the first approach that considers both evolutionary histories and tumor cell abundances by superimposing a replicator model onto a tumor mutation tree—an alternative representation of tumor phylogenies (Jahn et al., 2016). Yet, fitting a deterministic model to relative allele frequencies from a single patient inevitably leads to monotonically increasing fitness advantages toward the leaves of the tree. Therefore, SCIFIL is limited for inferring more complex fitness landscapes across patients. A more recent computational method infers the average fitness advantage of drive mutations by comparing simulated clone size vectors to the observed ones from multiple trees (Poon et al., 2025). Despite its improvements, this approach disregards potential epistatic interactions and is computationally expensive due to simulation-based inference.

In this paper, we present FiTree, a tree-structured branching process model for learning fitness landscapes from single-cell tumor mutation trees. For the first time, FiTree reconciles probabilistic graphical models of cancer progression with classical population genetics, providing a unified framework to analyze single-time-point intra-tumor data. In particular, we develop novel theory, extending the linear multitype branching processes (Athreya and Ney, 2004; Durrett and Moseley, 2010; Nicholson et al., 2023; Zhang and Bozic, 2024) to include a static initial wild-type population, tree structures, and a tumor-size-dependent censoring mechanism. We parametrize fitness landscapes with epistatic effects and leverage a Bayesian inference scheme to incorporate prior biological knowledge and quantify uncertainty in fitness estimates. Through simulations, we validate our branching process theory and demonstrate that FiTree outperforms state-of-the-art methods in recovering fitness landscapes (Skums et al., 2019; Watson et al., 2020; Salehi et al., 2021). Finally, we apply FiTree to a single-cell acute myeloid leukemia (AML) dataset (Morita et al., 2020). With the inferred posterior distribution of the fitness landscape, we provide uncertainty estimates of future events captured in longitudinal samples.

2. Materials and methods

FiTree is formulated as a probabilistic generative model for single-cell tumor mutation trees. We begin by outlining the model setup, followed by the derivation of tree probabilities. Lastly, we describe the Bayesian inference scheme for parameter estimation.

2.1. FiTree: model setup

A tumor mutation tree $\mathscr{T}$ is a rooted tree that represents the evolutionary history of a tumor (Jahn et al., 2016). Each node v in $\mathscr{T}$ contains a subset of mutations $\pi_v \subseteq [n] = \{1, 2, \ldots, n\}$. The cells attached to v form a subclone with genotype g_v, determined by mutations accumulated along the trajectory leading to v. The

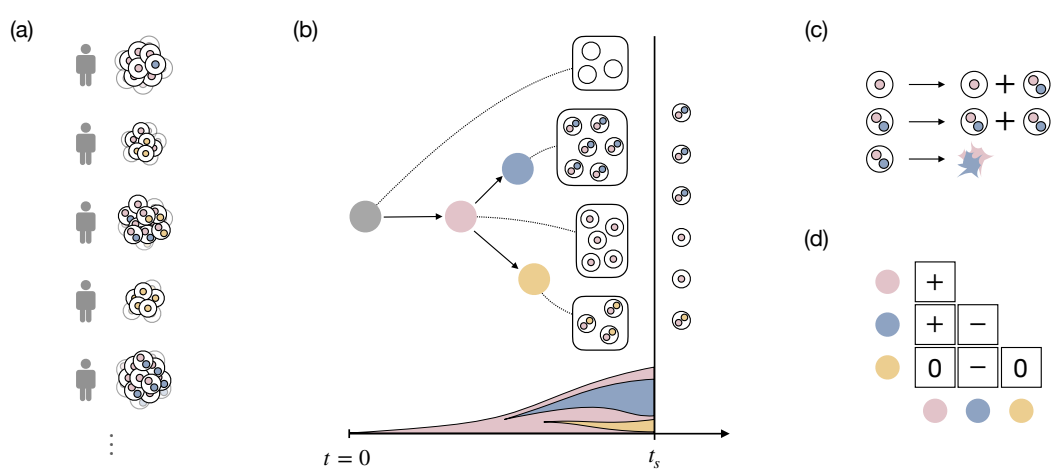

Fig. 1: Overview of FiTree. (a) Cross-sectional single-cell data. Genetically diverse tumor cells of different patients are collected at single time points. (b) Tumor evolutionary history represented as a tumor mutation tree and modeled using a probabilistic generative process in FiTree. The root node in gray corresponds to a static wild-type population at the start of evolution ($t = 0$). The colored nodes represent mutations accumulated over time. The sampling event occurs with rate proportional to the tumor size. At sampling time t_s, single cells are collected for sequencing. (c) Illustration of the growth dynamics in the branching process, where cells can mutate, divide, or die stochastically. (d) A fitness matrix where diagonal entries represent fitness effects of individual mutations and off-diagonal entries correspond to the epistatic effects.

root node v_0 corresponds to wild-type cells with no mutations. We denote the number of cells attached to v at time t by $C_v(t)$. To generate a tree $\mathscr{T}$, we consider a tree-structured multi-type branching process that starts with a static population of wild-type cells, $C_{v_0}(t) = C_0$ for all $t \geq 0$, and no mutant cells. Namely, the tree initially consists of only the root, representing stable populations such as hematopoietic stem cells in bone marrow (Catlin et al., 2011) or intestinal stem cells in colorectal crypts (Kozar et al., 2013). Then, a tree expansion rule R defines the set of subclones that could emerge next. One example of R is the Infinite Sites Assumption (ISA) (Kimura, 1969), which only allows mutations that have not yet occurred in any existing subclone to appear in the new subclones. Other rules relax ISA, permitting parallel, repeated, or back mutations (Kuipers et al., 2017). Note that while the input trees may follow tree expansion rules imposed by upstream phylogenetic methods or expert knowledge, our model is not restricted to any specific rules, thereby providing a flexible framework for different applications.

We define the growth dynamics for cells within a mutant subclone $v \neq v_0$ as follows (Fig. 1c):

$$\begin{aligned} C_{\mathrm{pa}(v)} &\xrightarrow{\nu_v} C_{\mathrm{pa}(v)} + C_v \\ C_v &\xrightarrow{\alpha_v} C_v + C_v \\ C_v &\xrightarrow{\beta_v} \emptyset \end{aligned} \tag{1}$$

Assuming each cell grows independently of each other, cells of subclone v emerge from the parent subclone $\mathrm{pa}(v)$ through mutations at rate ν_v, duplicate at rate α_v, and die off at rate β_v. We introduce a sampling event S, which occurs at a rate proportional to the total number of mutant cells $C_{\mathrm{tumor}} = \sum_{v \in \mathscr{T}, v \neq v_0} C_v$,

$$\emptyset \xrightarrow{C_{\mathrm{tumor}}/C_{\mathrm{sampling}}} S \tag{2}$$

where C_{sampling} is a scaling factor representing the tumor size at which sampling is likely to occur. For instance, setting $C_{\mathrm{sampling}} = 10^{11}$ implies that, on average, a tumor is sampled when it reaches around 10^{11} cells. The process terminates either at the first arrival of S with sampling time t_s, or upon reaching a pre-specified maximum observation time $t_{\max}$. The underlying principle is that larger tumors are more likely to be sampled, as patients are more prone to exhibit phenotypic symptoms and seek medical attention. Some patients, however, may develop cancer late in life and never get diagnosed, *i.e.* $t_s > t_{\max}$. In such cases, their tumor mutation trees are unobservable, introducing selection bias that must be accounted for during inference. Finally, to mimic the sequencing step, we sample C_{seq} cells without replacement from the tumor cell population. Typically the number of sequenced cells C_{seq} is much smaller than the total number of tumor cells C_{tumor} and often varies by patient. Leaf subclones with zero sequenced cells are considered unobserved and are therefore removed, resulting in the observed tumor mutation tree $\mathscr{T}$.

Assuming that the rates of individual mutations, denoted by $\boldsymbol{\mu} = (\mu_i)_{i=1}^n$, are independent and stay constant throughout the evolution, the subclone mutation rate is defined as $\nu_v = \prod_{i \in \pi_v} \mu_i$, a product of the rates of the new mutations in v. For identifiability we assign a common death rate β to all cell types and allow the birth rates to vary. To capture both the fitness effects of individual mutations and potential epistasis, we parameterize the log birth-to-death ratio using a triangular fitness matrix $F = (f_{ij})_{i,j \in [n]} \in \mathbb{R}^{n \times n}$ (Fig. 1d),

$$\varphi_v := \log(\alpha_v) - \log(\beta) = \max_{i \in g_v} f_{ii} + \sum_{i \neq j \in g_v} f_{ij}. \tag{3}$$

Here, the diagonal elements f_{ii} represent the fitness contributions of individual mutations, and the off-diagonal elements f_{ij} account for the pairwise interactions between mutations. This formulation assumes that only the most advantageous mutation dominates the fitness of a subclone, while all epistatic effects are captured in the interaction terms. The net growth rate is defined as

$$\lambda_v := \alpha_v - \beta = \beta \underbrace{(\exp(\varphi_v) - 1)}_{:= s_v}, \tag{4}$$

with its sign indicating whether the mutations are collectively advantageous ($\lambda_v > 0$), neutral ($\lambda_v = 0$), or deleterious ($\lambda_v < 0$). By definition the wild-type cells have a net growth rate of zero, *i.e.* $\alpha_{v_0} = \beta$ and $\lambda_{v_0} = 0$. Hence, the term s_v can be interpreted as the relative increase from the wild-type division rate.

2.2. Calculation of tree probabilities

To compute the probability of observing a tumor mutation tree $\mathscr{T}$ at sampling time t_s, we first need to understand the subclonal population and sampling time distributions. However, finding the exact solution of the above system is a highly intricate problem (Athreya and Ney, 2004; Antal and Krapivsky, 2011), even under the assumption that all data and parameters are perfectly known, which will be relaxed in the later sections. Therefore, we resort to approximations in the limit of large times and small mutation rates (Durrett and Moseley, 2010; Nicholson et al., 2023), which is often a reasonable assumption in cancer studies. In particular, we generalize the results in (Nicholson et al., 2023), which considers a linear multi-type branching process initiating from a single cell that grows exponentially, to model tree structures with static initial wild-type populations followed by cell types of arbitrary fitness. Following (Nicholson et al., 2023), we show in Theorem 1 (Supplementary Section B.1) that for any mutant type $v \neq v_0$ in $\mathscr{T}$, the limiting distribution of the subclonal population $C_v(t)$ can be decomposed approximately into a time-independent random variable $\tilde{C}_v$ and a time-dependent deterministic function,

$$C_v(t) \approx \tilde{C}_v t^{r_v - 1} \exp(\delta_v t), \tag{5}$$

where r_v is the number of times the running-max net growth rate

$$\delta_v = \max\{\lambda'_{v'} : v' \in \mathrm{ancestor}(v) \cup \{v\}\} \tag{6}$$

is attained along the lineage of v. Intuitively, given the subclonal population sizes

$$\mathbf{C}_{\mathscr{T}}(t) = \{C_v(t)\}_{v \in \mathscr{T}, v \neq v_0} \tag{7}$$

at some large time t, this result enables us to approximate deterministically the unobserved evolutionary trajectories $\mathbf{C}_{\mathscr{T}}(s)$ for all $0 < s < t$, thereby significantly simplifying the calculation of probabilities.

At any given time t, the specific size distribution of a subclone v is determined by both its position in the tree and its net growth rate λ_v. If v comes directly after the root, *i.e.* $\mathrm{pa}(v) = v_0$, then its

size $C_v(t)$ follows a negative binomial distribution, which in the large time limit is decomposable as in Eq. (5),

$$t^{-(r_v-1)}\exp(-\delta_v t)C_v(t) \approx \tilde{C}_v \sim \begin{cases} \text{Negative-Binomial}\left(C_0\rho_v, \phi_v^{-1}\right) & \text{if } \lambda_v < 0 \\ \text{Gamma}\left(C_0\rho_v, \phi_v^{-1}\right) & \text{if } \lambda_v \geq 0 \end{cases} \tag{8}$$

where

$$\rho_v = \frac{\nu_v}{\alpha_v} \quad \text{and} \quad \phi_v = \begin{cases} -\beta_v/\lambda_v & \text{if } \lambda_v < 0 \\ \alpha_v & \text{if } \lambda_v = 0 \\ \alpha_v/\lambda_v & \text{if } \lambda_v > 0 \end{cases} \tag{9}$$

are the shape and scale, respectively, of the distribution (Lemma 1 & Theorem 1, Supplementary Section B.1).

For all remaining subclones v with $\text{pa}(v) \neq v_0$, the population distribution has no simple form. It is characterized by the Laplace transform conditioned on the parent subclone size,

$$\mathbb{E}\left[\exp(-\theta\tilde{C}_v) \mid \tilde{C}_{\text{pa}(v)} = \tilde{c}_{\text{pa}(v)}\right] = \exp\left(-h_v(\theta)\tilde{c}_{\text{pa}(v)}\right), \tag{10}$$

where $h_v(\theta)$ is a piece-wise function that distinguishes whether v is fitter ($\lambda_v > \delta_{\text{pa}(v)}$), equally fit ($\lambda_v = \delta_{\text{pa}(v)}$), or less fit ($\lambda_v < \delta_{\text{pa}(v)}$) than all its ancestor subclones (Theorem 2, Supplementary Section B.1). If v is not fitter than its parent, then on average, its size will remain proportional to that of its parent at large times. Otherwise, it expands at its own net growth rate. In other words, the growth dynamics of subclone v are dominated by the fittest subclone in its lineage. Thus, deleterious mutations can still persist in the cohort by hitchhiking alongside beneficial mutations already present in the subclone.

The sampling time T_s is the waiting time until the first arrival of the sampling event $S(t)$. By definition it follows a nonhomogeneous Poisson process with intensity function depending on the tumor size,

$$\int_0^t \frac{C_{\text{tumor}}(s)}{C_{\text{sampling}}}ds, \qquad C_{\text{tumor}}(s) = \sum_{v\in\mathscr{T}, v\neq v_0} C_v(s). \tag{11}$$

Hence, as proved in Theorem 3 (Supplementary Section B.2), the distribution function of T_s is given by

$$P(T_s > t) = P(S(t) = 0) = \mathop{\mathbb{E}}_{\mathbf{C}_{\mathscr{T}}(t)}\left[\exp\left(-\frac{1}{C_{\text{sampling}}}\sum_{v\in\mathscr{T}, v\neq v_0}\int_0^t C_v(s)ds\right)\right]. \tag{12}$$

In addition, the vector of sequenced cells, defined as

$$\mathbf{C}'_{\mathscr{T}}(t_s) = \{C'_v(t_s)\}_{v\in\mathscr{T}, v\neq v_0}, \tag{13}$$

are sampled without replacement from the subclonal population size vector (Eq. (7)), thereby following a multivariate hypergeometric distribution with density

$$\frac{\prod_{v\in\mathscr{T}, v\neq v_0}\binom{C_v(t_s)}{C'_v(t_s)}}{\binom{C_{\text{tumor}}(t_s)}{C_{\text{seq}}}}. \tag{14}$$

Putting everything together, the probability of observing a tumor mutation tree $\mathscr{T}$ at sampling time t_s is given by

$$\begin{aligned} p(\mathscr{T}, t_s) = & \underbrace{\prod_{v\in\mathscr{T}, \text{pa}(v)=v_0} p(c_v(t_s))}_{\text{subclones following the root}} \\ & \times \underbrace{\prod_{v\in\mathscr{T}, \text{pa}(v)\neq v_0} p\left(c_v(t_s) \mid c_{\text{pa}(v)}(t_s)\right)}_{\text{subclones further downstream}} \\ & \times \underbrace{p\left(t_s \mid \mathbf{c}_{\mathscr{T}}(t_s)\right)}_{\text{sampling time}} \times \underbrace{p\left(\mathbf{c}'_{\mathscr{T}}(t_s) \mid \mathbf{c}_{\mathscr{T}}(t_s)\right)}_{\text{sequenced cells}}. \end{aligned} \tag{15}$$

We refer to this probabilistic model of mutation trees as FiTree. Key notations are summarized in Supplementary Table S1.

2.3. Bayesian inference of fitness landscapes

The parameters of FiTree (Eq. 15) are

$$\Theta = \{C_0, C_{\text{sampling}}, C_{\text{seq}}, t_{\max}, \boldsymbol{\mu}, \beta, F\}. \tag{16}$$

In this paper, we focus on inferring the marginal posterior distribution of the fitness matrix F given a set of N tumor mutation trees $\boldsymbol{\mathscr{T}} = \{\mathscr{T}^j\}_{j\in N}$ with sampling times $\boldsymbol{t}_s = (t_s^j)_{j\in N}$. To this end, we adopt the following assumptions about the data and the parameters:

1. While the actual subclonal population sizes (Eq. (7)) are unobservable, the total tumor size C_{tumor} can be estimated from clinical measurements.
2. The sample proportions accurately reflect the true cancer cell fractions in patients,
$$\frac{C'_v(t_s)}{C_{\text{seq}}} \approx \frac{C_v(t_s)}{C_{\text{tumor}}(t_s)}, \qquad \forall v \in \mathscr{T}, v \neq v_0. \tag{17}$$
3. All parameters, except F, can be derived from either clinical data or literature (see Supplementary Section E.1 for an example).

Note that the sampling times $\boldsymbol{t}_s$, if unavailable, can still be readily integrated out, which is not a limiting factor for broad applications. The posterior of F is given by

$$p(F \mid \boldsymbol{\mathscr{T}}, \boldsymbol{t}_s, \Theta \setminus \{F\}) \propto p(\boldsymbol{\mathscr{T}}, \boldsymbol{t}_s \mid \Theta)p(F). \tag{18}$$

Here, the log-likelihood $p(\boldsymbol{\mathscr{T}}, \boldsymbol{t}_s \mid \Theta)$ is computed as

$$\sum_{j=1}^N \log p(\mathscr{T}^j, t_s^j \mid \Theta) - N\log P(T_s \leq t_{\max} \mid \Theta), \tag{19}$$

where the first term sums the log-likelihoods of all trees (Eq. (15)), and the second term adjusts for selection bias, as datasets can only contain observable trees and no negative controls. When the lifetime risk of developing and being diagnosed with the disease,

$$\zeta := P(T_s \leq t_{\max}), \tag{20}$$

can be reliably estimated from external databases (e.g., Surveillance, Epidemiology, and End Results (SEER) Program),

we can additionally impose a prior on $P(T_s \leq t_{\max} \mid \Theta)$, centered at ζ with a small width, to enforce consistency. Moreover, we assign the normal prior $\mathcal{N}(0, \sigma^2)$ to every distinct entry in the matrix. To explore the parameter space, we use Markov chain Monte Carlo sampling. Further details are provided in Supplementary Section C.

3. Results

3.1. Simulations

Our simulation studies consist of two parts: (1) validating the large-time approximations for subclonal population and sampling time distributions, and (2) benchmarking the performance of FiTree in learning the subclone fitness effects φ_v (Eq. (3)) from simulated tumor mutation trees against state-of-the-art methods.

3.1.1. Validating approximated distribution functions

We begin with verifying the large-time approximations of the subclonal population and sampling time distributions, which are the building blocks of the likelihood function (Eq. (15)). Specifically, we perform stochastic simulations of branching processes with three subclones and a sampling event using the explicit tau-leaping method (Gillespie, 2001). The simulations encompass seven different parameter settings, as summarized in Supplementary Table S2 and illustrated in Fig. 2a. While most parameters follow those in (Nicholson et al., 2023), covering a range of net growth rates and mutation rates, we extend the analysis to include neutral and deleterious cases for the first cell type, as well as linear versus branching structures. Further simulation details are provided in Supplementary Section D.1.

Across all scenarios (Fig. 2b), the empirical marginal cumulative distribution functions for each cell type and sampling time closely match their analytical approximations (Corollary 1 and Theorem 3, Supplementary Section B). To our knowledge, this is the first formulation to integrate a static wild-type population, tree structures, and a size-dependent stopping mechanism into the multitype branching process framework, extending the foundations laid by Athreya and Ney (2004), Durrett and Moseley (2010), and Nicholson et al. (2023). Accurate approximations of the distribution functions ensure reliable fitness landscape inference in downstream analysis.

3.1.2. Benchmarking FiTree performance

The second part of the simulations benchmarks FiTree's performance in learning fitness landscapes from tumor mutation trees against state-of-the-art methods, including SCIFIL (Skums et al., 2019), fitclone (Salehi et al., 2021), and the diffusion approximation of a one-type branching process (Watson et al., 2020). The pseudocode for tree generation, along with details on ground-truth parameter selection and method execution, are provided in Supplementary Section D.

Since fitness definitions differ across models, direct comparisons of absolute fitness effects are generally not possible, except for the method in (Watson et al., 2020). For this method, we separately compare the deviation between the subclone fitness estimates φ'_v and the ground truth φ_v using the following metrics:

$$\text{mean absolute error} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{|\mathscr{T}^j|}\sum_{v\in\mathscr{T}^j}|\varphi'_v - \varphi_v|, \tag{21}$$

$$\text{sign agreement} = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{|\mathscr{T}^j|}\sum_{v\in\mathscr{T}^j}\mathbb{1}\left(\text{sign}(\varphi'_v) = \text{sign}(\varphi_v)\right). \tag{22}$$

For all methods, we evaluate Spearman correlations between ground-truth and estimated fitnesses to measure each model's ability to capture relative variations in fitness landscapes. Here, we include a simple baseline that ranks subclone fitnesses based on their average frequencies within the cohort. Moreover, we evaluate the generalizability of all methods on the unique genotypes of both observed and unobserved subclones, with the latter potentially undetected due to late emergence or technological limits. We focus on unobserved subclones with up to two mutations, as including more provides little additional insight given that none of the methods capture higher-order epistasis.

For varying number of mutations ($n \in \{5, 10, 15\}$) and number of samples ($N \in \{200, 500\}$), FiTree consistently outperforms all alternative methods in learning the fitness values of both observed and unobserved subclones (Fig. 3). With additional experiments, we show that removing less frequently mutated genes from inference does not create strong bias on the estimated fitness landscape (Supplementary Figure S12) and that imposing a strong informative prior on the lifetime risk (Eq. (20)) mitigates overestimation of fitness effects (Supplementary Figure S13).

The diffusion approximation approach in (Watson et al., 2020) performs reasonably well, as it is also based on a branching process model, albeit a much simpler one, and aggregates information across patients. However, being designed for bulk sequencing data, it cannot fully leverage the evolutionary trajectories and subclone sizes provided by single-cell tumor mutation trees. While its simplicity offers some stability as the problem dimensions increase (Fig. 3a), this comes at the cost of ignoring interactions between mutations, leaving little room for improvement with additional data (Fig. 3bc). In contrast, FiTree captures these interactions and provides a more comprehensive model of the evolutionary process, resulting in better performance despite requiring more data and computational effort, particularly in higher-dimensional settings.

SCIFIL and fitclone, both designed for single-patient analyses, assign fitness values to subclones based primarily on their sizes within individual trees. Under these models, a subclone may have high fitness if it is large in one tree but low fitness if it is small or absent in another. While one could argue that differences in tumor microenvironments across patients might lead to varying fitness for the same subclone, this does not fully explain the common patterns observed across trees in real data. A subclone may appear small in one tree not due to lower fitness but because of later emergence, reflecting the stochastic nature of the evolutionary process, which deterministic models like SCIFIL cannot capture. While SCIFIL performs well for observed subclones, likely due to its consideration of tree structures, its performance is similar to the baseline method when unobserved subclones are included (Fig. 3c).

Even as a stochastic method, fitclone struggles when it is provided with only two time points: a trivial initial point at $t = 0$, with no mutations present, and a second point at diagnosis, reflecting the limitations of many real-world datasets. The poor results suggest that fitclone requires more time points to be effective and cannot reliably handle such sparse data. As a Bayesian approach, fitclone consumes significant computational resources, and given its poor performance for $n = 5$, we did not evaluate it for larger scenarios.

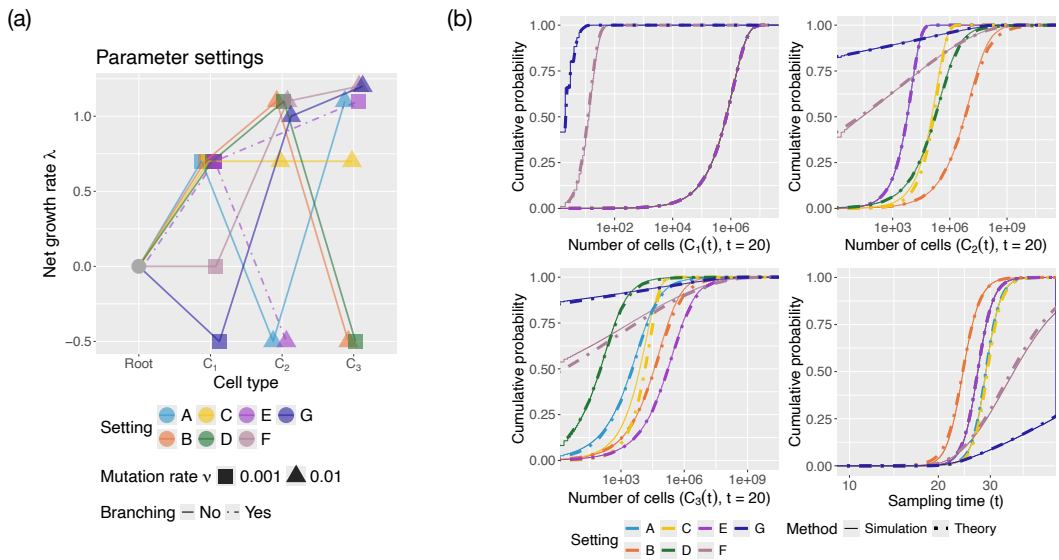

Fig. 2: Accurate approximations of subclonal population and sampling time distributions in tree-structured branching processes. (a) Illustration of seven different parameter settings in stochastic simulations. The x-axis is the cell type, and the y-axis is the net growth rate. Except for setting E, where a branching tree structure is used (dash lines), all other settings use the linear structure (solid lines). We use different shapes to represent the varying mutation rates (rectangle: 0.001; triangle: 0.01). The root has a net growth rate of 0 and does not have a mutation rate. (b) Comparison of the cumulative distribution functions between simulated values (solid lines) and theoretical approximations (dashed lines). Note that some curves are overlapping, *e.g.* curves for setting A through E for $C_1(t)$, since the displayed results are faceted by variables and colored by parameter settings. The same results faceted by parameter settings and colored by variables are provided in Supplementary Figures S2-S8.

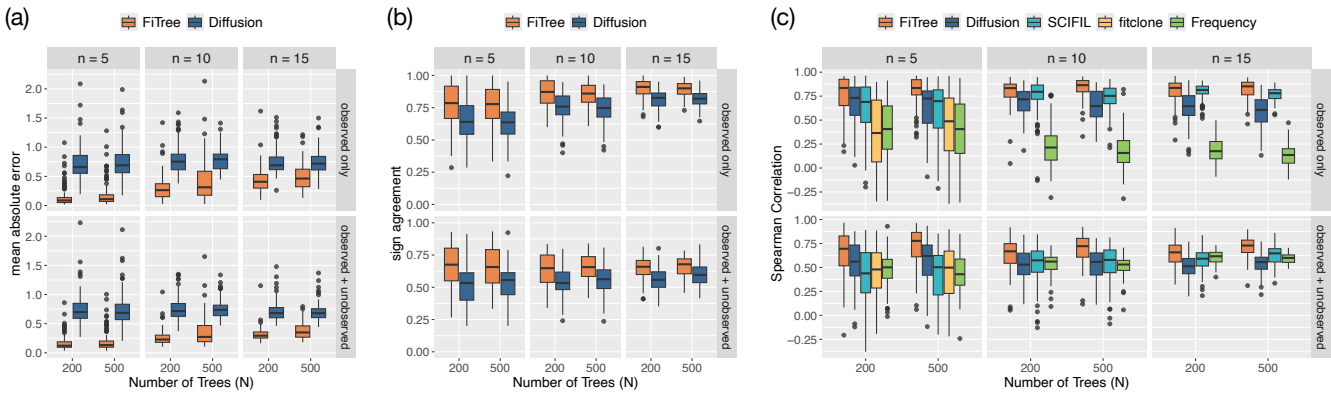

Fig. 3: Performance comparison of FiTree with state-of-the-art methods in learning fitness landscapes on simulated data. Boxplots display the mean absolute errors (a), sign agreement (b), and Spearman correlations (c) across 100 simulation runs, grouped by $n \in \{5, 10, 15\}$, $N \in \{200, 500\}$, and evaluation scope (observed subclones only or including unobserved subclones). For FiTree and fitclone, each point is derived from the posterior medians, while for other methods, points correspond to maximum likelihood estimates. Boxes represent the interquartile range (IQR) with medians marked inside, whiskers extend to 1.5 times the IQR, and outliers are shown as individual points.

3.2. Application to a single-cell AML dataset

We apply FiTree to a single-cell AML dataset (Morita et al., 2020) with $N = 123$ tumor mutation trees reconstructed using SCITE (Jahn et al., 2016), which contains 530 somatic mutations across $n = 31$ genes. To address the challenge of high dimensionality with a limited number of samples and enhance the statistical power to detect non-random patterns, we summarize the variants at the gene level and focus on mutations in genotypes present in at least three patients. This filtering reduces the analysis to $n = 15$ genes, which were also identified by TreeMHN (Luo et al., 2023) as having statistically stable exclusivity patterns. Importantly, less frequently mutated genes are not removed from the trees but are excluded from the inference of epistatic fitness effects, with their corresponding off-diagonal entries in the fitness matrix F assigned zeros. As validated in simulations, this approach does not introduce significant bias in the estimates. We provide details on data preparation, including tumor size estimation, and static parameter estimation, such as mutation rates and other model parameters in Eq. (16), Supplementary Section E.

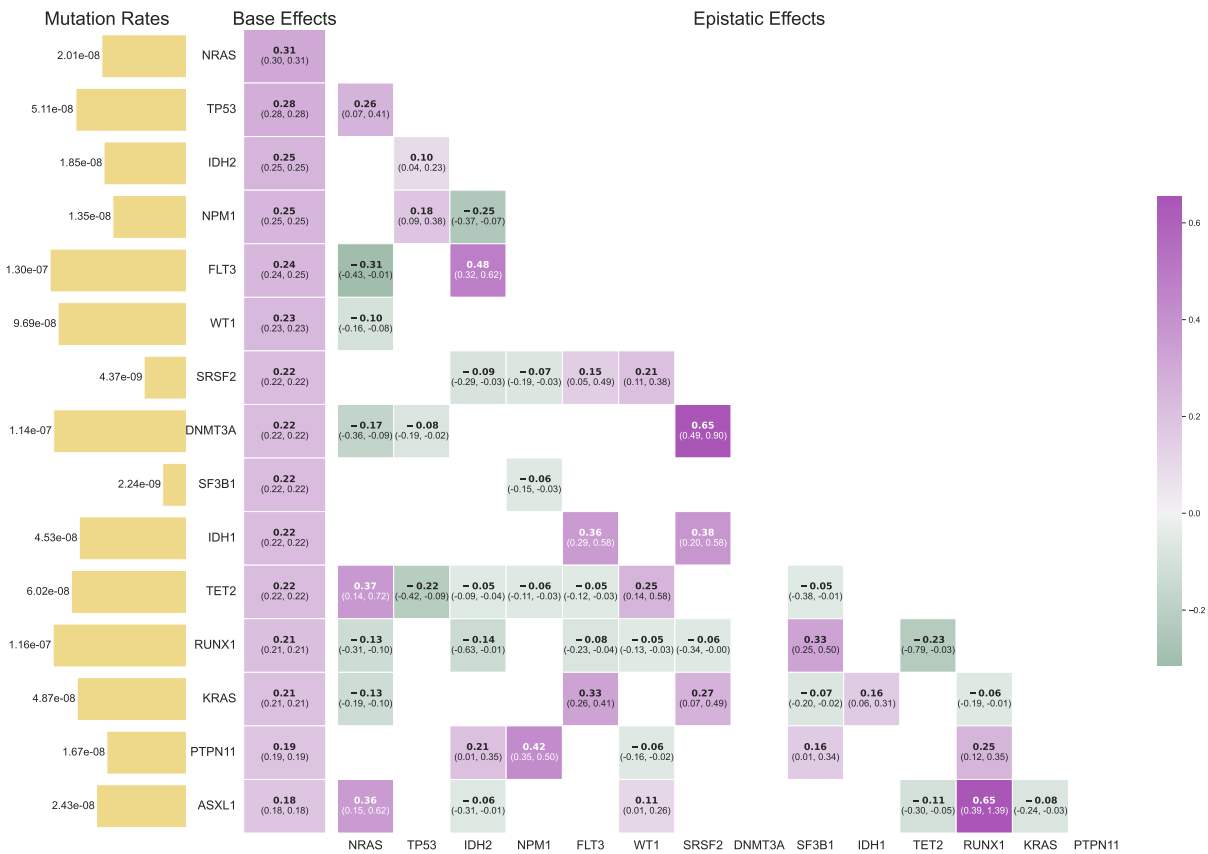

Fig. 4: Mutation rates and fitness matrix for the AML dataset. The mutation rates are displayed as bar plots on the left (yellow), with bar lengths scaled logarithmically according to the labeled mutation rates. The inferred fitness matrix is shown as a heatmap, where base effects are separately arranged in a column on the left. Matrix rows and columns are ordered by decreasing base effects. Each cell displays the posterior median in bold, with brackets indicating the corresponding 95% highest density interval (HDI). Cells with absolute epistatic effects lower than 0.05 or HDIs covering zero are masked. Full fitness matrix along with standard deviations are provided in Supplementary Figure S15 and S16.

The mutation rates and the posterior distribution of the fitness matrix inferred by FiTree are shown in Fig. 4. The results reveal a complex interplay of positive and negative epistatic effects, suggesting that assuming uniform (Poon et al., 2025) or independent fitness effects for individual mutations (Watson et al., 2020) is insufficient to capture the underlying evolutionary dynamics.

All base fitness effects of the frequently mutated genes are positive, with increases relative to the wild-type division rate ranging from 20% to 36% (s_v in Eq. (4)), confirming their role as driver mutations that confer a significant growth advantage to tumor cells. *DNMT3A* and *FLT3*, two commonly mutated genes in AML (Miles et al., 2020; Morita et al., 2020), have similar mutation rates, yet *FLT3* exhibits higher base fitness. One plausible explanation is that mutations in epigenetic modifiers (*e.g. DNMT3A*) are likely initiating events (Dai et al., 2017), whereas those in signaling genes (*e.g. FLT3*) drive rapid tumor growth, resulting in diagnosis with fewer subsequent mutations (Miles et al., 2020). In contrast, TreeMHN, which does not account for subclone sizes, estimates the highest fixation rate for *DNMT3A*. This does not contradict our findings; rather, it suggests that *DNMT3A* mutations are expected to appear more frequently in the cohort but do not necessarily result in larger subclones. By modeling mutation and selection separately while incorporating size information, FiTree provides more biologically meaningful insights.

With our Bayesian inference framework, we observe that the 95% highest density intervals for base effects are narrower than those for epistatic effects, indicating greater uncertainty in estimating epistasis due to the limited sample size. We observe large positive interactions with less uncertainty for several pairs of genes such as *NPM1-PTPN11*, *RUNX1-ASXL1*, and *FLT3-IDH1/2*. They correspond to known co-occurrence, and some are associated with worsened clinical outcomes (DiNardo et al., 2015; Falini et al., 2020; Gaidzik et al., 2016). Additionally, FiTree identifies a negative epistatic effect between *FLT3* and *NRAS*, both of which exhibit high base fitness. While each mutation provides a strong selective advantage individually, their combined presence may result in reduced fitness, making them less likely to co-occur within the same genome. This finding aligns with the mutual exclusivity observed in previous studies (Kuipers et al., 2021; Luo et al., 2023) and offers an explicit evolutionary explanation for the reported secondary resistance to *FLT3* inhibitors, which may arise from off-target RAS pathway mutations (*e.g. NRAS*) present in small populations before treatment (McMahon et al., 2019).

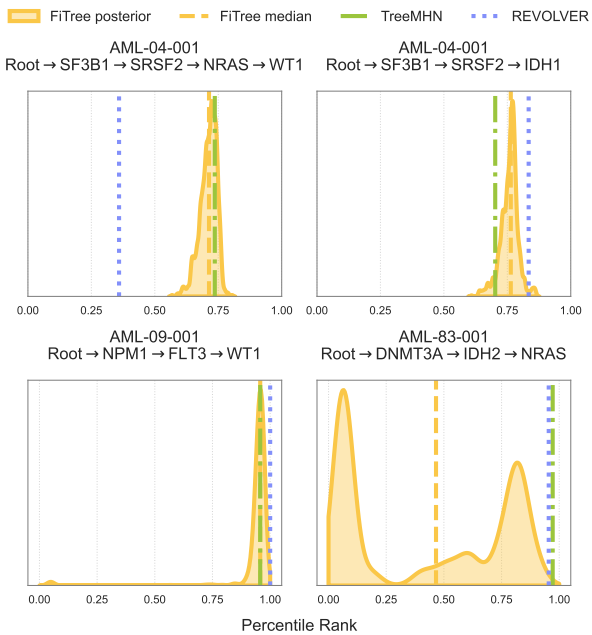

Fig. 5: Estimated percentile ranks of the four new events in the first post-treatment samples for patients AML-04-001, AML-09-001, and AML-83-001. Each subplot corresponds to one event. The density curves are estimated based on 1000 samples drawn from the FiTree posterior distribution. The lines represent the posterior median of FiTree (yellow, dashed), as well as the maximum likelihood estimates by TreeMHN (green, dash-dotted), and REVOLVER (blue, dotted).

By sampling from the posterior distribution of the fitness matrix, we assess FiTree's ability to predict the next likely mutational events given a primary tumor tree. We validate our evolutionary forecasts using longitudinal samples from the same cohort, which include clinical data for three patients with four new mutational events (*i.e.* subclones with new genotypes) in the first post-treatment samples (Morita et al., 2020). The predicted likelihood of an event is computed as the instantaneous growth rate, defined as the product of the parent subclone size at sampling, the mutation rate, and the estimated running-max growth rate (Eq. (6)). Following Luo et al. (2023), we rank the instantaneous growth rates among all possible future events for a given tree, where a higher percentile rank indicates a greater chance of occurrence. FiTree assigns higher rankings to three out of four events with small credible intervals (Fig. 5), consistent with previous estimates by TreeMHN and REVOLVER (Caravagna et al., 2018). For the event Root→*DNMT3A*→*IDH2*→*NRAS* in patient AML-83-001, the posterior predictive distribution is bimodal and spans the entire support, indicating substantial uncertainty and potential non-identifiability due to limited data. Unlike point estimation methods, FiTree offers a probabilistic view that reveals the inherent complexity of the evolutionary process.

4. Discussion

We have introduced FiTree, a novel tree-structured branching process model with a Bayesian inference scheme, to infer complex fitness landscapes from single-cell tumor mutation trees. Through simulations, FiTree demonstrated improved performance in recovering fitness landscapes compared to state-of-the-art methods. When applied to a single-cell AML dataset, FiTree identified epistatic fitness effects associated with known biological findings and quantified the uncertainty associated with future mutational events in longitudinal samples, offering valuable insights into tumor progression and enabling evolutionary forecasting. More accurate predictions will require additional data and explicit modeling of treatment effects, which can now be achieved with FiTree's design. Specifically, treatment-induced changes can be incorporated into the fitness parameterization, enabling relapse simulation through forward sampling. Such analyses are not feasible with previous methods, which lack a growth model and an explicit notion of fitness.

FiTree is particularly well-suited for analyzing AML, as branching processes are inherently designed for homogeneously mixed populations with non-spatial growth (Durrett, 2015). While incorporating spatial constraints is important for capturing different modes of tumor evolution (Noble et al., 2022), especially in solid tumors, it remains a technical challenge in the context of single-cell tumor mutation trees. One should be mindful of potential model misspecification when applying FiTree to different cancer types.

The current fitness parameterization considers only pairwise interactions and ignores clonal cooperation (Tabassum and Polyak, 2015). While constraining the parameter space is desirable when dealing with limited data, capturing higher-order epistasis and nonlinear interactions could provide a more accurate representation of tumor evolution and improve predictive power as more data becomes available.

Moreover, FiTree could be further improved by incorporating patient-specific factors, such as age, sex, and pre-existing conditions (*e.g.* smoking history), which are known to significantly impact tumor evolution. Integrating these factors would result in a more comprehensive model, where personalized fitness estimates could support downstream analyses such as patient stratification, survival prediction, and treatment planning (Bayer et al., 2023).

In summary, FiTree is the first approach to bridge probabilistic graphical models of cancer progression with classical population genetics for analyzing single-cell data. It opens up opportunities for more precise modeling of tumor progression and drug response, enabling improved treatment strategies and personalized interventions.

5. Competing interests

No competing interest is declared.

6. Author contributions statement

Conceptualization: XL, JK, KR, NB; Data Curation: XL, KT; Formal Analysis: XL; Methodology: XL, JK, KR, NB; Software: XL; Supervision: JK, NB; Visualization: XL; Writing – Original Draft Preparation: XL; Writing – Review & Editing: XL, JK, KR, KT, NB.

7. Acknowledgments

This work was supported by ETH Zurich [Open ETH project SKINTEGRITY.CH (N.B.)] and the European Union's Horizon 2020 research and innovation program [grant number 951970, OLISSIPO project (N.B.)].

References

T. Antal and P. Krapivsky. Exact solution of a two-type branching process: models of tumor progression. *Journal of Statistical Mechanics: Theory and Experiment*, 2011:P08018, 2011.

K. B. Athreya and P. E. Ney. *Branching processes*. Courier Corporation, 2004.

M.-A. Baciu-Drăgan and N. Beerenwinkel. Oncotree2vec—a method for embedding and clustering of tumor mutation trees. *Bioinformatics*, 40:i180–i188, 2024.

B. C. Bastian. The molecular pathology of melanoma: an integrated taxonomy of melanocytic neoplasia. *Annual Review of Pathology: Mechanisms of Disease*, 9:239–271, 2014.

F. Bayer et al. Network-based clustering unveils interconnected landscapes of genomic and clinical features across myeloid malignancies. *bioRxiv*, pages 2023–10, 2023.

N. Beerenwinkel et al. Cancer evolution: mathematical models and computational inference. *Systematic Biology*, 64:e1–e25, 2015.

G. Caravagna et al. Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nature Methods*, 15:707–714, 2018.

S. N. Catlin et al. The replication rate of human hematopoietic stem cells in vivo. *Blood, The Journal of the American Society of Hematology*, 117:4460–4466, 2011.

S. Christensen et al. Detecting evolutionary patterns of cancers using consensus trees. *Bioinformatics*, 36:i684–i691, 2020.

S. Ciccolella et al. Triplet-based similarity score for fully multilabeled trees with poly-occurring labels. *Bioinformatics*, 37:178–184, 2021.

I. Dagogo-Jack and A. T. Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15:81–94, 2018.

Y.-J. Dai et al. Conditional knockin of DNMT3A R878H initiates acute myeloid leukemia with mTOR pathway involvement. *Proceedings of the National Academy of Sciences*, 114(20): 5237–5242, 2017.

J. A. G. De Visser and J. Krug. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, 15: 480–490, 2014.

C. D. DiNardo et al. Characteristics, clinical outcome, and prognostic significance of IDH mutations in AML. *American Journal of Hematology*, 90:732–736, 2015.

K. N. Dinh et al. Statistical inference for the evolutionary history of cancer genomes. *Statistical Science*, 35:129–144, 2020.

H. Döhner et al. Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. *Blood, The Journal of the American Society of Hematology*, 140:1345–1377, 2022.

R. Durrett. *Probability models for DNA sequence evolution*, volume 2. Springer, 2008.

R. Durrett. *Branching process models of cancer*. Springer, 2015.

R. Durrett and S. Moseley. Evolution of resistance and progression to disease during clonal expansion of cancer. *Theoretical Population Biology*, 77:42–48, 2010.

W. J. Ewens. *Mathematical population genetics: theoretical introduction*, volume 27. Springer, 2004.

B. Falini et al. NPM1-mutated acute myeloid leukemia: from bench to bedside. *Blood, The Journal of the American Society of Hematology*, 136:1707–1721, 2020.

V. Gaidzik et al. RUNX1 mutations in acute myeloid leukemia are associated with distinct clinico-pathologic and genetic features. *Leukemia*, 30:2160–2168, 2016.

D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115:1716–1733, 2001.

M. Greaves and C. C. Maley. Clonal evolution in cancer. *Nature*, 481:306–313, 2012.

D. Hanahan and R. A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144:646–674, 2011.

E. Hodzic et al. Identification of conserved evolutionary trajectories in tumors. *Bioinformatics*, 36:i427–i435, 2020.

S. Ivanovic and M. El-Kebir. Modeling and predicting cancer clonal evolution with reinforcement learning. *Genome Research*, 33:1078–1088, 2023.

K. Jahn, J. Kuipers, and N. Beerenwinkel. Tree inference for single-cell data. *Genome biology*, 17:86, 2016.

K. Jahn, N. Beerenwinkel, and L. Zhang. The bourque distances for mutation trees of cancers. *Algorithms for Molecular Biology*, 16:9, 2021.

S. Kang et al. Sieve: joint inference of single-nucleotide variants and cell phylogeny from single-cell dna sequencing data. *Genome Biology*, 23:248, 2022.

S. Khakabimamaghani et al. Collaborative intra-tumor heterogeneity detection. *Bioinformatics*, 35:i379–i388, 2019.

M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61:893, 1969.

S. Kozar et al. Continuous clonal labeling reveals small numbers of functional stem cells in intestinal crypts and adenomas. *Cell Stem Cell*, 13:626–633, 2013.

A. Kozlov et al. CellPhy: accurate and fast probabilistic inference of single-cell phylogenies from scDNA-seq data. *Genome Biology*, 23:37, 2022.

J. Kuipers, K. Jahn, et al. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Research*, 27:1885–1894, 2017.

J. Kuipers et al. Single-cell copy number calling and event history reconstruction. *BioRxiv*, pages 2020–04, 2020.

J. Kuipers et al. Statistical tests for intra-tumour clonal co-occurrence and exclusivity. *PLoS Computational Biology*, 17: e1009036, 2021.

B. M. Lang, J. Kuipers, et al. Predicting colorectal cancer risk from adenoma detection via a two-type branching process model. *PLoS Computational Biology*, 16:e1007552, 2020.

M. Lässig, V. Mustonen, and A. M. Walczak. Predicting evolution. *Nature Ecology & Evolution*, 1:0077, 2017.

J.-M. Lee, L. Minasian, and E. C. Kohn. New strategies in ovarian cancer treatment. *Cancer*, 125:4623–4629, 2019.

N. D. Lee and I. Bozic. Inferring parameters of cancer evolution in chronic lymphocytic leukemia. *PLOS Computational Biology*, 18:e1010677, 2022.

X. G. Luo, J. Kuipers, and N. Beerenwinkel. Joint inference of exclusivity patterns and recurrent trajectories from tumor mutation trees. *Nature Communications*, 14:3676, 2023.

M. Markowska et al. CONET: copy number event tree model of evolutionary tumor history for single-cell data. *Genome Biology*, 23:128, 2022.

N. McGranahan and C. Swanton. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*, 168:613–628, 2017.

C. M. McMahon et al. Clonal selection with RAS pathway activation mediates secondary clinical resistance to selective FLT3 inhibition in acute myeloid leukemia. *Cancer Discovery*, 9:1050–1063, 2019.

L. A. Miles et al. Single-cell mutation analysis of clonal evolution in myeloid malignancies. *Nature*, 587:477–482, 2020.

D. C. Minussi et al. Breast tumours maintain a reservoir of subclonal diversity during expansion. *Nature*, 592:302–308, 2021.

K. Morita et al. Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nature Communications*, 11:5327, 2020.

M. D. Nicholson, D. Cheek, and T. Antal. Sequential mutations in exponentially growing populations. *PLoS Computational Biology*, 19:e1011289, 2023.

R. Noble et al. Spatial structure governs the mode of tumour evolution. *Nature Ecology & Evolution*, 6:207–217, 2022.

P. C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194:23–28, 1976.

L. Pellegrina and F. Vandin. Discovering significant evolutionary trajectories in cancer phylogenies. *Bioinformatics*, 38:ii49–ii55, 2022.

P. C. Phillips. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9:855–867, 2008.

G. Poon et al. Single-cell dna sequencing reveals pervasive positive selection throughout preleukemic evolution. *Cell Genomics*, 2025.

S. Salehi et al. Clonal fitness inferred from time-series modelling of single-cell cancer genomes. *Nature*, 595:585–590, 2021.

L. Salichos et al. Estimating growth patterns and driver effects in tumor evolution from individual samples. *Nature Communications*, 11:732, 2020.

P. Sashittal et al. ConDoR: tumor phylogeny inference with a copy-number constrained mutation loss model. *Genome Biology*, 24:272, 2023.

P. Skums, V. Tsyvina, and A. Zelikovsky. Inference of clonal selection in cancer populations using single-cell sequencing data. *Bioinformatics*, 35:i398–i407, 2019.

E. Sollier et al. COMPASS: joint copy number and mutation phylogeny reconstruction from amplicon single-cell sequencing data. *Nature Communications*, 14:4921, 2023.

Surveillance, Epidemiology, and End Results (SEER) Program. Cancer Stat Facts: Leukemia—Acute Myeloid Leukemia (AML). National Cancer Institute. Available at: https://seer.cancer.gov/statfacts/html/amyl.html. Accessed: 2024-10-21.

D. P. Tabassum and K. Polyak. Tumorigenesis: it takes a village. *Nature Reviews Cancer*, 15:473–483, 2015.

C. J. Watson et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science*, 367:1449–1454, 2020.

B. Werner et al. Measuring single cell divisions in human tissues from multi-region sequencing data. *Nature Communications*, 11:1035, 2020.

M. J. Williams et al. Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics*, 50:895–903, 2018.

S. Wright. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the Sixth International Congress of Genetics*, 1:356–366, 1932.

H. Zafar et al. SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome Research*, 29:1847–1859, 2019.

H. Zhang et al. Application of high-throughput single-nucleus dna sequencing in pancreatic cancer. *Nature Communications*, 14:749, 2023.

R. Zhang and I. Bozic. Accumulation of oncogenic mutations during progression from healthy tissue to cancer. *Bulletin of Mathematical Biology*, 86:142, 2024.