Joint inference of repeated evolutionary trajectories and patterns of clonal exclusivity or co-occurrence from tumor mutation trees

Xiang Ge Luo[1,2], Jack Kuipers[1,2], Niko Beerenwinkel[1,2*]
[1]Department of Biosystems Science and Engineering, ETH Zurich, Mattenstrasse 26, 4058 Basel, Switzerland
[2]SIB Swiss Institute of Bioinformatics, Mattenstrasse 26, 4058 Basel

Abstract Cancer progression is an evolutionary process shaped by both deterministic and stochastic forces. Multi-region and single-cell sequencing of tumors empower high-resolution reconstruction of the mutational history of each tumor. At the same time, it also highlights the extensive diversity across tumors and patients. Resolving the interactions among mutations and recovering the recurrent evolutionary processes may offer greater opportunities for successful therapeutic strategies. To this end, we present a novel probabilistic framework, called TreeMHN, for joint inference of repeated evolutionary trajectories and patterns of clonal exclusivity or co-occurrence from a cohort of intra-tumor phylogenetic trees. Through simulation studies, we show that TreeMHN outperforms existing alternatives that can only focus on one aspect of the task. By applying our method to an acute myeloid leukemia dataset, we find the most likely evolutionary trajectories and mutational patterns, consistent with and enriching known findings.

1 Introduction

Tumors emerge and develop malignancy through a somatic evolutionary process of accumulating selectively advantageous mutations in cells (Nowell, 1976). The genetic and phenotypic clonal diversity within a tumor, also known as intratumor heterogeneity (ITH), enables tumor cells to quickly adapt to micro-environmental changes, including those induced by treatment, often leading to lethal outcomes due to metastases or drug resistance (Marusyk and Polyak, 2010; Gerlinger *et al.*, 2012; Burrell *et al.*, 2013; McGranahan and Swanton, 2017; Dagogo-Jack and Shaw, 2018). Despite the inherent stochasticity of tumor evolution, recent evidence supported by increasingly available data and new computational methods has revealed at least some repeated features in tumor progression, such as frequently mutated genes (Lawrence *et al.*, 2013; Vogelstein *et al.*, 2013), specific order constraints on mutations (Ortmann *et al.*, 2015; Kent and Green, 2017; Levine *et al.*, 2019), and repeated evolutionary trajectories (Caravagna *et al.*, 2018; Khakabimamaghani *et al.*, 2019; Christensen *et al.*, 2020; Hodzic *et al.*, 2020). The ability to recover reproducible features of cancer evolution, and more importantly, to make reliable predictions of future evolutionary steps, is crucial for the development of successful therapeutic interventions (Lipinski *et al.*, 2016; Lässig *et al.*, 2017; Hosseini *et al.*, 2019; Diaz-Uriarte and Vasallo, 2019).

Recent advances in multi-region sequencing (Gerlinger *et al.*, 2012; Yates *et al.*, 2015; Turajlic *et al.*, 2018), single-cell sequencing (Lawson *et al.*, 2018; Lähnemann *et al.*, 2020), and phylogenetic tree inference (Kuipers *et al.*, 2017a; Schwartz and Schäffer, 2017) enable more precise characterization of clonal architecture and provide a clearer picture of tumor evolution. However, the increased resolution further substantiates the extensive variability in the subclonal compositions and mutational histories between tumors, making it more challenging to infer repeatable elements. For example, two parallel studies of acute myeloid leukemia (AML) using single-cell panel sequencing (Miles *et al.*, 2020; Morita *et al.*, 2020) show that the reconstructed trees typically contain a small number of clones and can vary considerably between any two patients. In particular, both studies report over-represented pairs of co-occurring or clonally exclusive mutations in the subclones. These relationships have recently been validated with a customized statistical test, called GeneAccord (Kuipers *et al.*, 2021), which accounts for the clonal dependency structures. Mutations that co-occur more frequently in the same clonal lineage may indicate synergistic effects on cell proliferation and survival (Park and Lehner, 2015). Clonally exclusive mutations, on the other hand, may suggest (1) clonal cooperation, where cell populations harboring complementary sets of mutations collaborate to promote tumor growth (Tabassum and Polyak, 2015), or (2) synthetic lethality, meaning that acquiring both mutations on the same genome will significantly reduce the viability of the clone (O'Neil *et al.*, 2017). Since clonal interactions have a major impact on intratumor heterogeneity and the observed evolutionary trajectories, it would be beneficial to exploit the pairwise frequencies of mutational events within and between subclones when modelling tumor progression.

*Correspondence: niko.beerenwinkel@bsse.ethz.ch

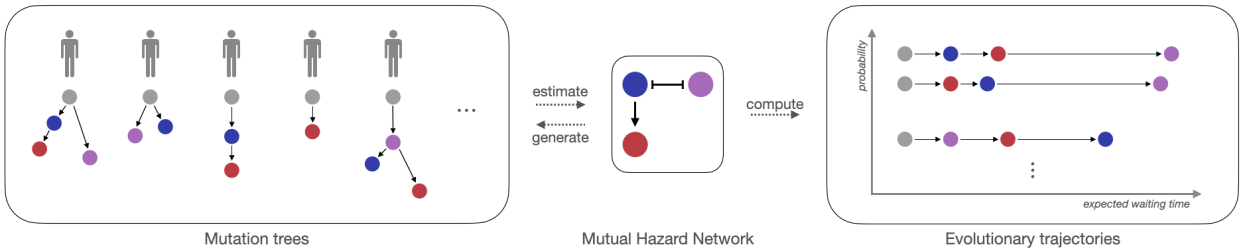

Figure 1: Schematic representation of TreeMHN. The middle panel is a Mutual Hazard Network (MHN) with three distinct mutations, and the edges represent the co-occurring ($\rightarrow$) or exclusive ($\dashv$) stochastic dependencies among the mutations. Given an MHN, it is possible to generate heterogeneous mutation trees (left panel) following the process described in Section 2.1. The grey circles (root nodes) represent the unmutated germline genotype, and the colored nodes represent the mutations. Given a set of trees for a patient cohort, we estimate the dependency parameters of an unknown underlying MHN using methods provided in Section 2.2. Conditioned on the estimated MHN, one can compute the probability and the expected waiting time of different evolutionary trajectories (right panel).

For cross-sectional bulk sequencing data, a collection of probabilistic methods for inferring the temporal order of mutations and predicting tumor progression is called cancer progression models (CPMs) (Beerenwinkel *et al.*, 2015; Diaz-Colunga and Diaz-Uriarte, 2020). In light of the observation that mutually exclusive mutations are often associated with genes in the same functional pathway (Yeang *et al.*, 2008; Vandin, 2017), Raphael and Vandin (2015) developed the first CPM that jointly infers sets of mutually exclusive mutations and the temporal order among them using a chain model. Later, pathTiMEx (Cristea *et al.*, 2017) generalized the linear structure with a continuous-time Conjunctive Bayesian Network (CBN) (Beerenwinkel and Sullivant, 2009; Gerstung *et al.*, 2009, 2011). The most recent method, Mutual Hazard Networks (MHNs), does not explicitly group exclusive mutations but re-parameterizes mutational waiting times using a matrix that encodes co-occurrence and exclusivity (Schill *et al.*, 2020).

However, the above methods require independent samples of binary genotypes as input. As such they have not been designed for tree-structured data as they cannot capture the subclonal structure within a tumor nor utilize the existing order information from the tumor phylogenies. On the one hand, mutual exclusivity is defined at the patient level, whereas clonal exclusivity refers to pairs of mutations that occur less frequently in the same subclone but can still co-exist in the same tumor (Kuipers *et al.*, 2021). Hence, if the consensus genotype of the tumor contains both mutations, then the above CPMs will treat it as evidence of co-occurrence. On the other hand, using the set of subclonal genotypes (rather than one consensus genotype per tumor) can produce misleading results, because the evolutionary relationships among the subclonal genotypes are overlooked. For instance, consider two mutations that occurred sequentially in a tumor resulting in the subclonal genotypes $(1, 0)$ and $(1, 1)$. While it is clear from the structure that the two mutations are co-occurring in the same lineage, treating $(1, 0)$ and $(1, 1)$ as independent observations implies that they are only co-occurring in half of the samples, and hence obscures the true relationship.

Recently, a number of methods have been developed to infer recurrent mutational trajectories from cross-sectional multi-region or single-cell sequencing data (Caravagna *et al.*, 2018; Khakabimamaghani *et al.*, 2019; Christensen *et al.*, 2020; Hodzic *et al.*, 2020). Based on transfer learning, REVOLVER (Caravagna *et al.*, 2016) reconciles the heterogeneous phylogenetic trees of a cohort of patients using a matrix summarizing the frequencies of all pairwise ancestor-descendant relationships across tumors and outputs the trees having the smallest distance to the matrix. The entries in the normalized matrix are empirical estimates of the probability of one mutation being the ancestor of another mutation, which can be used to compute the probability of a possible evolutionary trajectory. HINTRA (Khakabimamaghani *et al.*, 2019) extends the idea of REVOLVER by relaxing the assumption that the occurrence of a mutation depends only on its direct ancestor. It considers all possible sets of ancestors, allowing for more complex dependencies. However, the number of rows of the modified count matrix is exponential in the number of mutations, which limits the scalability of the algorithm (Christensen *et al.*, 2020). RECAP (Christensen *et al.*, 2020) and CONETT (Hodzic *et al.*, 2020) are two integer linear programming approaches that learn a consensus tree for a set of tumors. These two methods do not provide probabilistic measures of potential evolutionary trajectories and thus are more suitable for the task of clustering tumors into subtypes (Christensen *et al.*, 2020). It is worth

noting that none of these state-of-the-art methods incorporate clonal co-occurrences and exclusivities.

Therefore, we develop a generative model, called TreeMHN, for simultaneous inference of repeated evolutionary trajectories and patterns of clonal exclusivity and co-occurrence from a cohort of intra-tumor phylogenetic trees (Figure 1). TreeMHN improves upon MHNs by considering the complete mutational histories of tumors and their subclonal structures rather than the average presence and absence of mutations. We also provide an efficient parameter estimation scheme that can scale beyond 25 mutations, which is not feasible with MHNs. Using simulated data, we evaluate the accuracy of TreeMHN in estimating the mutual exclusivity parameters and the probability distribution of mutational pathways and compare it against MHNs, REVOLVER, and HINTRA. Finally, we apply TreeMHN to a set of single-cell mutation trees (Jahn *et al.*, 2016) for 123 AML patients (Morita *et al.*, 2020), compare our estimated patterns of mutual exclusivity with the significant gene pairs identified by GeneAccord (Kuipers *et al.*, 2021), and further differentiate their mutation orders and the most probably evolutionary trajectories.

2 Methods

We formulate the TreeMHN algorithm as a tree generating process, where the underlying mutational dependencies are encoded by a Mutual Hazard Network (Schill *et al.*, 2020). To ensure the scalability of the algorithm, we provide two inference methods based on maximum likelihood estimation (MLE) and a hybrid Monte Carlo expectation-maximization (MC-EM) algorithm (Wei and Tanner, 1990).

2.1 TreeMHN: a tree generating process

Unlike existing CPMs, including MHNs, which take genotypes from bulk sequencing data as input, TreeMHN works directly with individual intra-tumor phylogenetic trees, either a single tree or a sample of trees for each patient in a cohort, in order to exploit the evidence of mutation ordering and correctly account for clonal dependencies.

Consider N mutation trees $\boldsymbol{\mathscr{T}} = \{\mathscr{T}^1, \ldots, \mathscr{T}^N\}$ with a total number of n mutations. Each tree $\mathscr{T}^l$ corresponds to the mutational history of a tumor and contains a subset of mutations from $[n] := \{1, \ldots, n\}$. We assume that the trees are realizations of a Markov process with the transition rate matrix parameterized by a Mutual Hazard Network $\Theta \in \mathbb{R}^{n \times n}$ (Schill *et al.*, 2020). We denote each subclone in a tree by the mutational pathway π that runs from the root 0 to the node where the subclone is attached. In other words, a subclone π is a sequence $(0, \sigma_1, \ldots, \sigma_d)$ with $1 \leq d \leq n$, and $\sigma_i \in [n]$ are non-duplicated elements. Let Π denote the space of all subclones, or equivalently, the space of all mutational pathways or evolutionary trajectories. The tree generating process is defined as follows (Figure 2 & Supplementary Figure S1):

1. Starting from a subclone π, the set of mutations that could happen next is $[n] \setminus \pi$. The initial wild-type subclone is $\pi = (0)$ at time $T_{(0)} = 0$.

2. The waiting time until a new subclone (π, i) with $i \in [n] \setminus \pi$ is born from π is an exponentially distributed random variable,

$$T_{(\pi,i)} \sim T_\pi + \mathrm{Exp}(\lambda_{(\pi,i)}), \qquad \lambda_{(\pi,i)} = \exp\left(\theta_{ii} + \sum_{j \in \pi} \theta_{ij}\right) = \Theta_{ii} \prod_{j \in \pi} \Theta_{ij} \tag{1}$$

   where $\theta_{ii} = \log \Theta_{ii}$ is the baseline rate of mutation i, and $\theta_{ij} = \log \Theta_{ij}$ determines the positive (denoted $j \to i$), negative ($j \dashv i$) or zero effect of an existing mutation j on mutation i (Schill *et al.*, 2020). If both $\theta_{ij} < 0$ and $\theta_{ji} < 0$ (i.e., $i \vdash\dashv j$), then the two mutations are called clonally exclusive, and if both signs are positive (i.e., $i \leftrightarrow j$), we say that the mutations are co-occurring. The collection of all waiting times is denoted by $\mathbf{T} = \{T_\pi\}_{\pi \in \Pi}$.

3. The observed tree structure $\mathscr{T}$ is co-determined by an independent sampling event $T_s \sim \mathrm{Exp}(\lambda_s)$. Typically, we assume $\lambda_s = 1$ for identifiability (Gerstung *et al.*, 2009). An edge from π to (π, i) exists in $\mathscr{T}$ if and only if $T_{(\pi,i)} < T_s$, i.e., if mutation i happened before sampling the tumor cells.

4. The process iterates until all subclones that could emerge next have a longer waiting time than the sampling time. We denote the augmented tree structure by $A(\mathscr{T})$, which includes the edges pointing

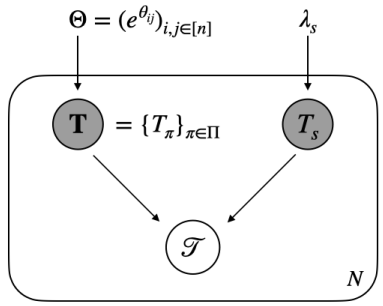

Figure 2: TreeMHN as a probabilistic graphical model. The waiting times of subclones $\mathbf{T} = \{T_\pi\}_{\pi\in\Pi}$ are exponentially distributed random variables parameterized by an MHN $\Theta = (e^{\theta_{ij}})_{i,j\in[n]}$. The tree structure $\mathscr{T}$ is jointly determined by $\mathbf{T}$ and an independent sampling event T_s, which is also an exponential random variable with rate λ_s. Both $\mathbf{T}$ and T_s are hidden variables, indicated by shaded nodes.

towards the events right after sampling. Events further downstream are not considered as they cannot influence the observed tree structure.

One advantage of this formulation is that mutations on the same gene can appear in different lineages of a tree. For computational convenience, most existing tree reconstruction algorithms adopt the infinite sites assumption (ISA), *i.e.* they require that each mutational event (*e.g.* an SNV) is gained at most once. However, it has been shown that allowing parallel mutations is, in general, a more realistic assumption (Kuipers *et al.*, 2017b). Even assuming the ISA for individual genomic bases, when summarising at the gene level the same gene may be affected in parallel lineages. In Section 3.2 below, we will demonstrate the applicability of TreeMHN using a real dataset which contains parallel mutations. Due to the introduction of clonally exclusive mutations, another advantage is the possibility to generate very distinct trees from given mutual hazards Θ. Therefore, TreeMHN can model extensive intra-tumor and inter-tumor heterogeneity without reconciling the trees or clustering patients into subgroups.

2.2 Efficient parameter estimation

The marginal probability of observing a tree $\mathscr{T}$ given Θ is equal to the probability that all observed mutational events happen before the sampling event, and all unobserved events that could happen next do not happen before the sampling event (Supplementary Figure S1),

$$p(\mathscr{T} \mid \Theta) = P\left(\max_{\pi\in\mathscr{T}} T_\pi < T_s < \min_{\substack{\pi'\notin\mathscr{T} \\ \pi'\in A(\mathscr{T})}} T_{\pi'} \;\middle|\; \Theta \right). \tag{2}$$

Both this term and the gradients $\partial p(\mathscr{T} \mid \Theta)/\partial\lambda_\pi$ can be computed efficiently by inverting specific triangular matrices, which are constructed using the rates associated with the events in $A(\mathscr{T})$, and the dimensions depend only on the number of subtrees in $\mathscr{T}$ (Supplementary Section A.1). Given N mutation trees $\boldsymbol{\mathscr{T}} = \{\mathscr{T}^1, \dots, \mathscr{T}^N\}$, we can follow Schill *et al.* (2020) and estimate Θ by maximizing the penalized log-likelihood,

$$\hat{\Theta} = \arg\max_{\Theta} \sum_{l=1}^{N} \log p(\mathscr{T}^l \mid \Theta) - \gamma \sum_{i\neq j} |\log \Theta_{ij}|, \tag{3}$$

where $\gamma > 0$ controls the sparsity of Θ in order to mitigate overfitting.

However, when some or all of the trees in $\boldsymbol{\mathscr{T}}$ have many subtrees (*e.g.* more than 500 subtrees), MLE can still be very slow or even infeasible. In this case, we resort to a hybrid EM and Monte Carlo EM algorithm based on importance sampling (Montazeri *et al.*, 2016) as follows (see also Supplementary Section A.2). In the E step, given the observed trees $\boldsymbol{\mathscr{T}}$ and the current estimate $\Theta^{(k)}$ for iteration $k \in \{0, 1, 2, \dots\}$, we compute the expected value of the complete-data log-likelihood with respect to the unobserved collection of

waiting times as

$$g(\Theta, \Theta^{(k)}) = \sum_{l=1}^{N} \sum_{i=1}^{n} \sum_{\pi \in \mathscr{T}^l : (\pi,i) \in A(\mathscr{T}^l)} \left[\log \Theta_{ii} + \sum_{j \in \pi} \log \Theta_{ij} - \Theta_{ii} \prod_{j \in \pi} \Theta_{ij} \mathbb{E}_{\boldsymbol{T}^l, T_s^l \mid \mathscr{T}^l, \Theta^{(k)}} (T^l_{(\pi,i)} - T^l_{\pi}) \right] + C, \quad (4)$$

where C is a constant. For small trees, we can calculate the expected time difference in exact form,

$$\mathbb{E}_{\boldsymbol{T}, T_s \mid \mathscr{T}, \Theta}[T_{(\pi,i)} - T_{\pi}] = \frac{1}{\lambda_{(\pi,i)}} - \frac{1}{P(\mathscr{T} \mid \Theta)} \frac{\partial P(\mathscr{T} \mid \Theta)}{\partial \lambda_{(\pi,i)}}. \quad (5)$$

For large trees, we approximate the expectation by drawing M samples from the following proposal distribution. First, the sampling time $T_s \sim \mathrm{Exp}(\lambda_s)$ with $\lambda_s = 1$ is drawn independently. Using the equation $T_{(\pi,i)} = T_{\pi} + Z_{(\pi,i)}$, we then follow the topological order in $\mathscr{T}$ and sample the difference in waiting times between subclones π and (π, i) recursively as

$$Z_{(\pi,i)} \sim \begin{cases} \mathrm{TExp}(\lambda_{(\pi,i)}, 0, t_s - t_{\pi}) & \text{if } (\pi,i) \in \mathscr{T} \\ \mathrm{TExp}(\lambda_{(\pi,i)}, t_s - t_{\pi}, \infty) & \text{if } (\pi,i) \in A(\mathscr{T}) \setminus \mathscr{T} \end{cases} \quad (6)$$

where $\mathrm{TExp}(\lambda, a, b)$ is the truncated exponential distribution with parameter λ and bounds $0 \le a < b < \infty$. Thus, the time points $\boldsymbol{t}^{(m)}$ and $t_s^{(m)}$ for $m = 1, \ldots, M$ generated from our proposal distribution are by definition compatible with the tree structure. The approximation is then

$$\begin{aligned} \mathbb{E}_{\boldsymbol{T}, T_s \mid \mathscr{T}, \Theta}[T_{(\pi,i)} - T_{\pi}] &\approx \frac{\frac{1}{M} \sum_{m=1}^{M} w^{(m)} (t^{(m)}_{(\pi,i)} - t^{(m)}_{\pi})}{\frac{1}{M} \sum_{m=1}^{M} w^{(m)}}, \\ w^{(m)} &= \prod_{i=1}^{n} \left[\prod_{(\pi,i) \in \mathscr{T}} (1 - e^{-\lambda_{(\pi,i)}(t_s^{(m)} - t_{\pi}^{(m)})}) \times \prod_{\substack{(\pi,i) \notin \mathscr{T} \\ (\pi,i) \in A(\mathscr{T})}} e^{-\lambda_{(\pi,i)}(t_s^{(m)} - t_{\pi}^{(m)})} \right]. \end{aligned} \quad (7)$$

In the M step, we update Θ by maximizing the penalized expected complete-data log-likelihood

$$g(\Theta, \Theta^{(k)}) - \gamma \sum_{i \neq j} |\log \Theta_{ij}|. \quad (8)$$

In the case where the number of observations is much smaller than the number of parameters to estimate, we can run TreeMHN with stability selection (Meinshausen and Bühlmann, 2010). In particular, the parameters in Θ are estimated over many subsamples of the trees $\boldsymbol{\mathscr{T}}$, and only those having a high probability of being non-zero are kept. This procedure can significantly improve the precision of identifying the true relationships among the mutations (see also Section 3.1 and Supplementary Section B).

2.3 Probability and expected waiting time of a pathway

Let $\Pi_d \subseteq \Pi$ be the set of mutational pathways of a fixed length d with $1 \le d \le n$. Given Θ, we compute the probability of a particular pathway $\pi \in \Pi_d$ as the product of competing exponentials (Hosseini *et al.*, 2019),

$$P_{\Theta}(\pi) = \prod_{i=1}^{d} \frac{\lambda_{(\pi_{i-1}, \sigma_i)}}{\sum_{j \in [n] \setminus \pi_{i-1}} \lambda_{(\pi_{i-1}, j)}}, \quad (9)$$

where $\pi_i = (0, \sigma_1, \ldots, \sigma_i) \subset \pi$. It follows that $(P_{\Theta}(\pi))_{\pi \in \Pi_d}$ is the probability distribution over Π_d with respect to Θ, since $\sum_{\pi \in \Pi_d} P_{\Theta}(\pi) = 1$. Suppose Q is another probability distribution over Π_d, then the Kullback-Leibler (KL) divergence from Q to P_{Θ},

$$D_{\mathrm{KL}}(P_{\Theta} \parallel Q) = \sum_{\pi \in \Pi_d} P_{\Theta}(\pi) \log \frac{P_{\Theta}(\pi)}{Q(\pi)} \quad (10)$$

measures the dissimilarity between the two distributions. In particular $D_{\mathrm{KL}}(P_\Theta \parallel Q) = 0$ if and only if $P_\Theta = Q$. Likewise, the expected waiting time of π is the sum of the expected time interval lengths along the pathway (Beerenwinkel and Sullivant, 2009),

$$\mathbb{E}_\Theta[T_\pi] = \sum_{i=1}^{d} \frac{1}{\sum_{j\in[n]\setminus\pi_{i-1}} \lambda_{(\pi_{i-1},j)}}. \tag{11}$$

2.4 Simulations

We use simulation studies to assess the performance of TreeMHN in estimating the network parameters Θ and the associated pathway probabilities from a set of mutation trees. Following Schill *et al.* (2020), we first randomly generate a ground truth network Θ with n mutations. The diagonal entries $\log \Theta_{ii}$ are drawn from a uniform distribution ranging from -6 to -1. A random half of the off-diagonal entries $\log \Theta_{ij}$ are set to zero, and another half are sampled from a Gamma distribution $\Gamma(\alpha, \beta)$ with $\alpha = 4$ and $\beta = 2.5$. The non-zero entries are then multiplied by -1 with a 50% chance. These values are chosen to be consistent with the real dataset in Section 3.2 (Supplementary Figure S9). Given Θ, we then generate N mutation trees (Section 2.1, Supplementary Section A.3), from which we obtain an estimate $\hat{\Theta}$ using TreeMHN. The performance metrics we use are:

1. Precision and recall of $\hat{\Theta}$ in identifying the edges ($i \quad j$, $i \to j$, $i \dashv j$) in Θ. We call an off-diagonal entry in $\hat{\Theta}$ true positive if and only if it is non-zero and has the correct sign (Supplementary Section C.1).

2. KL divergence from an estimated pathway distribution $P_{\hat{\Theta}}$ to P_Θ over Π_d with $d = 4$, which quantifies the deviation of the estimated pathway probabilities (Section 2.3). Note that longer pathway lengths ($d > 4$) can dramatically increase the computational complexity as the number of pathways in Π_d is $\frac{n!}{(n-d)!}$, especially for large n.

For the first metric, we compare TreeMHN against the baseline MHN method of Schill *et al.* (2020), the only previously published cancer progression model that can estimate Θ. Since MHNs cannot handle tree structures, we use the subclonal genotypes as input, weighted by the number of subclones within a tree. Even though the subclonal structure is lost, using multiple genotypes per tumor is still more informative for MHNs than a consensus genotype. In addition, we implement the stability selection procedure (Supplementary Section B) for both TreeMHN and MHNs to reduce the number of false positives. For the second metric, we also include REVOLVER (Caravagna *et al.*, 2018) and HINTRA (Khakabimamaghani *et al.*, 2019) for comparison. Since these two alternatives do not output an estimate of Θ, we compute their key matrices (denoted $\boldsymbol{w}$ in REVOLVER and β in HINTRA) using the edge frequencies directly from the trees, followed by constructing the probability distributions over Π_d with the matrix elements (Supplementary Section C.2).

We consider different configurations of simulation experiments, including varying the number of mutations $n \in \{10, 15, 20\}$, sample size $N \in \{200, 300, 500\}$, network sparsity (proportion of zero entries in Θ), and the ratio between positive and negative entries in Θ. To evaluate the robustness of TreeMHN against the uncertainty in estimated tree structures, we also introduce a noise level $\epsilon \in \{1\%, 5\%, 10\%, 20\%\}$, the probability of perturbing individual nodes in the simulated trees (Supplementary Section C.3). For each configuration, we repeat 100 times the process of simulating the data and estimating the parameters over a set of regularization parameters $\gamma \in \{0.05, 0.1, 0.5, 1, 1.5, 2, 2.5, 3\}$.

3 Results

3.1 Performance assessment on simulated data

Using simulated data (Section 2.4), we first evaluate how well TreeMHN can estimate the patterns of clonal co-occurrence and exclusivity against the baseline MHN method (Figure 3). As expected, the performance of both methods improves as sample size increases and mutation network size decreases. At the same regularization level, the solutions with stability selection can achieve much higher precision although at the cost of lower recall. We also notice that both methods perform better in recovering the pairwise interactions between mutations with higher baseline rates than those with lower baseline rates (Figure 3 vs. Supplementary Figure S3). The reason is that interactions between rare mutations can hardly be observed in the trees,

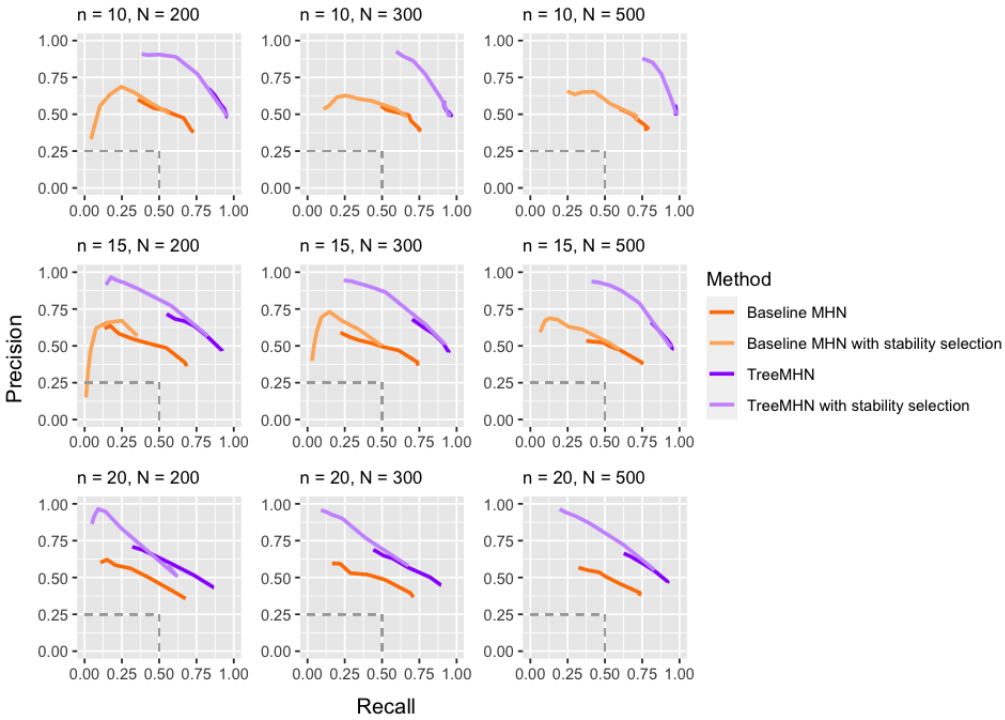

Figure 3: Performance of TreeMHN and the baseline MHN method (with and without stability selection) in estimating the true network Θ on simulated data for $n \in \{10, 15, 20\}$ mutations in total and $N \in \{200, 300, 500\}$ samples. The precision and recall curves averaged over 100 simulation runs are plotted for the top half of mutations ranked by baseline rates. Each point on the curves corresponds to a penalization level $\gamma \in \{0.05, 0.1, 0.5, 1, 1.5, 2, 2.5, 3\}$. The dash lines indicate the performance of randomly guessing the edge directions in Θ.

leading to a lack of statistical power to correctly estimate the corresponding entries in Θ. This general, method-independent limitation can also occur in practice (Section 3.2 and Supplementary Figure S8).

In all cases, TreeMHN clearly and consistently outperforms the baseline MHN approach in terms of both average precision and recall. This is because the problem of inferring an MHN from binary genotypes is in general underspecified, meaning that multiple different MHNs share similar likelihoods (Gotovos *et al.*, 2021), which cannot be alleviated by taking subclonal genotypes directly as input. This result highlights the benefit of utilizing the existing ordering information encoded in the trees to resolve the underlying network. In fact, this benefit is maintained even if high levels of noise ϵ are added to the trees (Supplementary Figure S4 and S5), as well as for varying proportions of zero entries (Supplementary Figure S6) and negative entries in Θ (Supplementary Figure S7).

Next, we look at the performance of TreeMHN in estimating pathway probabilities (Figure 4, Section 2.4). Across different sample sizes and network sizes, TreeMHN again outperforms all alternatives. HINTRA has the worst performance possibly because of over-parameterization in the matrix β given the limited number of observations for every possible ancestry set (Supplementary Section C.2). REVOLVER is only slightly worse than the baseline MHNs even though it does not explicitly model co-occurrence and exclusivity between mutations. Note that the baseline MHN method is unable to handle the case $n = 30$ due to exponentially increasing space and time complexity (Schill *et al.*, 2020). For TreeMHN, the runtime is rather limited by the tree with the maximum number of subtrees (Supplementary Table S1).

3.2 Application to AML tree data

We apply TreeMHN to a cohort of $N = 123$ AML patient samples analyzed by high-throughput single-cell panel sequencing, which involves 543 somatic mutations in $n = 31$ cancer-associated genes (Morita *et al.*, 2020). The mutation trees reconstructed by SCITE (Jahn *et al.*, 2016), a single-cell phylogenetic method, represent the complete evolutionary histories of the tumors. Despite the fact that different point mutations

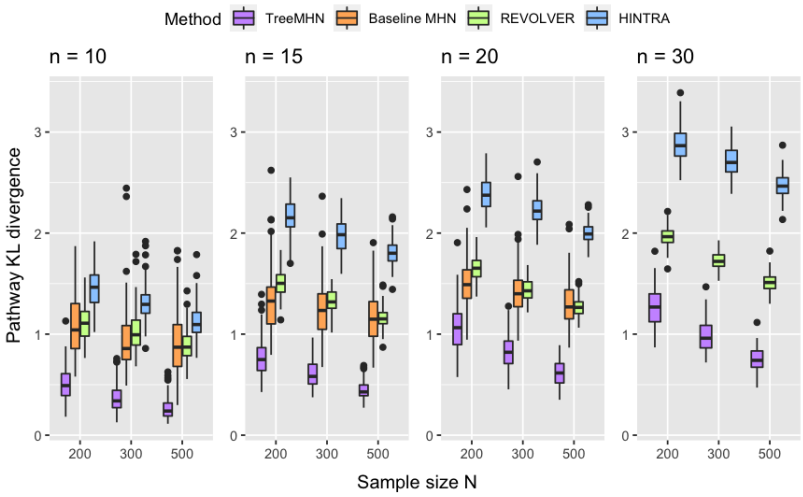

Figure 4: KL divergence from the distributions estimated using TreeMHN, baseline MHN, REVOLVER and HINTRA respectively to the ground truth distribution P_Θ over Π_d with $d = 4$.

in the same gene may have different effects, we summarize the mutations at the gene level and observe many mutated genes appearing more than once in two separately evolved subclones of the same tree, such as *FLT3*, *KRAS*, and *PTPN11* (Morita *et al.*, 2020). This allows us to substantially increase the statistical power for detecting recurrent gene-level trajectories and patterns of clonal exclusivity.

To avoid overfitting, we train TreeMHN with stability selection and obtain a network with only 15 events having non-zero off-diagonal entries (Figure 5 and Supplementary Figure S8). As also explained in Section 3.1, this is because the number of observations for the pairwise interactions between rare events is so small that the corresponding entries are filtered out during stability selection. In this case, even if an interaction exists, there is not enough power to detect it.

Among the 31 AML genes, *DNMT3A* has the highest baseline rate, followed by *IDH2*, *FLT3*, *NRAS*, *NPM1*, and *TET2*, which are known to be more frequently mutated in AML patients (Cancer Genome Atlas Research Network, 2013). However, these genes do not necessarily appear in every mutation tree or become the initiating events, since the appearance of a mutation depends not only on the baseline rate but also whether other mutations happened upstream with promoting or inhibiting effects, which further explains the high degree of heterogeneity in the trees. By comparing the off-diagonal entries against the list of gene pairs found by GeneAccord (Kuipers *et al.*, 2021) on the same dataset, we observe highly consistent but more informative results. In particular, most significant pairs of co-occurring or exclusive genes can be confirmed in the estimated network with additional directional strengths. For example, GeneAccord identifies *NRAS* and *FLT3* as significantly clonally exclusive. TreeMHN further tells that the effect of *FLT3* inhibiting the occurrence of *NRAS* in the same lineage is much stronger than the other way around. In other words, if a subclone has already acquired mutations in *FLT3*, then it is less likely to get another mutation in *NRAS*, which may still occur in separately-evolved subclones. On the contrary, subclones that acquire mutations in *NRAS* first are still very likely to hit *FLT3* subsequently. Apart from *NRAS*, such relationship also appears in *PTPN11* and *IDH1/2*. This observation aligns with previous studies, where some patients developed secondary resistance to *FLT3* inhibitors due to off-target mutations (*e.g.* genes in the RAS pathways), which were present in small populations prior to treatment (McMahon *et al.*, 2019; Kennedy and Smith, 2020). Another example consistent with the GeneAccord results is the two clonally exclusive *IDH* genes, which are directly co-occurring with *NPM1* and indirectly with mutations affecting the RAS/RTK/MAPK signalling pathways (*NRAS*, *KRAS*, *FLT3*, *PTPN11*). This result helps explain the two functionally similar subclones co-existing in patient sample AML-38-001 (Supplementary Figure S10), where the main difference is the divergence in *IDH1* and *IDH2*. Instead of accumulating both *IDH* genes in one clonal population, evolving separately can be more advantageous for the tumor, since the two subclones may cooperate to accelerate tumor progression (Tabassum and Polyak, 2015).

From the most likely mutational trajectories of length 4 inferred from the estimated network (Figure 6),

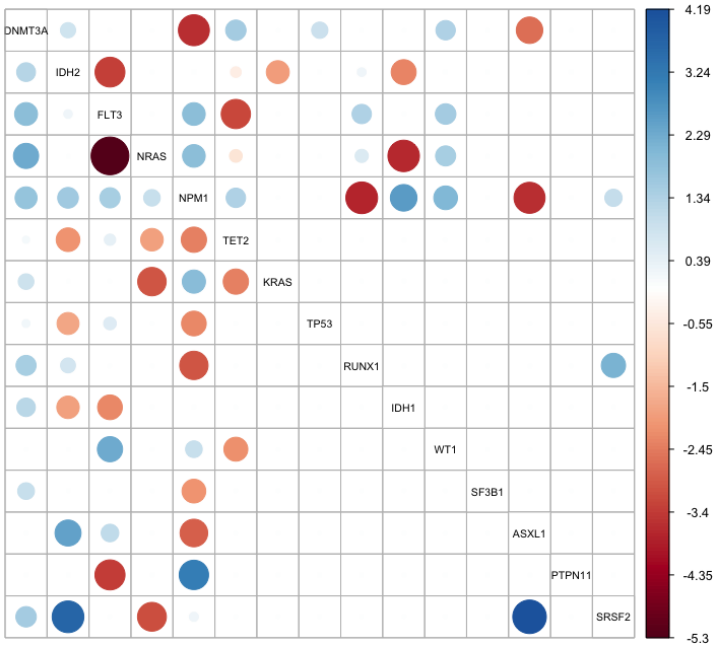

Figure 5: Partial Mutual Hazard Network learned from a cohort of 123 AML patient samples (Morita *et al.*, 2020). The full network is shown in Supplementary Figure S8. The columns and rows of the matrix is ordered by decreasing baseline rates. The empty off-diagonal entries represent the case $\theta_{ij} = 0$, meaning that there is no edge from mutation j to mutation i. The red circles correspond to $\theta_{ij} < 0$ and the blue $\theta_{ij} > 0$. The darker the colors, the stronger the effects.

we observe that *DNMT3A*, *NPM1*, and *FLT3* often co-occur with a relatively high probability. This three-way interaction is found to be associated with poor prognosis (Loghavi *et al.*, 2014; Bezerra *et al.*, 2020). With *DNMT3A* $\rightarrow$ *NPM1* and *NPM1* $\dashv$ *DNMT3A*, the ordering between them is more likely to be *DNMT3A* first followed by *NPM1*, which is consistent with previous findings (Shlush *et al.*, 2014; Papaemmanuil *et al.*, 2016). Moreover, conditioned on the estimated network and a tree structure, we can compute the next most probable mutational event (Supplementary Section A.4). Take again patient sample AML-38-001 as an example. The next mutation will be more likely to hit *WT1* on the two *FLT3* branches, followed by *SRSF2* in the subclone led by *IDH2* (Supplementary Figure S10). Note that if we assess the next possible mutations based on baseline rates only, neither *WT1* nor *SRSF2* will appear on top of the list.

4 Discussion

We have developed TreeMHN, a novel cancer progression model for the joint inference of repeated evolutionary trajectories and patterns of clonal co-occurrence or exclusivity from a cohort of intra-tumor phylogenetic trees reconstructed using multi-region or single-cell sequencing data. As an extension of Mutual Hazard Networks (Schill *et al.*, 2020), TreeMHN can take as input heterogeneous tree structures, rather than per-tumor or per-clone consensus genotypes. Importantly, with our efficient parameter estimation procedure, it is the maximum tree size, rather than the total number of mutations, which is typically much larger, that limits the computation time of TreeMHN.

Through simulation studies, we demonstrated the superior performance of TreeMHN in estimating both patterns of clonal exclusivity or co-occurrence and pathway probabilities in comparison with MHNs, REVOLVER (Caravagna *et al.*, 2018), and HINTRA (Khakabimamaghani *et al.*, 2019). Moreover, we have shown that TreeMHN is robust against uncertainty in the phylogenetic trees. Alternatively, one may sample trees for each patient proportionally to their posterior and refit TreeMHN, followed by averaging the corresponding hazard networks. By exploiting the evidence of temporal ordering among mutations contained in the tree topologies and properly accounting for clonal dependencies, TreeMHN can better resolve the underlying network structure. Apart from utilizing the tree structures, another remedy is to use longitudinal

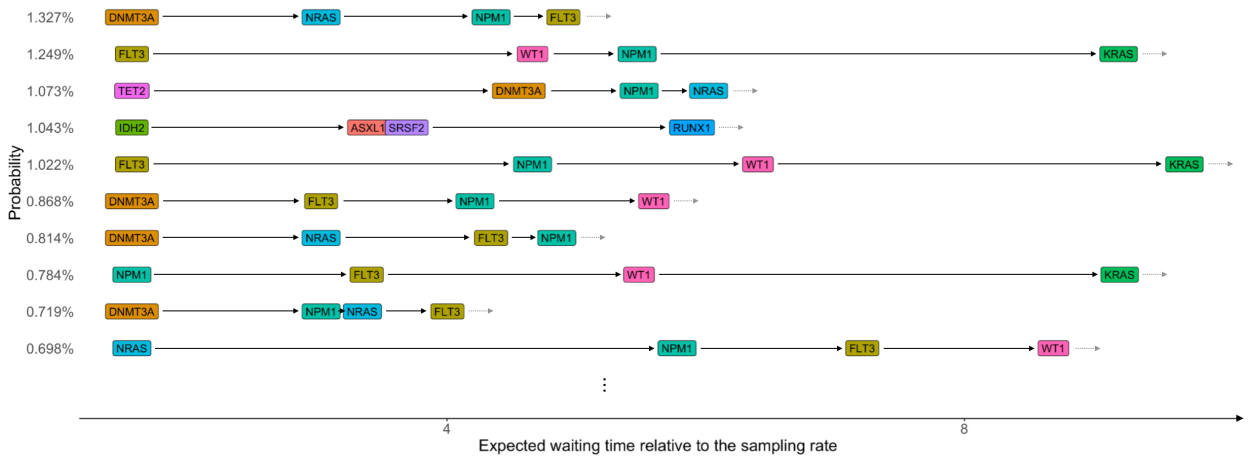

Figure 6: Top 10 most likely mutational pathways of length 4 inferred from the Mutual Hazard Network of the AML dataset (Morita *et al.*, 2020) ordered by pathway probabilities. The arrows represent the mutational ordering, and their lengths correspond to the expected waiting times of the mutations relative to the sampling rate of $\lambda_s = 1$, which are computed based on Eq. (11).

data, where the sampling time is either provided or can be inferred from data. Including observed sampling times is technically a straightforward task (Montazeri *et al.*, 2015, 2016), but such data is often difficult to obtain without having treatment interventions. Modelling drug response data is a crucial but challenging direction to explore, and TreeMHN may serve as a basis for future solutions.

Unlike REVOLVER and HINTRA, our method embraces the heterogeneity among the trees and incorporates clonal exclusivity and co-occurrence into the analysis of recurrent evolutionary processes. Also, TreeMHN does not rely on any particular phylogenetic method. It is possible to compare or combine different phylogenies from various sources (*e.g.* Jahn *et al.* (2016); Singer *et al.* (2018); Zafar *et al.* (2017)) in order to increase statistical power and correct for input bias. Future developments in phylogenetic methods together with more available data can further improve the estimate of TreeMHN. Another advantage of TreeMHN is the ability to model parallel mutations in distinct lineages, which are not uncommon in real data (Morita *et al.*, 2020), while existing alternatives require the infinite sites assumption. Like all other progression models, however, TreeMHN is currently not considering back mutations, *i.e.* situations in which a mutation is acquired at first but subsequently lost (Kuipers *et al.*, 2017b). A possible extension along this line is to include additional parameters and use as input phylogenetic trees inferred by methods such as SCARLET (Satas *et al.*, 2020), which views a decrease in copy numbers that overlap a mutated locus as evidence of back mutations. Moreover, TreeMHN is not designed for a specific type of mutation, such as SNVs. In other words, it is possible to detect recurrent trajectories at the level of copy number alterations (Kuipers *et al.*, 2020), mixed types of mutations (Chen *et al.*, 2020), or even functional pathways (Moore *et al.*, 2021).

We have applied our method to a real dataset containing single-cell mutation trees for 123 AML patients (Morita *et al.*, 2020). The inferred patterns of clonal exclusivity are highly consistent with the results from a statistical testing framework called GeneAccord (Kuipers *et al.*, 2021) but now includes directional effects and relate closely to known biological findings. Beyond detecting the existence of pairwise interactions, TreeMHN estimates the baseline rates of mutational events, directional strengths of the enabling or inhibiting effects, as well as the probabilistic temporal ordering among mutations. Nevertheless, TreeMHN does not take into account the subclone sizes, which can be viewed as consequences of clonal selection (Skums *et al.*, 2019). As noted by Tsyvina *et al.* (2020), inferring mutation rates and fitness values jointly may be a promising but challenging solution. On the one hand, larger subclones can be attributed to both their earlier appearance or higher fitness (Tsyvina *et al.*, 2020). On the other hand, the mutation rates in cancer progression models are implicitly referring to rates of evolution. Thus, any attempts to modify TreeMHN to model clonal selection need to be taken with caution. Given its flexibility, TreeMHN can be applied to tree data of different cancer types and the results may be useful for further experimental studies as well as treatment designs.

Acknowledgements

The authors would like to thank Katharina Jahn for providing the phylogenetic trees from Morita *et al.* (2020).

Availability of data and materials

The R package for TreeMHN, the processed AML mutation trees, and the associated analysis are available at `https://github.com/cbg-ethz/TreeMHN`.

Author contributions

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

Competing interests

The authors declare that they have no competing interests.

References

Beerenwinkel, N. and Sullivant, S. (2009). Markov models for accumulating mutations. *Biometrika*, **96**, 645–661.

Beerenwinkel, N., Schwarz, R. F., Gerstung, M., and Markowetz, F. (2015). Cancer evolution: Mathematical models and computational inference. *Systematic Biology*, **64**, e1–e25.

Bezerra, M. F., Lima, A. S., Piqué-Borràs, M.-R., Silveira, D. R., Coelho-Silva, J. L., Pereira-Martins, D. A., Weinhäuser, I., Franca-Neto, P. L., Quek, L., Corby, A., Oliveira, M. M., Lima, M. M., de Assis, R. A., de Melo Campos, P., Duarte, B. K., Bendit, I., Rocha, V., Rego, E. M., Traina, F., Saad, S. T., Beltrão, E. I., Bezerra, M. A., and Lucena-Araujo, A. R. (2020). Co-occurrence of DNMT3A, NPM1, FLT3 mutations identifies a subset of acute myeloid leukemia with adverse prognosis. *Blood*, **135**, 870–875.

Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, **501**, 338–345.

Cancer Genome Atlas Research Network (2013). Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *New England Journal of Medicine*, **368**, 2059–2074.

Caravagna, G., Graudenzi, A., Ramazzotti, D., Sanz-Pamplona, R., De Sano, L., Mauri, G., Moreno, V., Antoniotti, M., and Mishra, B. (2016). Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, E4025–E4034.

Caravagna, G., Giarratano, Y., Ramazzotti, D., Tomlinson, I., Graham, T. A., Sanguinetti, G., and Sottoriva, A. (2018). Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nature Methods*, **15**, 707–714.

Chen, Z., Gong, F., Wan, L., and Ma, L. (2020). BiTSC^2: Bayesian inference of Tumor clonal Tree by joint analysis of Single-Cell SNV and CNA data. *bioRxiv/2020.11.30.380949*.

Christensen, S., Kim, J., Chia, N., Koyejo, O., and El-Kebir, M. (2020). Detecting evolutionary patterns of cancers using consensus trees. *Bioinformatics*, **36**, i684–i691.

Cristea, S., Kuipers, J., and Beerenwinkel, N. (2017). pathTiMEx: joint inference of mutually exclusive cancer pathways and their progression dynamics. *Journal of Computational Biology*, **24**, 603–615.

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

Diaz-Colunga, J. and Diaz-Uriarte, R. (2020). Conditional prediction of consecutive tumor evolution using cancer progression models: What genotype comes next? *bioRxiv/2020.12.16.423099*.

Diaz-Uriarte, R. and Vasallo, C. (2019). Every which way? On predicting tumor evolution using cancer progression models. *PLoS Computational Biology*, **15**, e1007246.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N. Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C. R., Nohadani, M., Eklund, A. C., Spencer-Dene, B., Clark, G., Pickering, L., Stamp, G., Gore, M., Szallasi, Z., Downward, J., Futreal, P. A., and Swanton, C. (2012). Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine*, **366**, 883–892.

Gerstung, M., Baudis, M., Moch, H., and Beerenwinkel, N. (2009). Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics*, **25**, 2809–2815.

Gerstung, M., Eriksson, N., Lin, J., Vogelstein, B., and Beerenwinkel, N. (2011). The temporal order of genetic and pathway alterations in tumorigenesis. *PLoS ONE*, **6**, e27136.

Gotovos, A., Burkholz, R., Quackenbush, J., and Jegelka, S. (2021). Scaling up Continuous-Time Markov Chains Helps Resolve Underspecification. *arXiv/2107.02911*.

Hodzic, E., Shrestha, R., Malikic, S., Collins, C. C., Litchfield, K., Turajlic, S., and Sahinalp, S. C. (2020). Identification of conserved evolutionary trajectories in tumors. *Bioinformatics*, **36**, i427–i435.

Hosseini, S.-R., Diaz-Uriarte, R., Markowetz, F., and Beerenwinkel, N. (2019). Estimating the predictability of cancer evolution. *Bioinformatics*, **35**, i389–i397.

Jahn, K., Kuipers, J., and Beerenwinkel, N. (2016). Tree inference for single-cell data. *Genome Biology*, **17**, 1–17.

Kennedy, V. E. and Smith, C. C. (2020). FLT3 Mutations in Acute Myeloid Leukemia: Key Concepts and Emerging Controversies. *Frontiers in Oncology*, **10**, 2927.

Kent, D. G. and Green, A. R. (2017). Order matters: The order of somatic mutations influences cancer evolution. *Cold Spring Harbor Perspectives in Medicine*, **7**, a027060.

Khakabimamaghani, S., Malikic, S., Tang, J., Ding, D., Morin, R., Chindelevitch, L., and Ester, M. (2019). Collaborative intra-tumor heterogeneity detection. *Bioinformatics*, **35**, i379–i388.

Kuipers, J., Jahn, K., and Beerenwinkel, N. (2017a). Advances in understanding tumour evolution through single-cell sequencing. *Biochimica et Biophysica Acta - Reviews on Cancer*, **1867**, 127–138.

Kuipers, J., Jahn, K., Raphael, B. J., and Beerenwinkel, N. (2017b). Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Research*, **27**, 1885–1894.

Kuipers, J., Tuncel, M. A., Ferreira, P., Jahn, K., and Beerenwinkel, N. (2020). Single-cell copy number calling and event history reconstruction. *bioRxiv/2020.04.28.065755*.

Kuipers, J., Moore, A. L., Jahn, K., Schraml, P., Wang, F., Morita, K., Futreal, A., Takahashi, K., Beisel, C., Moch, H., *et al.* (2021). Statistical tests for intra-tumour clonal co-occurrence and exclusivity. *bioRxiv/2021.05.05.442732*.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S. O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. d., Cappuccio, A., Corleone, G., Dutilh, B. E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T. J., Keizer, E. M., Khatri, I., Kielbasa, S. M., Korbel, J. O., Kozlov, A. M., Kuo, T. H., Lelieveldt, B. P., Manoptdoiu, I. I., Marioni, J. C., Marschall, T., Mölder, F., Niknejad, A., Raczkowski, L., Reinders, M., Ridder, J. d., Saliba, A. E., Somarakis, A., Stegle, O., Theis, F. J., Yang, H., Zelikovsky, A., McHardy, A. C., Raphael, B. J., Shah, S. P., and Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, **21**, 1–35.

Lässig, M., Mustonen, V., and Walczak, A. M. (2017). Predicting evolution. *Nature Ecology and Evolution*, **1**, 1–9.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., Dicara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D. A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. A., Mora, J., Lee, R. S., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R., Gordenin, D. A., Sunyaev, S., Lander, E. S., and Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214–218.

Lawson, D. A., Kessenbrock, K., Davis, R. T., Pervolarakis, N., and Werb, Z. (2018). Tumour heterogeneity and metastasis at single-cell resolution. *Nature Cell Biology*, **20**, 1349–1360.

Levine, A. J., Jenkins, N. A., and Copeland, N. G. (2019). The Roles of Initiating Truncal Mutations in Human Cancers: The Order of Mutations and Tumor Cell Type Matters. *Cancer Cell*, **35**, 10–15.

Lipinski, K. A., Barber, L. J., Davies, M. N., Ashenden, M., Sottoriva, A., and Gerlinger, M. (2016). Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. *Trends in Cancer*, **2**, 49–63.

Loghavi, S., Zuo, Z., Ravandi, F., Kantarjian, H. M., Bueso-Ramos, C., Zhang, L., Singh, R. R., Patel, K. P., Medeiros, L. J., Stingo, F., Routbort, M., Cortes, J., Luthra, R., and Khoury, J. D. (2014). Clinical features of De Novo acute myeloid leukemia with concurrent DNMT3A, FLT3 and NPM1 mutations. *Journal of Hematology & Oncology*, **7**, 1–10.

Marusyk, A. and Polyak, K. (2010). Tumor heterogeneity: Causes and consequences. *Biochimica et Biophysica Acta - Reviews on Cancer*, **1805**, 105–117.

McGranahan, N. and Swanton, C. (2017). Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*, **168**, 613–628.

McMahon, C. M., Ferng, T., Canaani, J., Wang, E. S., Morrissette, J. J., Eastburn, D. J., Pellegrino, M., Durruthy-Durruthy, R., Watt, C. D., Asthana, S., Lasater, E. A., DeFilippis, R., Peretz, C. A., McGary, L. H., Deihimi, S., Logan, A. C., Luger, S. M., Shah, N. P., Carroll, M., Smith, C. C., and Perl, A. E. (2019). Clonal Selection with RAS Pathway Activation Mediates Secondary Clinical Resistance to Selective FLT3 Inhibition in Acute Myeloid Leukemia. *Cancer Discovery*, **9**, 1050–1063.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**, 417–473.

Miles, L. A., Bowman, R. L., Merlinsky, T. R., Csete, I. S., Ooi, A. T., Durruthy-Durruthy, R., Bowman, M., Famulare, C., Patel, M. A., Mendez, P., Ainali, C., Demaree, B., Delley, C. L., Abate, A. R., Manivannan, M., Sahu, S., Goldberg, A. D., Bolton, K. L., Zehir, A., Rampal, R., Carroll, M. P., Meyer, S. E., Viny, A. D., and Levine, R. L. (2020). Single-cell mutation analysis of clonal evolution in myeloid malignancies. *Nature*, **587**, 477–482.

Montazeri, H., Günthard, H. F., Yang, W. L., Kouyos, R., and Beerenwinkel, N. (2015). Estimating the dynamics and dependencies of accumulating mutations with applications to HIV drug resistance. *Biostatistics*, **16**, 713–726.

Montazeri, H., Kuipers, J., Kouyos, R., Böni, J., Yerly, S., Klimkait, T., Aubert, V., Günthard, H. F., Beerenwinkel, N., and Study, S. H. C. (2016). Large-scale inference of conjunctive bayesian networks. *Bioinformatics*, **32**, 727–735.

Moore, A. L., Batavia, A. A., Kuipers, J., Singer, J., Burcklen, E., Schraml, P., Beisel, C., Moch, H., and Beerenwinkel, N. (2021). Spatial distribution of private gene mutations in clear cell renal cell carcinoma. *Cancers*, **13**, 2163.

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., and Takahashi, K. (2020). Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nature communications*, **11**, 1–17.

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

O'Neil, N. J., Bailey, M. L., and Hieter, P. (2017). Synthetic lethality and cancer. *Nature Reviews Genetics*, **18**, 613–623.

Ortmann, C. A., Kent, D. G., Nangalia, J., Silber, Y., Wedge, D. C., Grinfeld, J., Baxter, E. J., Massie, C. E., Papaemmanuil, E., Menon, S., Godfrey, A. L., Dimitropoulou, D., Guglielmelli, P., Bellosillo, B., Besses, C., Döhner, K., Harrison, C. N., Vassiliou, G. S., Vannucchi, A., Campbell, P. J., and Green, A. R. (2015). Effect of Mutation Order on Myeloproliferative Neoplasms. *New England Journal of Medicine*, **372**, 601–612.

Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V. I., Paschka, P., Roberts, N. D., Potter, N. E., Heuser, M., Thol, F., Bolli, N., Gundem, G., Van Loo, P., Martincorena, I., Ganly, P., Mudie, L., McLaren, S., O'Meara, S., Raine, K., Jones, D. R., Teague, J. W., Butler, A. P., Greaves, M. F., Ganser, A., Döhner, K., Schlenk, R. F., Döhner, H., and Campbell, P. J. (2016). Genomic classification and prognosis in acute myeloid leukemia. *New England Journal of Medicine*, **374**, 2209–2221.

Park, S. and Lehner, B. (2015). Cancer type-dependent genetic interactions between cancer driver alterations indicate plasticity of epistasis across cell types. *Molecular Systems Biology*, **11**, 824.

Raphael, B. J. and Vandin, F. (2015). Simultaneous Inference of Cancer Pathways and Tumor Progression from Cross-Sectional Mutation Data. *Journal of Computational Biology*, **22**, 510–527.

Satas, G., Zaccaria, S., Mon, G., and Raphael, B. J. (2020). SCARLET: Single-Cell Tumor Phylogeny Inference with Copy-optnumber Constrained Mutation Losses. *Cell Systems*, **10**, 323–332.

Schill, R., Solbrig, S., Wettig, T., and Spang, R. (2020). Modelling cancer progression using Mutual Hazard Networks. *Bioinformatics*, **36**, 241–249.

Schwartz, R. and Schäffer, A. A. (2017). The evolution of tumour phylogenetics: Principles and practice. *Nature Reviews Genetics*, **18**, 213–229.

Shlush, L. I., Zandi, S., Mitchell, A., Chen, W. C., Brandwein, J. M., Gupta, V., Kennedy, J. A., Schimmer, A. D., Schuh, A. C., Yee, K. W., McLeod, J. L., Doedens, M., Medeiros, J. J. F., Marke, R., Kim, H. J., Lee, K., McPherson, J. D., Hudson, T. J., Pan-Leukemia Gene Panel Consortium, T. H., Brown, A. M. K., Yousif, F., Trinh, Q. M., Stein, L. D., Minden, M. D., Wang, J. C. Y., and Dick, J. E. (2014). Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature*, **506**, 328–333.

Singer, J., Kuipers, J., Jahn, K., and Beerenwinkel, N. (2018). Single-cell mutation identification via phylogenetic inference. *Nature Communications*, **9**, 1–8.

Skums, P., Tsyvina, V., and Zelikovsky, A. (2019). Inference of clonal selection in cancer populations using single-cell sequencing data. *Bioinformatics*, **35**(14), i398–i407.

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

Tsyvina, V., Zelikovsky, A., Snir, S., and Skums, P. (2020). Inference of mutability landscapes of tumors from single cell sequencing data. *PLOS Computational Biology*, **16**, e1008454.

Turajlic, S., Xu, H., Litchfield, K., Rowan, A., Chambers, T., Lopez, J. I., Nicol, D., O'Brien, T., Larkin, J., Horswell, S., Stares, M., Au, L., Jamal-Hanjani, M., Challacombe, B., Chandra, A., Hazell, S., Eichler-Jonsson, C., Soultati, A., Chowdhury, S., Rudman, S., Lynch, J., Fernando, A., Stamp, G., Nye, E., Jabbar, F., Spain, L., Lall, S., Guarch, R., Falzon, M., Proctor, I., Pickering, L., Gore, M., Watkins, T. B., Ward, S., Stewart, A., DiNatale, R., Becerra, M. F., Reznik, E., Hsieh, J. J., Richmond, T. A., Mayhew, G. F., Hill, S. M., McNally, C. D., Jones, C., Rosenbaum, H., Stanislaw, S., Burgess, D. L., Alexander, N. R., and Swanton, C. (2018). Tracking Cancer Evolution Reveals Constrained Routes to Metastases: TRACERx Renal. *Cell*, **173**, 581–594.

Vandin, F. (2017). Computational methods for characterizing cancer mutational heterogeneity. *Frontiers in genetics*, **8**, 83.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer Genome Landscapes. *Science*, **339**, 1546–1558.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, **85**, 699–704.

Yates, L. R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., Aas, T., Alexandrov, L. B., Larsimont, D., Davies, H., Li, Y., Ju, Y. S., Ramakrishna, M., Haugland, H. K., Lilleng, P. K., Nik-Zainal, S., McLaren, S., Butler, A., Martin, S., Glodzik, D., Menzies, A., Raine, K., Hinton, J., Jones, D., Mudie, L. J., Jiang, B., Vincent, D., Greene-Colozzi, A., Adnet, P. Y., Fatima, A., Maetens, M., Ignatiadis, M., Stratton, M. R., Sotiriou, C., Richardson, A. L., Lønning, P. E., Wedge, D. C., and Campbell, P. J. (2015). Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature Medicine*, **21**, 751–759.

Yeang, C.-H., McCormick, F., and Levine, A. (2008). Combinatorial patterns of somatic gene mutations in cancer. *The FASEB Journal*, **22**, 2605–2622.

Zafar, H., Tzen, A., Navin, N., Chen, K., and Nakhleh, L. (2017). SiFit: Inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biology*, **18**, 1–20.