Joint inference of exclusivity patterns and recurrent trajectories 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 enable high-resolution reconstruction of the mutational history of each tumor and highlight the extensive diversity across tumors and patients. Resolving the interactions among mutations and recovering recurrent evolutionary processes may offer greater opportunities for successful therapeutic strategies. To this end, we present a novel probabilistic framework, called TreeMHN, for the joint inference of exclusivity patterns and recurrent trajectories from a cohort of intra-tumor phylogenetic trees. Through simulations, we show that TreeMHN outperforms existing alternatives that can only focus on one aspect of the task. By analyzing datasets of blood, lung, and breast cancers, we find the most likely evolutionary trajectories and mutational patterns, consistent with and enriching our current understanding of tumorigenesis. Moreover, TreeMHN facilitates the prediction of tumor evolution and provides probabilistic measures on the next mutational events given a tumor tree, a prerequisite for evolution-guided treatment strategies.

Introduction

Tumors emerge and develop malignancy through a somatic evolutionary process of accumulating selectively advantageous mutations in cells [1]. 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 [2, 3]. 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 [4], specific order constraints on mutations [5, 6], and repeated evolutionary trajectories [7–10]. 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 [11–14].

Recent advances in multi-region sequencing [15, 16], single-cell sequencing [17], and phylogenetic tree inference [18, 19] 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 [20, 21] 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 examined with a customized statistical test, called GeneAccord [22], which is designed for evolutionary-related samples from the same tumor. Mutations that co-occur more frequently in the same clonal lineage may indicate synergistic effects on cell proliferation and survival [4]. Clonally exclusive mutations, on the other hand, occur more frequently in different lineages. They may suggest either clonal cooperation, where cell populations harboring complementary sets of mutations collaborate to promote tumor growth [23], or synthetic lethality, meaning that acquiring both mutations on the same genome will significantly reduce the viability of the clone [24]. Since clonal interactions have a major impact on intratumor heterogeneity and the observed evolutionary trajectories, it would be natural to incorporate mutational interdependencies within and between subclones when modelling tumor progression.

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) [25, 26]. In light of the observation that mutually exclusive mutations are often associated with genes in the same

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

functional pathway, Raphael & Vandin developed the first CPM that jointly infers sets of mutually exclusive mutations and the temporal order among them using a chain model [27]. Later, pathTiMEx [28] generalized the linear structure with a continuous-time Conjunctive Bayesian Network (CBN) [29]. A recent method, called Mutual Hazard Networks (MHNs), does not explicitly group exclusive mutations but re-parameterizes mutational waiting times using a matrix that encodes both co-occurrence and exclusivity [30]. However, these 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. In addition, 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 [22]. Hence, the consensus genotype of tumor with clonally exclusive mutations will contain both mutations, but the above CPMs will treat it as evidence of co-occurrence.

First efforts have been made to infer recurrent mutational trajectories from cross-sectional multi-region or single-cell sequencing data. Based on transfer learning, REVOLVER [7] 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 [8] 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 [9]. RECAP [9] and CONETT [10] 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 [9]. It is worth noting that none of these state-of-the-art methods incorporate clonal co-occurrences and exclusivities.

Here, we present a novel generative model, called TreeMHN, for simultaneous inference of patterns of clonal exclusivity and co-occurrence and repeated evolutionary trajectories from a cohort of intra-tumor phylogenetic trees. TreeMHN considers the complete mutational histories of tumors and their subclonal

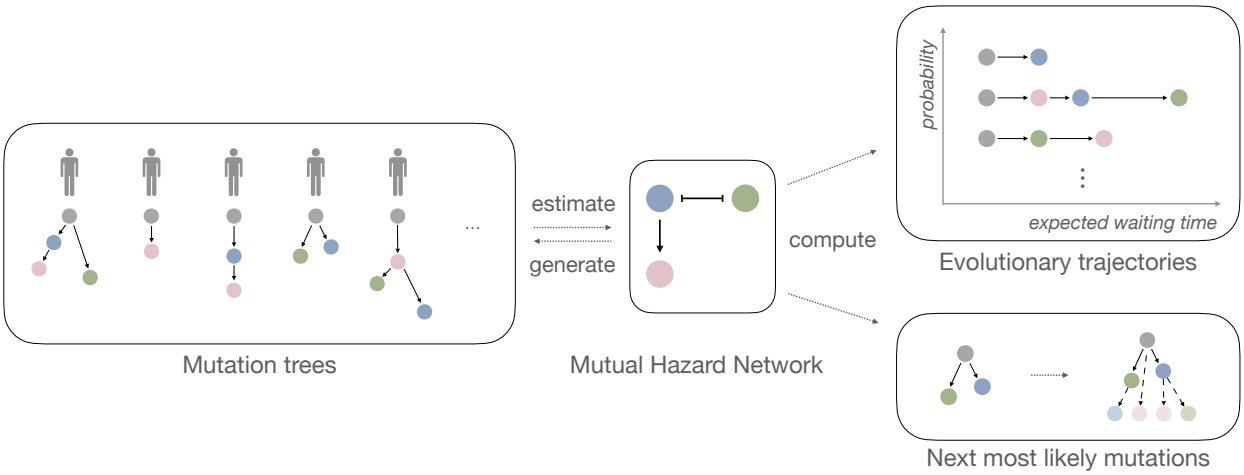


Figure 1: Schematic representation of TreeMHN. The middle panel is a Mutual Hazard Network (MHN) with three distinct mutations, and the edges represent the co-occurring (→) or exclusive (⊣) stochastic dependencies among the mutations. Given an MHN, TreeMHN generates heterogeneous mutation trees (left panel). The grey circles (root nodes) represent the start of the evolution, and the colored nodes represent the mutations. Given a set of trees for a patient cohort (left panel), we estimate the dependency parameters of an unknown underlying MHN. Conditioned on the estimated MHN, we can compute the probability and the expected waiting time of any evolutionary trajectory (upper right panel). Additionally, we can compute the next most probable mutational events given an existing tree (lower right panel).

structures rather than the overall presence and absence of mutations. Using simulated data, we demonstrate the superior performance of TreeMHN over consensus genotype MHNs [30], REVOLVER [7], and HINTRA [8] in estimating the mutual exclusivity parameters and the probability distribution of mutational pathways. Finally, we apply TreeMHN to three cancer cohorts: acute myeloid leukemia (AML) [21], non-small-cell lung cancer (NSCLC) [15], and breast cancer [31]. Our estimated exclusivity patterns and most probable mutational pathways not only confirm previous findings but also provide new insights into the interdependencies of mutations.

Results

TreeMHN overview. We developed TreeMHN, a probabilistic model for inferring exclusivity patterns of mutations and recurrent evolutionary trajectories from tumor mutation trees. In particular, we postulate an underlying tree-generating process where the mutational dependencies are encoded by a Mutual Hazard Network $\Theta \in \mathbb{R}^{n \times n}$ for n mutations. Existing mutations can have positive (co-occurring), negative (exclusive), or zero (no) effects on downstream mutations. We also provide an efficient parameter estimation scheme based on regularized maximum likelihood estimation and a hybrid Monte Carlo expectation-maximization algorithm. The computational complexity depends on the maximum tree size, which is often much smaller than the number of mutations. The estimated model allows us to compute the probabilities of different trajectories or evaluate the next most likely mutational events given a tumor tree. We provide an overview of our method in Figure 1 and more details can be found in the Methods section.

Performance assessment on simulated data. Through simulations (Methods), we first evaluate how well TreeMHN can estimate the patterns of clonal co-occurrence and exclusivity compared to the genotype MHN method (Figure 2). As expected, the performance of both methods improves as sample size increases

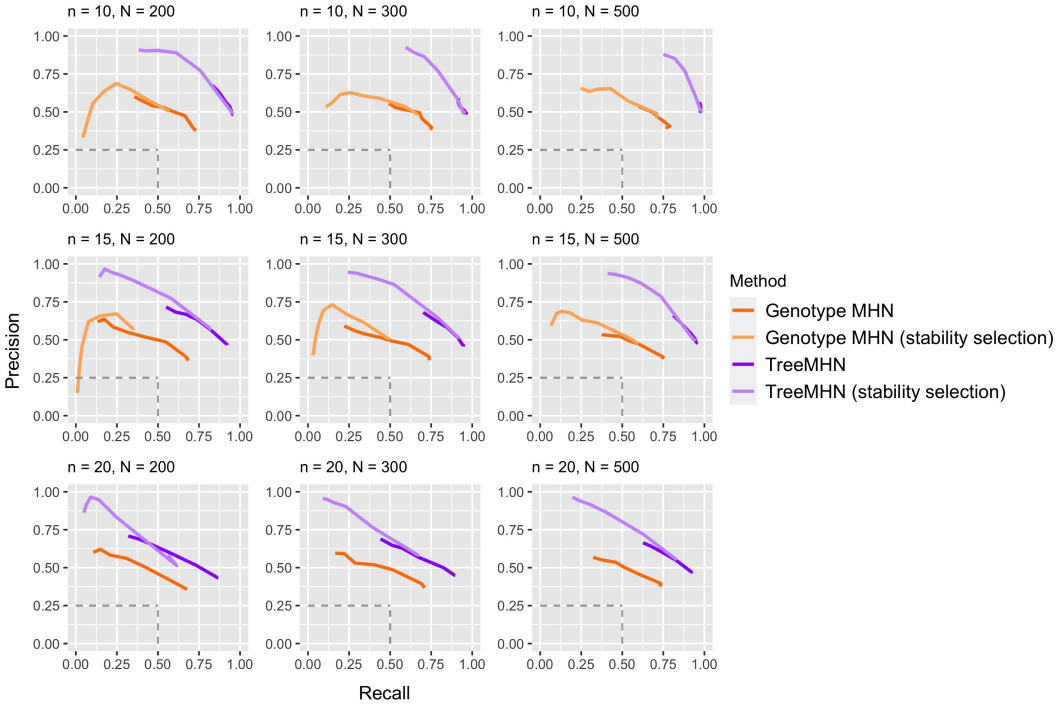


Figure 2: Performance of TreeMHN and the genotype MHN method (with and without stability selection) in estimating the true network Θ on simulated data for $n \in \{10, 15, 20\}$ mutations and $N \in \{200, 300, 500\}$ samples. The precision and recall curves averaged over 100 simulation runs are plotted 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.

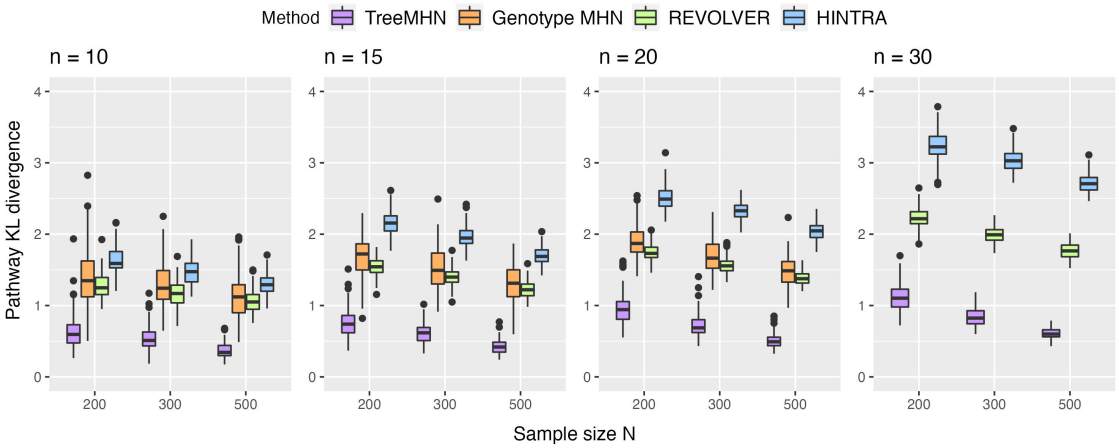


Figure 3: Performance of TreeMHN in estimating the probability distribution of mutational pathways compared to genotype MHNs, REVOLVER, and HINTRA (Methods). We display the Kullback-Leibler divergence between the estimated distributions and the ground truth distributions for $n \in \{10, 15, 20, 30\}$ mutations and $N \in \{200, 300, 500\}$ samples over 100 simulation runs.

and mutation network size decreases. At the same regularization level, the solutions with stability selection 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 2, Supplementary Figure S4). The reason is that interactions between rare mutations can hardly be observed in the trees, leading to a lack of statistical power to correctly estimate the corresponding entries in Θ. This limitation is in general method-independent.

In all cases, TreeMHN clearly and consistently outperforms the genotype 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 [32], 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 are added to the trees (Supplementary Figures S5 & S6), as well as for varying proportions of zero entries (Supplementary Figure S7) and negative entries in Θ (Supplementary Figure S8). The difference between the networks estimated using MLE and MC-EM is small and decreases with increasing number of Monte Carlo samples (Supplementary Figure S9).

Next, we assess the performance of TreeMHN in estimating pathway probabilities (Figure 3). Across different sample and network sizes, TreeMHN outperforms all alternatives. HINTRA has the worst performance, possibly because of over-parameterization given the limited number of observations for every possible ancestry set. REVOLVER has similar performance as genotype MHNs even though it does not explicitly model co-occurrence and exclusivity between mutations. The genotype MHN method is unable to handle the case $n = 30$ due to exponentially increasing space and time complexity [30]. For TreeMHN, the runtime is instead limited by the tree with the maximum number of subtrees (Supplementary Figure S10).

Application to acute myeloid leukemia data. We apply TreeMHN to the cohort of $N = 123$ AML patient samples analyzed in [21] by high-throughput single-cell panel sequencing, which involves 543 somatic mutations in $n = 31$ cancer-associated genes. The mutation trees reconstructed by SCITE [33], a single-cell phylogenetic method, represent the complete evolutionary histories of the tumors. Assuming all point mutations in the same gene have similar 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* [21] (Supplementary Section E.1). This recurrence increases 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 over the 15 mutated genes which have non-zero off-diagonal entries (Figure 4 and Supplementary Figure S11). The sparseness

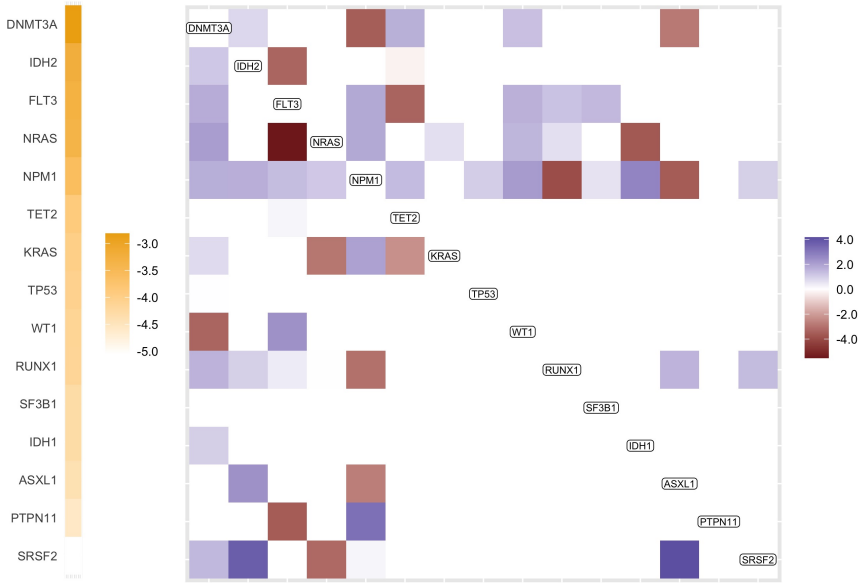


Figure 4: Partial Mutual Hazard Network for the 123 AML patient samples. The full network is shown in Supplementary Figure S11, and only mutations with non-zero off-diagonal entries are shown here. The columns and rows of the matrix are ordered by decreasing baseline rates, which are shown on the left in yellow scale. The empty off-diagonal entries represent the case of no effect ($\theta_{ij} = 0$), meaning that there is no edge from mutation j to mutation i. The red blocks correspond to inhibiting effects ($\theta_{ij} < 0$, $j \dashv i$) and the blue promoting effects ($\theta_{ij} > 0$, $j \rightarrow i$). The darker the colors, the stronger the effects.

of the network 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 may not be 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 [34]. However, these genes do not necessarily appear in every mutation tree or are the initiating events, since the appearance of a mutation depends not only on the baseline rate but also on 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 [22] 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 reveals 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 accumulate another mutation in *NRAS*, which may still occur in separately-evolved subclones. On the contrary, subclones that acquire mutations in *NRAS* first are still relatively likely to hit *FLT3* subsequently. Apart from with *NRAS*, such a relationship also appears with *PTPN11* and *IDH2*. 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 cell populations prior to treatment [35, 36].

Next, we infer the most probable mutational pathways from the estimated network (Methods & Figure 5). In comparison to alternative methods, the probabilities estimated by TreeMHN match closer in rankings with the relative frequencies of the observed AML trajectories (Supplementary Figures S12 & S13). Also, mutations in *DNMT3A*, *NPM1*, and *FLT3* often co-occur with a relatively high probability. This three-way interaction is found to be associated with poor prognosis [37, 38]. 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 [39, 40].

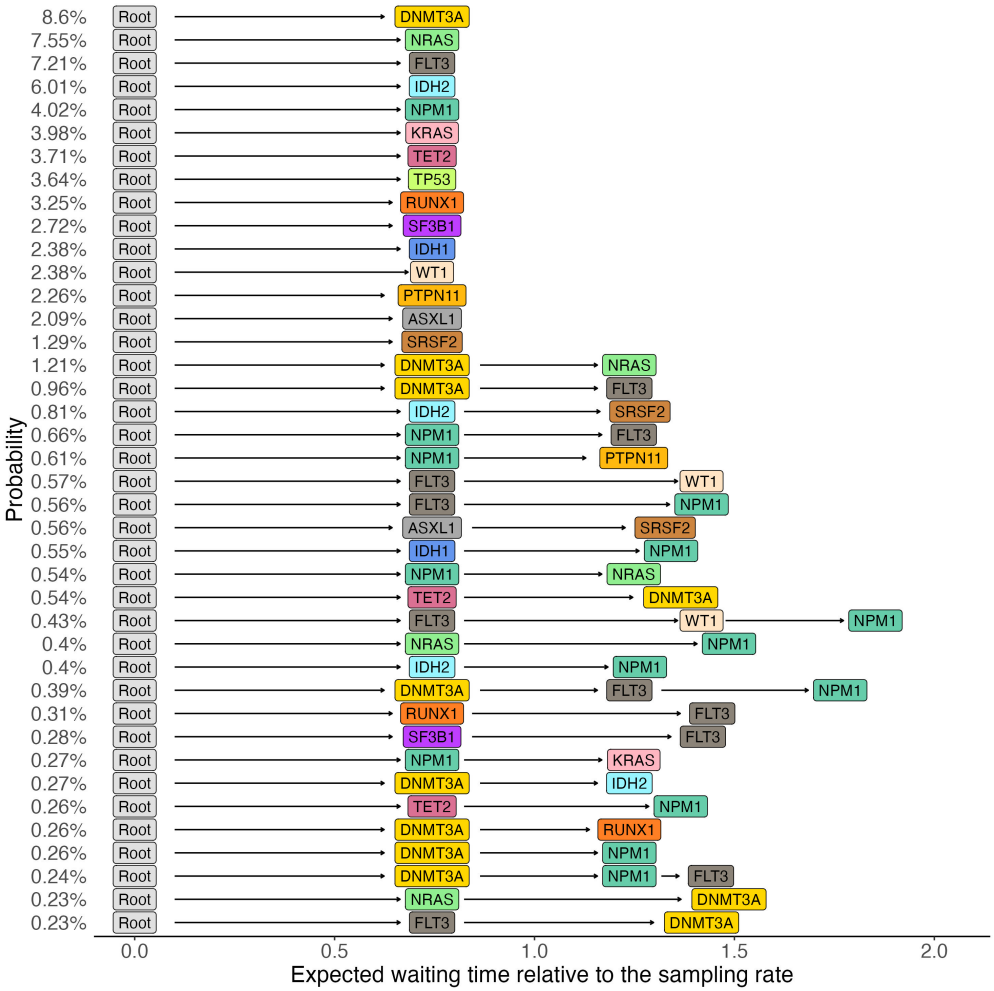
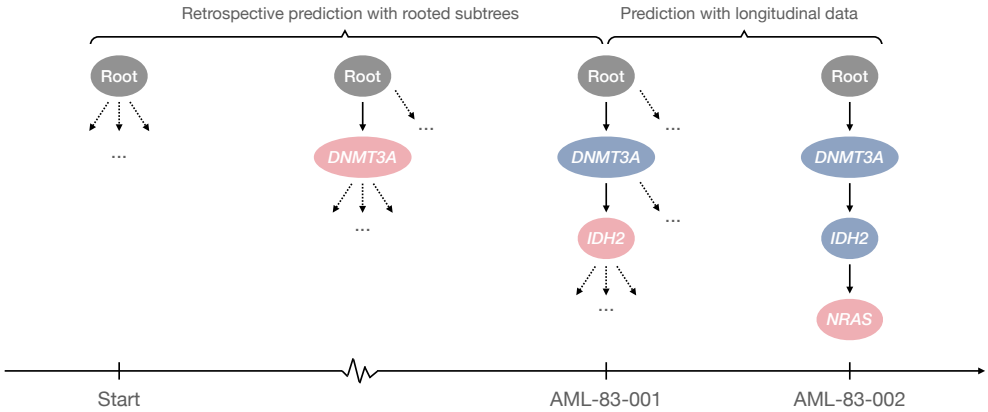
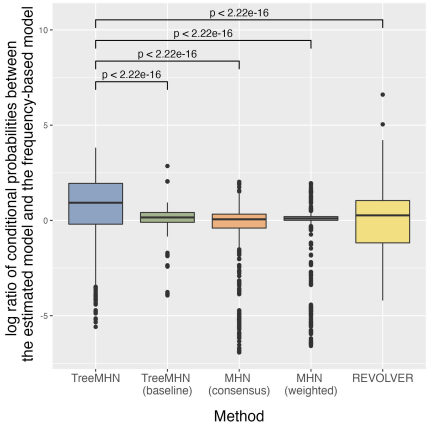


Figure 5: Top 40 most probable mutational pathways inferred from the partial Mutual Hazard Network of the AML dataset (Figure 4). Each row represents a mutational pathway, labeled and ordered by pathway probabilities, and the horizontal positions of the mutations correspond to their expected waiting times relative to the sampling rate of $\lambda_s = 1$ (Methods).

Moreover, conditioned on the estimated network and a given tumor tree, we can predict the next most probable mutational event (Methods & Supplementary Section C.2). For the majority of the 123 primary tumor trees, TreeMHN can reliably capture the events downstream of any of the rooted subtrees with high probabilities (Figure 6a, 6b). To further validate the results, we analyze the longitudinal samples from 15 patients in the same dataset. We consider only those tree pairs on consecutive time points, where the second tree has mutations not present in the first tree (Figure 6a). Under TreeMHN, seven out of nine new events have higher predicted probabilities than over 95% of the other possible mutational events, whereas alternative methods give varying predictions which are worse on average than TreeMHN (Figure 6c). For instance, given the tree of patient sample AML-38-002, TreeMHN successfully identifies the two relapse events (Root→*NPM1*→*IDH2*→*PTPN11* and Root→*NPM1*→*IDH1*→*FLT3* in AML-38-003) in the top 5% most probable mutational events. With treatment effects drastically altering the fitness landscape [41], it is possible that subclones that were too small to detect at diagnosis became more abundant in a relapse. The consistent predictions highlight the ability of TreeMHN to unravel the interplay among the mutational events from the heterogeneous mutation trees, offering opportunities for evolution-guided treatment strategies.



(a)



(b)

From	To	New Event	Estimated Percentile Ranks					
			TreeMHN	TreeMHN (baseline)	MHN (consensus)	MHN (weighted)	REVOLVER	Frequency
AML-38-002	AML-38-003	Root -> *NPM1* -> *IDH2* -> *PTPN11*	**98.4**	72.8	70.8	74.0	84.0	60.8
AML-63-003	AML-63-004	Root -> *IDH2* -> *NPM1*	98.3	**100**	94.9	96.6	98.3	96.6
AML-83-001	AML-83-002	Root -> *DNMT3A* -> *IDH2* -> *NRAS*	**97.7**	93.1	86.2	94.3	95.4	92.0
AML-99-003	AML-99-004	Root -> *DNMT3A* -> *IDH2* -> *RUNX1* -> *FLT3*	**97.6**	96.4	75.4	92.2	95.8	95.8
AML-38-002	AML-38-003	Root -> *NPM1* -> *IDH1* -> *FLT3*	96.4	**100**	97.6	92.8	78.4	93.6
AML-09-001	AML-09-002	Root -> *NPM1* -> *FLT3* -> *WT1*	95.7	79.1	82.6	82.6	**100**	75.7
AML-99-003	AML-99-004	Root -> *DNMT3A* -> *IDH2* -> *NRAS*	95.2	92.8	91.6	**95.8**	94.6	90.4
AML-04-001	AML-04-002	Root -> *SF3B1* -> *SRSF2* -> *NRAS* -> *WT1*	74.4	77.2	**79.5**	77.2	71.2	72.6
AML-04-001	AML-04-002	Root -> *SF3B1* -> *SRSF2* -> *IDH1*	71.2	68.0	76.7	70.8	**84.9**	64.8
		Average Percentile Ranks	**91.7**	86.6	83.9	86.3	89.2	82.5

(c)

Figure 6: Predictions with AML trees. (a) A schematic representation of evaluating the predictions of next events given a tree structure. In this example, AML-83-001 is a tumor sample, and the associated mutation tree is a chain with two mutations. Given this tree, we can retrospectively enumerate all its rooted subtrees and compute the probability of the existing downstream events given an estimated network. Since a consecutive sample AML-83-002 is available for the same patient, we can also evaluate how well TreeMHN can predict the next events after treatment. The root node is in grey, the events that are common from a preceding tree are in blue, and the new events are in pink. (b) Performance assessment on retrospective predictions. For the 123 primary tumor trees, we enumerate all combinations of rooted subtrees and their corresponding downstream events. Then, for each combination, we compute the log ratio of the conditional probabilities between the estimated TreeMHN model and a frequency-based model that predicts the next events using only the relative frequencies of the mutations in the cohort. We include four other methods for comparison: TreeMHN with only the baseline rates, genotype MHN using consensus genotypes, genotype MHN using weighted genotypes, and REVOLVER (Supplementary Section C.2). The p-values for the one-sided Wilcoxon signed-rank tests between TreeMHN and alternative methods are also displayed. (c) Table of estimated percentile ranks for the new events in consecutive tree pairs in the longitudinal samples. An event having a percentile rank of 90 has a higher probability than 90% of the other possible mutational events. The highest percentile rank for each row is in bold.

Application to non-small-cell lung cancer data. Next, we analyze the TRACERx NSCLC multi-region whole-exome sequencing data for $N = 99$ patients [15]. By applying TreeMHN to the phylogenetic trees with $n = 79$ driver mutations from [7] (Supplementary Section F.1), we detect stable signals of interdependencies among 18 mutations (Figure 7 & Supplementary Figures S14). The majority of these interactions are from *KRAS*, *TP53*, and *EGFR* to other genes, suggesting their essential roles as recurrent initial events in tumorigenesis and progression [6, 15]. In particular, the exclusive relationship between the oncogenic drivers

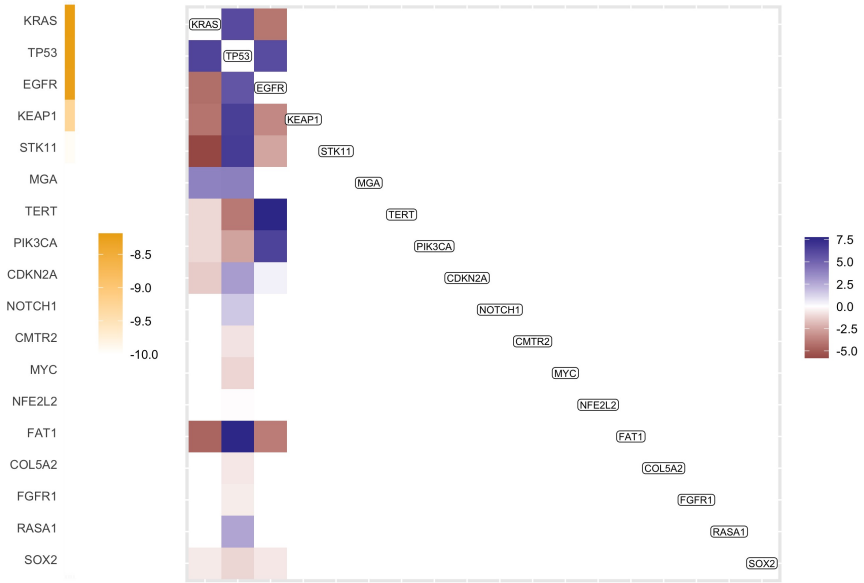


Figure 7: Partial Mutual Hazard Network for the 99 TRACERx NSCLC patient samples. The full network is shown in Supplementary Figure S14, and here only mutations with non-zero off-diagonal entries are shown. The columns and rows of the matrix are ordered by decreasing baseline rates.

KRAS and *EGFR* aligns with previous studies and could be associated with different clinical features (*e.g.* smoking exposure) between *KRAS*-mutant and *EGFR*-mutant subgroups [42, 43]. Each of these two mutations is co-occuring with *TP53*, one of the most commonly mutated tumor suppressor genes in NSCLC [44]. *EGFR*-mutant tumors with co-occuring *TP53* were found to have higher degrees of genomic instability and shorter progression-free survival after EGFR TKI therapy [45]. Co-mutations in *KRAS*/*TP53*, on the other hand, are predictive for favorable clinical response to PD-1 inhibitors [46]. Moreover, existing *EGFR* mutations may encourage the occurrence of mutations in *TERT* and *PIK3CA*, where the former can influence telomere maintenance mechanisms in tumor cells [47], and the latter promotes cellular invasion and migration in vitro [43]. Both are associated with poor overall survival [43, 47].

Without reconciling the trees or clustering the trees into subgroups, the most probable mutational pathways inferred by TreeMHN capture most of the repeated trajectories found by REVOLVER (Supplementary Figure S15). Notably, we recover 7 of the 10 REVOLVER clusters (C2, C3, C4, C6, C7, C9, C10) in the top 50 pathways (Supplementary Figure S16), underlining the ability of TreeMHN to disentangle noisy signals. For cluster C7, TreeMHN assigns a higher probability to the trajectory *KRAS*→*TP53*→*MGA* as compared to *TP53*→*KRAS*→*MGA*, whereas REVOLVER cannot tell them apart. Interestingly, the trajectory *CDKN2A*→*TP53* in cluster C5, identified as highly robust by REVOLVER, has a lower probability than *TP53*→*CDKN2A*, where the ordering is reversed. As pointed out in [15], most of the mutations in *TP53* were clonal for both adenocarcinoma and squamous-cell carcinoma in the TRACERx dataset, whereas mutations in *CDKN2A* appeared mostly in the latter subtype and often late. Hence, it is more likely that mutations in *TP53* precede those in *CDKN2A*, hinting at how they cooperate on cell-cycle deregulation.

Application to breast cancer data. Finally, TreeMHN is not restricted to tumor trees reconstructed from single-cell or multi-region sequencing data, and we analyze a bulk sequencing dataset of 1918 tumors for 1756 advanced breast cancer patients, where clinical data is also available [31]. Based on hormone receptor and HER2 status (HR+/HER2+, HR+/HER2-, HR-/HER2+, Triple Negative) and sample type (treatment-free primary vs. metastasis), we can segregate the patients into eight subgroups. Considering only copy-neutral autosomal regions, we restrict the analysis to the union of SNVs that appear in at least 10% of the patients in each subgroup. In total we have $n = 19$ mutations and $N = 1152$ patients with 1232 phylogenetic trees inferred by SPRUCE [48], a phylogenetic method based on bulk mixture deconvolution (Supplementary Section G.1).

The estimated network on all trees captures combined signals, which are mainly driven by the largest subgroups with HR+/HER2- status (Supplementary Figure S17). We observe that mutations in *TP53* and *PIK3CA*, which were identified as early clonal events [6, 8, 31], have the highest baseline rates of mutations and are co-occurring. We also detect exclusivity between *PIK3CA* and *PIK3R1*, suggesting that mutations in one of these drivers may suffice to trigger abnormal regulation of PI3K activation in breast cancer [49]. While there are no large deviations from the combined network, we obtain some subgroup-specific interdependencies by applying TreeMHN separately to each subgroup (Supplementary Figures S18 & S19). For example, *NF1* mutations, which are found to be associated with endocrine resistance [31], have higher baseline rates in hormone receptor-negative tumors and co-occur with *TP53* more frequently in metastatic samples. Also, we detect co-occurrence of *PIK3CA* and *GATA3* in hormone receptor-positive tumors but not in the negative ones. It has been observed that tumors with *GATA3* mutations depend more on estrogen level and may be predictive for positive response to aromatase inhibitors [50]. The subgroup-level most probable mutational pathways not only confirm existing findings reported by HINTRA [8] and RECAP [9], such as *CDH1*→*PIK3CA* for the HR+/HER2- samples, but also provide new insights (Supplementary Figures S20-S23). For instance, the combination of *TP53*, *PTEN*, and *PIK3CA*/*PIK3R1* ranks higher in the metastatic triple-negative subgroup than in other subgroups. Cooperatively, these mutations can lead to hyperactivation of the PI3K-AKT pathway and uncontrolled proliferation of cells, ultimately reducing the overall survival rate [51].

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. Unlike Mutual Hazard Networks [30], TreeMHN can take as input heterogeneous tree structures estimated from multi-region, single-cell, or bulk sequencing data, 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, typically much larger, number of mutations, that limits the computation time of TreeMHN.

Through simulation studies, we have demonstrated the superior performance of TreeMHN in estimating both patterns of clonal exclusivity or co-occurrence and pathway probabilities in comparison with MHNs, REVOLVER, and HINTRA. Moreover, we have shown that TreeMHN is robust against uncertainty in the phylogenetic trees for varying noise levels. Alternatively, one may handle such uncertainty by sampling the trees for each patient proportionally to their posterior probabilities and taking them as weighted inputs to TreeMHN. 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 data, where the sampling time is either provided or can be inferred from data. Including observed sampling times is technically a straightforward task [52], but such data is often difficult to obtain without having treatment interventions. Modelling drug response data is a crucial but challenging direction to explore, for which TreeMHN may serve as a basis.

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 combine different phylogenies from various sources (*e.g.* [33, 53, 54]) to take into account different modelling assumptions. 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 [21], while existing alternatives require the infinite sites assumption. Like all other progression models, however, TreeMHN currently does not consider back mutations, *i.e.* situations in which a mutation is acquired at first but subsequently lost [55]. A possible extension along this line is to include additional parameters and use as input phylogenetic trees inferred by methods such as SCARLET [56], 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 [57], mixed types of mutations [58], or even functional pathways [59].

Nevertheless, TreeMHN does not take into account the subclone sizes, which can be viewed as consequences of clonal selection [60]. As noted by [61], inferring mutation rates and fitness values jointly may have promise but is challenging. On the one hand, larger subclone sizes can be attributed to both their earlier appearance or higher fitness [61]. On the other hand, the mutation rates in cancer progression models are rates of evolution, which implicitly involve subclone fitness. Thus, any attempts to modify TreeMHN to model clonal selection need to be taken with caution. In this work we assumed all trees in the same cohort are generated from one MHN, but it is possible to extend TreeMHN to perform model-based unsupervised clustering akin to [62].

We have applied TreeMHN to data of three different cancer types and demonstrated its practicality and flexibility. Beyond reproducing existing results, TreeMHN provides additional insights, including baseline rates of mutational events, directional strengths of the promoting or inhibiting effects, as well as probabilistic measures of mutational pathways. In particular, we find good overlap between the longitudinal AML data [21] and our predictions of the next mutational events. The application and results of TreeMHN may therefore be useful for prioritizing personalized treatments.

Methods

TreeMHN algorithm. TreeMHN is a tree-generating process. Unlike existing CPMs, including genotype MHNs [30], 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 $\Theta \in \mathbb{R}^{n \times n}$. 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 $0 \leq d \leq n$, 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 8 & Supplementary Section A.1):

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

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 [30]. If both $\theta_{ij} < 0$ and $\theta_{ji} < 0$ (i.e., $j \vdash\dashv i$), then the two mutations are called clonally exclusive, and if both signs are positive (i.e., $j \leftrightarrow i$), 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 S with time $T_s \sim \mathrm{Exp}(\lambda_s)$. Typically, we assume $\lambda_s = 1$ for identifiability [29]. 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 towards the events right after sampling (Supplementary Figure S1(b)). 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

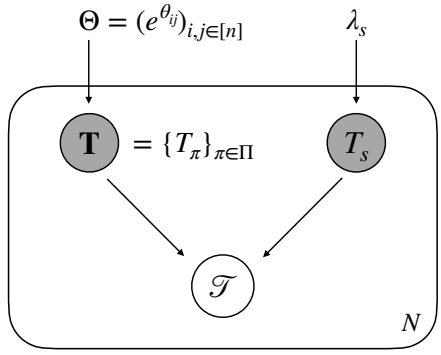


Figure 8: TreeMHN as a probabilistic graphical model. The waiting times of subclones $\mathbf{T} = \{T_\pi\}_{\pi\in\Pi}$ are exponentially distributed random variables parameterized by an MHN $\Theta = (e^{\theta_{ij}})_{i,j\in[n]}$. The tree structure $\mathscr{T}$ is jointly determined by $\mathbf{T}$ and an independent sampling time 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.

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 [55]. Even assuming the ISA for individual genomic bases, when summarising at the gene level the same gene may be affected in parallel lineages. In Results, we demonstrated the applicability of TreeMHN using the AML 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. Therefore, TreeMHN can model extensive intra-tumor and inter-tumor heterogeneity, while capturing common, re-occurring features.

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 $p(\mathscr{T} \mid \Theta)$ and its 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.2). Given N mutation trees $\boldsymbol{\mathscr{T}} = \{\mathscr{T}^1, \dots, \mathscr{T}^N\}$, we follow [30] 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.

When some or all of the trees in $\boldsymbol{\mathscr{T}}$ have many subtrees (*e.g.* more than 500 subtrees), the MLE procedure 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 as follows (see also Supplementary Section A.3). 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

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

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

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

For large trees, we approximate the expectation by drawing M samples from the following proposal distribution. First, the sampling time $T_s \sim \text{Exp}(\lambda_s)$ 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

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

where $\text{TExp}(\lambda, a, b)$ 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

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

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

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

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 [63]. 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 Supplementary Section B).

Probability and expected waiting time of a pathway. We consider the set of mutational pathways that end with the sampling event S,

$$\Pi_S := \{(0, \sigma_1, \ldots, \sigma_d, S) \mid \sigma_i \in [n] \text{ non-duplicated}, 0 \le d \le n\}. \tag{9}$$

Given Θ and λ_s, we can compute the probability of a pathway $\pi \in \Pi_S$ as the product of competing exponentials,

$$P_\Theta(\pi) = \prod_{i=1}^{d} \frac{\lambda_{(\pi_{i-1},\sigma_i)}}{\lambda_s + \sum_{j\in[n]\setminus\pi_{i-1}} \lambda_{(\pi_{i-1},j)}} \times \frac{\lambda_s}{\lambda_s + \sum_{j\in[n]\setminus\pi} \lambda_{(\pi,j)}}, \tag{10}$$

where $\pi_i = (0, \sigma_1, \ldots, \sigma_i) \subset \pi$. It follows that $(P_\Theta(\pi))_{\pi\in\Pi_S}$ is the probability distribution over Π_S with respect to Θ, since $\sum_{\pi\in\Pi_S} P_\Theta(\pi) = 1$. Likewise, the expected waiting time of $\pi \in \Pi_S$ is the sum of the expected time interval lengths along the pathway,

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

However, computing $P_\Theta(\pi)$ for all $\pi \in \Pi_S$ is computationally infeasible for large n, since the number of pathways in Π_S is factorial in n. Instead, we can enumerate the most probable pathways that have at least one mutation with dynamic programming following the pseudocode in Supplementary Algorithm 2, the complexity of which is linear in n.

To compare with alternative methods that do not contain a sampling event, such as REVOLVER [7] and HINTRA [8], we can use another formulation with $\Pi_d \subseteq \Pi$, which is the set of mutational pathways of a fixed length d for $1 \le d \le n$. The probability of a pathway $\pi \in \Pi_d$ can be computed similarly as

$$P_\Theta(\pi) = \prod_{i=1}^{d} \frac{\lambda_{(\pi_{i-1},\sigma_i)}}{\sum_{j\in[n]\setminus\pi_{i-1}} \lambda_{(\pi_{i-1},j)}}, \tag{12}$$

and $\sum_{\pi \in \Pi_d} P_\Theta(\pi) = 1$. Note that in this formulation, pathways of different lengths are not directly comparable, but it is still useful as a tool to validate an estimated pathway probability distribution. Suppose Q is another probability distribution over Π_d, then the Kullback-Leibler (KL) divergence from Q to P_Θ,

$$D_{\mathrm{KL}}(P_\Theta \parallel Q) = \sum_{\pi \in \Pi_d} P_\Theta(\pi) \log \frac{P_\Theta(\pi)}{Q(\pi)} \tag{13}$$

measures the distance between the two distributions. In particular $D_{\mathrm{KL}}(P_\Theta \parallel Q) = 0$ if and only if $P_\Theta = Q$.

Probability of a downstream event given a tree. Given a tree structure $\mathscr{T}$ and an estimated network Θ, we can compute the probabilities of the next mutational events. The events that could happen next are all events in the augmented tree $A(\mathscr{T})$ but not in $\mathscr{T}$. By competing exponentials, we calculate the probability of an event π happens before all the other events in $A(\mathscr{T}) \setminus \mathscr{T}$ as

$$p(\pi \mid \mathscr{T}, \Theta) = \frac{\lambda_\pi}{\sum_{\pi' \in A(\mathscr{T}) \setminus \mathscr{T}} \lambda_{\pi'}} \times \mathbb{1}\{\pi \in A(\mathscr{T}) \setminus \mathscr{T}\}, \tag{14}$$

where λ_π is the rate associated with event π (Eq. (1)). Then, we can rank all events $\pi \in A(\mathscr{T}) \setminus \mathscr{T}$ by their probabilities. We describe alternative approaches in Supplementary Section C.2.

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 [30], 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 mimick the AML dataset. Given Θ, we then generate N mutation trees (see also Supplementary Section A.4), 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 ($j \quad i$, $j \to i$, $j \dashv i$) 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 D.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 (Eq. (12) and Eq. (13)). Note that longer pathway lengths can dramatically increase the computational complexity as the number of pathways in Π_d is $\frac{n!}{(n-d)!}$.

For the first metric, we compare TreeMHN against the genotype MHN method, the only previously published cancer progression model that can estimate Θ. However, the genotype MHN method cannot handle tree structures, since the input is one consensus genotype per tumor. As such, 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 the genotype MHN method 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 and HINTRA 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 D.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 D.3). For each configuration, we repeat 100 times the process of simulating the data and estimating the parameters over a set of regularization parameters $\gamma \in \{0.05, 0.1, 0.5, 1, 1.5, 2, 2.5, 3\}$. For TreeMHN, the choice of parameter estimation method is automatically determined by a pre-specified number of subtrees: if the maximum

number of subtrees of all trees in the set is below this threshold, then MLE is used. Otherwise, we use the hybrid MC-EM method instead. By default, the threshold is set to be 500 subtrees, and the number of Monte Carlo samples is $M = 300$.

Acknowledgements

The authors would like to thank Katharina Jahn for providing the phylogenetic trees from [21]. This work was supported by ETH Zurich - Open ETH project SKINTEGRITY.CH.

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

1. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194,** 23–28 (1976).
2. McGranahan, N. & Swanton, C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* **168,** 613–628 (2017).
3. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology* **15,** 81–94 (2018).
4. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339,** 1546–1558 (2013).
5. Ortmann, C. A. *et al.* Effect of mutation order on myeloproliferative neoplasms. *New England Journal of Medicine* **372,** 1865–1866 (2015).
6. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578,** 122–128 (2020).
7. Caravagna, G. *et al.* Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nature Methods* **15,** 707–714 (2018).
8. Khakabimamaghani, S. *et al.* Collaborative intra-tumor heterogeneity detection. *Bioinformatics* **35,** i379–i388 (2019).
9. Christensen, S., Kim, J., Chia, N., Koyejo, O. & El-Kebir, M. Detecting evolutionary patterns of cancers using consensus trees. *Bioinformatics* **36,** i684–i691 (2020).
10. Hodzic, E. *et al.* Identification of conserved evolutionary trajectories in tumors. *Bioinformatics* **36,** i427–i435 (2020).
11. Lipinski, K. A. *et al.* Cancer evolution and the limits of predictability in precision cancer medicine. *Trends in Cancer* **2,** 49–63 (2016).
12. Lässig, M., Mustonen, V. & Walczak, A. M. Predicting evolution. *Nature Ecology and Evolution* **1,** 0077 (2017).
13. Hosseini, S.-R., Diaz-Uriarte, R., Markowetz, F. & Beerenwinkel, N. Estimating the predictability of cancer evolution. *Bioinformatics* **35,** i389–i397 (2019).

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

15. Jamal-Hanjani, M. *et al.* Tracking the evolution of non–small-cell lung cancer. *New England Journal of Medicine* **376,** 2109–2121 (2017).

16. Turajlic, S. *et al.* Tracking cancer evolution reveals constrained routes to metastases: TRACERx renal. *Cell* **173,** 581–594 (2018).

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

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

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

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

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

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

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

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

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

26. Diaz-Colunga, J. & Diaz-Uriarte, R. Conditional prediction of consecutive tumor evolution using cancer progression models: What genotype comes next? *PLoS Computational Biology* **17,** e1009055 (2021).

27. Raphael, B. J. & Vandin, F. Simultaneous inference of cancer pathways and tumor progression from cross-sectional mutation data. *Journal of Computational Biology* **22,** 510–527 (2015).

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

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

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

31. Razavi, P. *et al.* The genomic landscape of endocrine-resistant advanced breast cancers. *Cancer Cell* **34,** 427–438 (2018).

32. Gotovos, A., Burkholz, R., Quackenbush, J. & Jegelka, S. Scaling up continuous-time Markov chains helps resolve underspecification. *Advances in Neural Information Processing Systems* **34** (2021).

33. Jahn, K., Kuipers, J. & Beerenwinkel, N. Tree inference for single-cell data. *Genome Biology* **17,** 86 (2016).

34. Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New England Journal of Medicine* **368,** 2059–2074 (2013).

35. McMahon, C. M. *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).

36. Kennedy, V. E. & Smith, C. C. FLT3 mutations in acute myeloid leukemia: key concepts and emerging controversies. *Frontiers in Oncology* **10,** 2927 (2020).

37. Loghavi, S. *et al.* Clinical features of de novo acute myeloid leukemia with concurrent DNMT3A, FLT3 and NPM1 mutations. *Journal of Hematology & Oncology* **7,** 74 (2014).

38. Bezerra, M. F. *et al.* Co-occurrence of DNMT3A, NPM1, FLT3 mutations identifies a subset of acute myeloid leukemia with adverse prognosis. *Blood* **135,** 870–875 (2020).

39. Shlush, L. I. *et al.* Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506,** 328–333 (2014).

40. Papaemmanuil, E. *et al.* Genomic classification and prognosis in acute myeloid leukemia. *New England Journal of Medicine* **374,** 2209–2221 (2016).

41. Bozic, I. Quantification of the selective advantage of driver mutations is dependent on the underlying model and stage of tumor evolution. *Cancer Research* **82,** 21–24 (2022).

42. Román, M. *et al.* KRAS oncogene in non-small cell lung cancer: clinical perspectives on the treatment of an old target. *Molecular Cancer* **17,** 33 (2018).

43. Skoulidis, F. & Heymach, J. V. Co-occurring genomic alterations in non-small-cell lung cancer biology and therapy. *Nature Reviews Cancer* **19,** 495–509 (2019).

44. Herbst, R. S., Morgensztern, D. & Boshoff, C. The biology and management of non-small cell lung cancer. *Nature* **553,** 446–454 (2018).

45. Labbé, C. *et al.* Prognostic and predictive effects of TP53 co-mutation in patients with EGFR-mutated non-small cell lung cancer (NSCLC). *Lung Cancer* **111,** 23–29 (2017).

46. Dong, Z.-Y. *et al.* Potential predictive value of TP53 and KRAS mutation status for response to PD-1 blockade immunotherapy in lung adenocarcinoma. *Clinical Cancer Research* **23,** 3012–3024 (2017).

47. Bordeira Gaspar, T. *et al.* Telomere maintenance mechanisms in cancer. *Genes* **9,** 241 (2018).

48. El-Kebir, M., Satas, G., Oesper, L. & Raphael, B. J. Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. *Cell Systems* **3,** 43–53 (2016).

49. Chen, L. *et al.* Characterization of PIK3CA and PIK3R1 somatic mutations in Chinese breast cancer patients. *Nature Communications* **9,** 1357 (2018).

50. Ma, C. X., Reinert, T., Chmielewska, I. & Ellis, M. J. Mechanisms of aromatase inhibitor resistance. *Nature Reviews Cancer* **15,** 261–275 (2015).

51. Chai, C., Wu, H. H., Abuetabh, Y., Sergi, C. & Leng, R. Regulation of the tumor suppressor PTEN in triple-negative breast cancer. *Cancer Letters* **527,** 41–48 (2022).

52. Montazeri, H. *et al.* Large-scale inference of Conjunctive Bayesian Networks. *Bioinformatics* **32,** 727–735 (2016).

53. Singer, J., Kuipers, J., Jahn, K. & Beerenwinkel, N. Single-cell mutation identification via phylogenetic inference. *Nature Communications* **9,** 5144 (2018).

54. Zafar, H., Tzen, A., Navin, N., Chen, K. & Nakhleh, L. SiFit: Inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biology* **18,** 178 (2017).

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

56. Satas, G., Zaccaria, S., Mon, G. & Raphael, B. J. SCARLET: single-cell tumor phylogeny inference with copy-number constrained mutation losses. *Cell Systems* **10,** 323–332 (2020).

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

58. Chen, Z., Gong, F., Wan, L. & Ma, L. BiTSC^2: Bayesian inference of tumor clonal tree by joint analysis of single-cell SNV and CNA data. *Briefings in Bioinformatics* **23,** bbac092 (2022).

59. Moore, A. L. *et al.* Spatial distribution of private gene mutations in clear cell renal cell carcinoma. *Cancers* **13,** 2163 (2021).

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

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

62. Kuipers, J. *et al.* Mutational interactions define novel cancer subgroups. *Nature Communications* **9,** 4353 (2018).

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