Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Discovering Significant Evolutionary Trajectories in Cancer Phylogenies

Leonardo Pellegrina [1] and Fabio Vandin [1,*]

[1] Department of Information Engineering, University of Padova, Padova, 35129, Italy.

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Tumors are the result of a somatic evolutionary process leading to substantial intra-tumor heterogeneity. Single-cell and multi-region sequencing enable the detailed characterization of the clonal architecture of tumors, and have highlighted its extensive diversity across tumors. While several computational methods have been developed to characterize the clonal composition and the evolutionary history of tumors, the identification of significantly conserved evolutionary trajectories across tumors is still a major challenge.

Results: We present a new algorithm, MASTRO, to discover significantly conserved evolutionary trajectories in cancer. MASTRO discovers all conserved trajectories in a collection of phylogenetic trees describing the evolution of a cohort of tumors, allowing the discovery of conserved complex relations between alterations. MASTRO assesses the significance of the trajectories using a conditional statistical test that captures the coherence in the order in which alterations are observed in different tumors. We apply MASTRO to data from non-small-cell lung cancer bulk sequencing and to acute myeloid leukemia data from single-cell panel sequencing, and find significant evolutionary trajectories recapitulating and extending the results reported in the original studies.

Availability: MASTRO is available at `https://github.com/VandinLab/MASTRO`

Contact: fabio.vandin@unipd.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Cancer is the result of the accumulation of somatic alterations conferring selective advantage to cells (Nowell, 1976). The accumulation of such alterations is an evolutionary process, with subpopulations, or *clones*, of tumor cells having distinct genomic alterations that arise as a tumor grows. Such subpopulations are shaped by the processes of clonal expansion and selection, which lead to a substantial intra-tumor heterogeneity, arguably one of the main challenges in cancer treatment.

Recent advances in multi-region sequencing (Gerlinger *et al.*, 2012; Yates *et al.*, 2015; Turajlic *et al.*, 2018) and single-cell sequencing (Navin, 2014; Lawson *et al.*, 2018) have enabled the collection of data providing a more precise characterization of the clonal architecture of tumors. Such data has shown that, while there is an inherent stochastic component in tumor evolution, there are some features that are shared by the progression of certain tumors, such as some constraints in the order with which alterations arise (Ortmann *et al.*, 2015; Kent and Green, 2017; Levine *et al.*, 2019). The detection of such shared features is crucial for the development of effective therapeutic interventions (Lipinski *et al.*, 2016; Hosseini *et al.*, 2019; Diaz-Uriarte and Vasallo, 2019). However, intra-tumor heterogeneity makes it more challenging to reliably identify shared features from the complicated genomic landscapes resulting by the presence of several cell subpopulations with distinct genomic alterations.

A number of computational methods have been recently proposed to infer the subclonal composition or the evolutionary history of tumors from tumor sequencing data (El-Kebir *et al.*, 2015; Deshwar *et al.*, 2015; Popic *et al.*, 2015; Malikic *et al.*, 2015; El-Kebir *et al.*, 2016; Jahn *et al.*, 2016; Ross and Markowetz, 2016; Zaccaria *et al.*, 2018; Eaton *et al.*, 2018; Govek *et al.*, 2018; Malikic *et al.*, 2019a,b; Zafar *et al.*, 2019). These methods typically produce in output a phylogenetic tree (Schwartz and Schäffer, 2017) that represents (one of) the inferred order of the observed genomic alterations. The identification of *recurrent trajectories* from such trees is still challenging, due to the inter-tumor heterogeneity of genomic

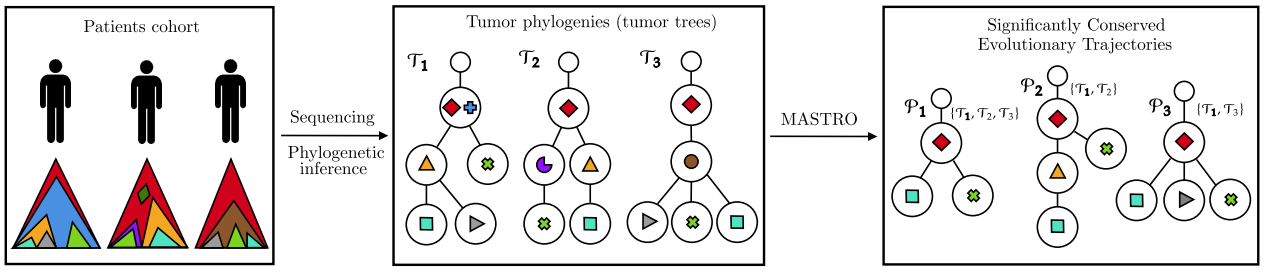

Fig. 1. High level description of MASTRO. Leveraging sequencing data from a cohort of patients and phylogenetic inference algorithms, it is possible to infer the tumor trees describing the clonal evolution of the tumors as the accumulation of different alterations (different alterations corresponds to different colored shapes within the nodes of the trees). MASTRO identifies conserved evolutionary trajectories, describing complex interactions among alterations, that are frequently observed in the tumor trees, . MASTRO assesses their statistical significance and provides sound control of false discoveries. In the example, we show three trajectories ($\mathcal{P}_1$, $\mathcal{P}_2$, $\mathcal{P}_3$) observed in at least two tumor trees (shown to the right of the root nodes).

alterations (Marusyk *et al.*, 2020) and the significant differences observed in the trees describing large cancer cohorts.

In recent years, several methods have been developed to identify recurrent relations between alterations from cross-sectional multi-region data or from single-cell sequencing data. REVOLVER (Caravagna *et al.*, 2018) uses a maximum-likelihood approach and transfer learning to infer a tumor model that jointly describes all tumors in a cohort. Such model is used to infer a tree for each individual tumor, describing the temporal order of clonal alterations in the tumor. However, REVOLVER does not identify significantly recurring trajectories of alterations, and focuses instead on clustering the trees describing individual tumors with the goal of detecting common edges within each cluster. HINTRA (Khakabimamaghani *et al.*, 2019) extends REVOLVER by allowing for more complex dependencies in the order of alterations. However, its scalability to the large number of alterations observed in tumors is limited by an exhaustive enumeration of an exponential number of (directed two-state perfect) phylogenies (Christensen *et al.*, 2020). RECAP is an integer programming approach to identify evolutionary patterns in cancer. RECAP models the problem as the identification a consensus tree for a set of tumors and does not focus on the identification of conserved evolutionary trajectories, and is thus more suitable to cluster tumors for subtype identification. cd-CAP (Hodzic *et al.*, 2019) identifies subnetworks of an interaction network with conserved alterations patterns across tumor samples. However, cd-CAP does not consider the order of alterations and, thus, does not provide information on the evolutionary processes common in a tumor type. GeneAccord (Kuipers *et al.*, 2021) proposes a statistical test to identify over-represented *pairs* of co-occurring or clonally exclusive mutations in subclones. However, it does not focus on conserved evolutionary trajectories and it considers only pairs of mutations. TreeMHN (Luo *et al.*, 2021) is a probabilistic framework that infers a mutual hazard network between alterations, which can be used to predict the most likely *linear pathway* of alterations for each tumor. TreeMHN is therefore focused on *linear trajectories* and assumes that there is one model which recapitulates all tumors and cancer subtypes.

CONETT (Hodzic *et al.*, 2020) is an integer programming approach to identify a consensus phylogenetic tree from the trees describing the evolution of a number of tumors. While CONETT does report a tree describing evolutionary patterns in a collection of tumors, its goal is building a single consensus phylogenetic tree whose topology describes ancestor-descendant relationships for the largest possible number of tumors, and it does not focus on finding all possible conserved evolutionary trajectories from the data, corresponding for example to different cancer subtypes. Moreover, in the tree reported by CONETT only paths from the root to any other event correspond to conserved trajectories, and, thus, CONETT cannot identify more complex evolutionary trajectories (e.g., describing two “sibling" clones). In addition, the problem formulation of CONETT tries to optimize the maximum total node depth (distance from the root), which favors long alteration trajectories potentially appearing in a small number of tumors. Finally, the statistical test used by CONETT to assess the significance of a phylogenetic tree permutes both the order of alterations and the set of tumors in which each alteration is found. As a result, the significance of the results may strongly depend on the co-occurrence of alterations in a given set of tumors, instead that on the order of the alterations being conserved in the set of samples.

In this article we present MASTRO (MAximal tumor treeS TRajectOries), a novel computational method for the identification of significant evolutionary trajectories in cancer (Figure 1). MASTRO takes in input a set of phylogenetic trees describing the evolution of tumors in a cohort, and produces in output *all* trajectories observed in at least σ tumors, where σ is a threshold set by the user. MASTRO can report trajectories of any structure, without restricting their topology (e.g., to linear paths). Moreover, it does not assume that the alterations in a trajectory are *consecutive* in the tumors where they are observed, but only that the order with which alterations appear is the tumor is the same as described by the trajectory. While, as we prove, the identification of conserved trajectories is computationally difficult (i.e., an NP-hard problem, see Section 2), MASTRO allows for the efficient identification of all conserved trajectories by formulating the problem as a *frequent itemset mining* problem (Han *et al.*, 2007) and leveraging the efficient tools that have been proposed to solve the problem. MASTRO assesses the significance of each trajectory by using a *conditional* statistical test, in which the set of alterations and the topology of the phylogenetic tree observed in each patient is fixed. In this way MASTRO assesses the degree with which *the order* of alterations in the trajectory is conserved in the tumors, and properly accounts for cancer subtypes with different complements of alterations. MASTRO uses resampling based methods such as the Westfall-Young permutation procedure (Westfall and Young, 1993) to control false discoveries, which allows to consider the complex relation between the various trajectories while properly correcting for multiple hypothesis testing.

We have applied MASTRO to simulated data, showing that it properly controls for false discoveries while identifying conserved trajectories even when they appear in a relatively low number of tumors. We also applied MASTRO to TRACERx non-small-cell lung cancer (NSCLC) multi-region whole exome sequencing data (Jamal-Hanjani *et al.*, 2017) from 99 tumors and to acute myeloid leukemia data from single-cell panel sequencing of 123 tumors (Morita *et al.*, 2020). MASTRO identifies a number of significant evolutionary trajectories recapitulating and extending the results reported in the original studies.

2 Methods

The input to MASTRO is a multiset of n rooted tumor trees $\mathcal{D} = \{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$, which may be obtained from one of the several computational methods which infer the evolutionary history of the corresponding tumors using (multi-region) bulk or single-cell data. Each

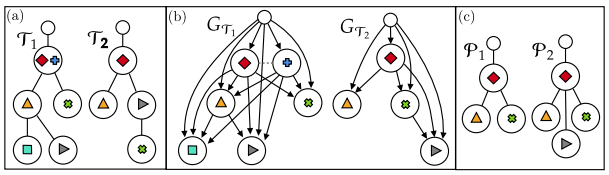

Fig. 2. (a): tumor trees $\mathcal{T}_1$ and $\mathcal{T}_2$. (b): the expanded tumor graphs $G_{\mathcal{T}_i}$ of $\mathcal{T}_i$ (the gray dashed edge denotes an anti-edge between red and blue alterations). (c): the importance of identifying induced trajectories: examples of a trajectory observed in both trees ($\mathcal{P}_1$) forming an induced subgraph on the expanded tumor graphs, and a trajectory ($\mathcal{P}_2$) which satisfies all partial orders among alterations but it is not induced in the expanded tumor graphs, suggesting a spurious clonal exclusivity relation between the grey and orange and green alterations.

tumor tree $\mathcal{T} = (V_\mathcal{T}, E_\mathcal{T}) \in \mathcal{D}$ comprises a set $V_\mathcal{T}$ of nodes and a set $E_\mathcal{T}$ of edges. Each node of $V_\mathcal{T}$ corresponds to a clone in the tumor, and *contains* (i.e., is labelled with) a (potentially empty) collection of alterations (e.g., single-nucleotide variants (SNVs), copy number aberrations (CNA)). The root of each tree $\mathcal{T}$ contains the empty set, and represents normal (germline) cells, while each non-root node $v \in V_\mathcal{T}$ contains a non-empty subset of alterations from a set $\mathcal{A}$ of m alterations. The set of alterations in a node are the alterations that appear in the corresponding clone but not in its ancestors, and the entire complement of alterations of a clone are given by the union of the sets of alterations found in the unique path from the corresponding node to the root. We assume that each alteration appears at most once in each tree, but note that the input can still encode different events on the same genomic region (e.g., a SNV and a CNA affecting the same gene, or the loss of a gene's mutation). For any tumor tree, we say that an alteration a, contained in the node v, is an ancestor of the alteration b, contained in the node $w \neq v$, if the node v belongs to the path from w to the root of the tumor tree.

We define a *trajectory* $\mathcal{P}$ as a tumor tree $\mathcal{P} = (V_\mathcal{P}, E_\mathcal{P})$. A trajectory $\mathcal{P}$ is *observed* in a tree $\mathcal{T}$ if the set of alterations contained in $\mathcal{P}$ is a subset of the alterations $V_\mathcal{T}$ in $\mathcal{T}$ and if all pairwise temporal orderings of the alterations in $\mathcal{P}$ are satisfied in $\mathcal{T}$. Formally, we say that the *trajectory* $\mathcal{P}$ is *observed* in a tree $\mathcal{T}$ if the following conditions hold: 1) for each pair a, b of alterations of $\mathcal{P}$ such that a is an ancestor of b in $\mathcal{P}$, then a is an ancestor of b in $\mathcal{T}$; 2) for all pairs a, b of alterations of $\mathcal{P}$ in the same node in $\mathcal{P}$, they belong to the same node in $\mathcal{T}$ (their ordering is not known); 3) for all pairs a, b of alterations of $\mathcal{P}$ in in different branches of $\mathcal{P}$, they belong to different branches of $\mathcal{T}$ (they are clonally exclusive).

An equivalent representation is given by the following. We define the *expanded tumor graph* $G_\mathcal{T} = (V^G_\mathcal{T}, E^G_\mathcal{T})$ of a tumor tree $\mathcal{T}$ as a directed graph such that: 1) for every alteration a of $\mathcal{A}$ contained in a node $v \in V_\mathcal{T}$ there is a node $v_a \in V^G_\mathcal{T}$ containing only the alteration a; 2) for each pair a, b of alterations of $\mathcal{A}$, there is a directed edge $(v_a, v_b) \in E^G_\mathcal{T}$ if and only if a is an ancestor of b in $\mathcal{T}$; 3) for each pair a, b of alterations belonging the the same node in $\mathcal{T}$, there is an (undirected) special edge (or *anti-edge*) $(v_a, v_b, \star) \in E^G_\mathcal{T}$, denoting the fact that the ordering between a and b is unknown; 4) $V^G_\mathcal{T}$ contains an empty node v_r (the root of $\mathcal{T}$) and $E^G_\mathcal{T}$ contains a directed edge from v_r to all other nodes of G. We show examples of tumor trees and their expanded tumor graphs in Figure 2 (a)-(b).

Given a tumor tree $\mathcal{T}$ and its expanded tumor graph $G_\mathcal{T}$, it is easy to observe that a trajectory $\mathcal{P}$ is observed in $\mathcal{T}$ if and only if $G_\mathcal{P}$ is an induced subgraph of $G_\mathcal{T}$. Our notion of tumor graph is similar to the one proposed by Hodzic *et al.* (2020), but with crucial differences. First, in our formulation the trajectory $\mathcal{P}$ must correspond to an *induced* subgraph of $\mathcal{T}$, which means that *all* pairwise orderings of alterations are preserved, rather than requiring only a *partial* order (as in Hodzic *et al.* (2020)). While our method can be adapted to trajectories composed by partial orderings among alterations, our rationale to enforce this stricter setting is based on the fact that, by definition, a partial order does not specify all relations among alterations, suggesting trajectories that may not be conserved on the underlying tumor trees. For example, consider the trajectories shown in Figure 2 (c). Without the requirement on induced subgraphs, trajectory $\mathcal{P}_2$ is considered conserved in both tumor trees $\mathcal{T}_1$ and $\mathcal{T}_2$, since all partial orders involving the red alteration are satisfied. However, such trajectory describes the grey alteration as belonging to a clone that does not contain neither the orange nor the green alteration, which is not supported by either tumor tree $\mathcal{T}_1, \mathcal{T}_2$. On the contrary, trajectory $\mathcal{P}_1$ describes the clonal exclusivity of the orange and green alterations without any ambiguity. Moreover, our formulation strictly differentiates between directed edges, in which a known temporal ordering among alterations is known, and anti-edges: we do not allow trajectories with a directed edge between two alterations to be observed in a tree containing such alterations and whose order is not known, thus focusing only on temporal relationships supported by the tumor trees (i.e., the data). (In the example of Figure 2, a trajectory with a directed edge, in any direction, between the red and blue alterations cannot be observed in $\mathcal{T}_1$.) We will show in our experimental evaluation that these differences lead to sensibly different results in both simulated and cancer data.

We denote with $\mathcal{P} \in \mathcal{T}$ the fact that the trajectory $\mathcal{P}$ is observed in the tumor tree $\mathcal{T}$. We define the *support* $s_\mathcal{P}$ of $\mathcal{P}$ in the dataset $\mathcal{D} = \{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$ as the number of trees of $\mathcal{D}$ in which $\mathcal{P}$ is observed: $s_\mathcal{P} = \sum_{i=1}^{n} \mathbb{1}[\mathcal{P} \in \mathcal{T}_i]$, where $\mathbb{1}[\cdot]$ is the indicator function ($\mathbb{1}[\cdot] = 1$ if its argument is true, and $\mathbb{1}[\cdot] = 0$ otherwise). Furthermore, we say that a trajectory $\mathcal{P}$ is *maximal* if adding any alteration (in any node of $\mathcal{P}$ or in a new node) not already contained in $\mathcal{P}$ decreases its support $s_\mathcal{P}$. Note that a non-maximal trajectory describes only part of the conserved evolutionary trajectory in a subset of the trees in $\mathcal{D}$.

Our goal is to discover *frequent maximal trajectories* from the tumor trees, as we formalize with the following computational problem.

Definition 1. *(Frequent Maximal Trajectories (FMT) problem)*
Instance: A multiset $\{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$ of $n \geq 1$ tumor trees and a support threshold $\sigma \in [1, n]$.
Solution: All maximal trajectories observed in at least σ tumor trees.

The parameter σ provides a way to control the (minimum) number of tumors in which a maximal trajectory is observed. For example, by setting $\sigma = 2$ (as done in all our experiments), we consider all trajectories appearing in at least two tumors.

The FMT problem is closely related to the problem of enumerating *maximal cliques* from an undirected graph and to the problem of finding *common induced subgraphs* from a collection of graphs (Section 5.1). We prove that the FMT problem is NP-Hard even if we restrict to the case $\sigma = n$ (the proof is in Supplementary Material Section 5.1).

Theorem 1. *The FMT problem with $\sigma = n$ is NP-Hard.*

Given this negative result, it is unlikely that the FMT problem can be solved efficiently in the worst-case. However, we develop a practical solution to this problem by reducing it to a variant of frequent itemset mining, a well studied problem in data mining (Han *et al.*, 2007). We show that, by doing so, frequent maximal trajectories can be enumerated quickly in practice, exploiting already available efficient algorithms for frequent itemsets mining.

2.1 MASTRO: Finding Frequent Maximal Trajectories

In this Section we describe MASTRO, our algorithm to discover frequent maximal trajectories from tumor trees. We first introduce the frequent itemset mining problem, and then describe how MASTRO solves the FMT problem by using frequent itemset mining.

Let $\mathcal{I}$ be universe of *items*, and let a transaction t be a subset of $\mathcal{I}$. A dataset $\mathcal{S}$ is a multiset of n transactions $\mathcal{S} = \{t_1, \ldots, t_n\}$. An *itemset*

A is a subset of $\mathcal{I}$, and its support set $S(A)$ is the set of transactions containing A: $S(A) = \{t_i : A \subseteq t_i, i \in [1,n]\}$. The support s_A of A is defined as the cardinality of $S(A)$: $s_A = |S(A)|$. The problem of *frequent itemset mining* is to compute the set of itemsets with support $\geq \sigma$, that is to compute the set $FI(\mathcal{I},\mathcal{S},\sigma)$ defined as $FI(\mathcal{I},\mathcal{S},\sigma) = \{A \subseteq \mathcal{I} : s_A \geq \sigma\}$. Note that enumerating all itemsets to find the frequent ones is not feasible (their number is $2^{|\mathcal{I}|}$). However, this set can often be computed efficiently by leveraging the *anti-monotonicity* of the support, a key property of itemsets: the support s_A of A is an upper bound to the support of all itemsets containing A. This allows to *prune* large portions of the search space that is explored to find frequent itemsets.

We now show that the set of frequent maximal trajectories in a multiset $\mathcal{D} = \{\mathcal{T}_1,\ldots,\mathcal{T}_n\}$ of tumor trees can be obtained efficiently from $FI(\mathcal{I},\mathcal{S},\sigma)$, for appropriately defined $\mathcal{I}$ and $\mathcal{S}$. For a tumor tree $\mathcal{T} \in \mathcal{D}$, denote the set of *different-branch* edges $E'^{G}_{\mathcal{T}}$ as undirected labelled edges $(v_a, v_b, /)$ between nodes v_a and v_b if alterations a and b belong to different branches of $\mathcal{T}$. Equivalently, such edges do not belong to $E^{G}_{\mathcal{T}}$: $E'^{G}_{\mathcal{T}} = \big\{(v_a, v_b, /) : a \neq b, (v_a, v_b) \notin E^{G}_{\mathcal{T}}, (v_b, v_a) \notin E^{G}_{\mathcal{T}}, (v_a, v_b, \star) \notin E^{G}_{\mathcal{T}}\big\}$. We define the *complete tumor graph* $G^{c}_{\mathcal{T}} = (V^{G}_{\mathcal{T}}, E^{G^c}_{\mathcal{T}})$, where $E^{G^c}_{\mathcal{T}} = E^{G}_{\mathcal{T}} \cup E'^{G}_{\mathcal{T}}$. The complete tumor graph $G^{c}_{\mathcal{T}}$ is a complete graph composed of three types of edges: directed edges between alterations with a known order, and labelled undirected edges between alterations on different branches ($/$) or on the same node ($\star$). We define $\mathcal{I}$ as the union of the sets of edges of all complete tumor graphs: $\mathcal{I} = \bigcup_{\mathcal{T}\in\mathcal{D}} E^{G^c}_{\mathcal{T}}$. We note that the edges of each complete tumor graph of $\mathcal{D}$ is a subset of $\mathcal{I}$; therefore, we define $\mathcal{S} = \{E^{G^c}_{\mathcal{T}} : \mathcal{T} \in \mathcal{D}\}$, with the i-th transaction t_i equal to the edges of the i-th complete tumor graph $t_i = E^{G^c}_{\mathcal{T}_i}$. For an itemset $A \subseteq \mathcal{I}$, denote by $|A|$ the number of edges in A; then, define the set of nodes $V(A)$ that are adjacent to at least one edge of A: $V(A) = \big\{v : \exists(v,w) \in A \vee \exists(w,v) \in A \vee \exists(v,w,\ell) \in A, \ell \in \{/,\star\}\big\}$. We define the set of frequent trajectories $FT(\mathcal{D},\sigma)$ as the set of frequent itemsets in $\mathcal{S}$ such that $|A| = \binom{|V(A)|}{2}$: $FT(\mathcal{D},\sigma) = \left\{A \in FI(\mathcal{I},\mathcal{S},\sigma) : |A| = \binom{|V(A)|}{2}\right\}$. The reason for requiring $|A| = \binom{|V(A)|}{2}$ is that an itemset $A \subseteq \mathcal{I}$, observed in at least $\sigma \geq 1$ complete tumor graphs, represents the set of edges of a complete subgraph with node set $V(A)$ if and only if $|A| = \binom{|V(A)|}{2}$, since it is a subgraph of at least one tumor graph (with unique alterations) and there is an edge (either directed or undirected) between every pair of nodes of $V(A)$. In accordance with the definition of an observed trajectory, $A \subseteq E^{G^c}_{\mathcal{T}}$ and $A \in FT(\mathcal{D},\sigma)$ imply that $(V(A), A \cap E^{G}_{\mathcal{T}})$ is an induced subgraph of $G_{\mathcal{T}}$, and viceversa; therefore, there is a unique mapping between an itemset $\in FT(\mathcal{D},\sigma)$ and a frequent trajectory. We can conclude that any itemset $\notin FT(\mathcal{D},\sigma)$ does not represent a frequent trajectory and is safely discarded.

We now describe our algorithm MASTRO, which is based on the relation between frequent trajectories and frequent itemsets described above. Given in input a multiset $\mathcal{D} = \{\mathcal{T}_1,\ldots,\mathcal{T}_n\}$ of tumor trees and a minimum support threshold $\sigma \in [1,n]$, MASTRO builds the corresponding transactional dataset $\mathcal{S}$ on a universe of items $\mathcal{I}$ as described above. It then uses a known algorithm for frequent itemset mining to extract all frequent itemsets $FI(\mathcal{I},\mathcal{S},\sigma)$, and discards itemsets A such that $|A| \neq \binom{k}{2}$ for every k, obtaining the set $FT(\mathcal{D},\sigma)$. It then finds the set *maximal* frequent trajectories $MFT(\mathcal{D},\sigma)$ from $FT(\mathcal{D},\sigma)$ as follows. For each element A of $FT(\mathcal{D},\sigma)$, if there is no other frequent trajectory $A' \in FT(\mathcal{D},\sigma)$ with the same support set $S(A) = S(A')$ and such that $A' \supseteq A$, then A is the complete tumor graph of a frequent maximal trajectory (otherwise, we could add additional nodes, and corresponding edges, in A without reducing its support, in contrast with the definition of maximal trajectories). We note that this filtering is done efficiently once $FT(\mathcal{D},\sigma)$ has been computed, and that an element of $MFT(\mathcal{D},\sigma)$ always contains the empty root node (and its corresponding edges), since it is common to any subset of the set of complete tumor graphs and, thus, it can be always included without reducing the support of any trajectory. MASTRO then assesses the significance of the trajectories in $MFT(\mathcal{D},\sigma)$ as described in Section 2.2. We implemented MASTRO in Python, using the implementation of LCM (Uno *et al.*, 2004) (version 5.3)[1] to extract frequent itemsets. The implementation of MASTRO is available at `https://github.com/VandinLab/MASTRO`.

2.2 Significance of MASTRO's trajectories

To assess the significance of the support of a given trajectory $\mathcal{P}$, we consider how likely it is to observe $\mathcal{P}$ in a tumor tree $\mathcal{T}$ under the assumption that alterations in $\mathcal{T}$ are randomly assigned to its nodes, i.e., assuming that alterations in the tumor are the same but arise independently of any temporal order. Therefore, we design a statistical test that conditions on the observed set of alterations of each patient, so to directly evaluate the role of ordering among alterations. To do so, we quantify the *expected* number of trees in which we observe $\mathcal{P}$, and how likely it is to observe a *frequent* trajectory just by chance. We compute the significance of trajectory $\mathcal{P}$ by considering the probability of observing $\mathcal{P}$ in each tree under the assumption that the tumor trees have been generated from three different null distributions. In particular, for each tree $\mathcal{T}$ we consider the following three null models: i) each alteration of $\mathcal{T}$ is assigned to one of the nodes of $\mathcal{T}$ chosen independently and uniformly at random; ii) alterations of $\mathcal{T}$ are randomly permuted over the nodes of $\mathcal{T}$, preserving the number of alterations in each node; iii) we sample uniformly at random a topology from $\mathcal{D}$ and assign the alterations of $\mathcal{T}$ independently and uniformly at random to such topology, thus relaxing the conditioning on the observed topology of $\mathcal{T}$. We compute such probabilities for general trajectories, and show that they depend on the topologies of both the trajectory and the tumor trees; in particular, we obtain a dependence on the number of distinct occurrences of the induced subgraph $G_{\mathcal{P}}$ in $G_{\mathcal{T}}$ and the number of automorphism of $G_{\mathcal{P}}$. Due to space constraints, we defer the details on the computation of such probabilities to the Supplementary Material (Section 5.2).

We now describe how to use the probabilities described above (i.e., computed according to one of the null models above) to assess the statistical significance of a trajectory. Let p_i be the probability that a trajectory $\mathcal{P}$ is observed in the i-th tumor tree $\mathcal{T}_i$, with $p_i = 0$ if the set of alterations $\mathcal{A}_{\mathcal{P}}$ contained in nodes of $\mathcal{P}$ is not a subset of $\mathcal{A}_{\mathcal{T}}$. Let $X_1,\ldots,X_n$ be independent Bernoulli random variables such that $\Pr(X_i = 1) = \mathbb{E}[X_i] = p_i$, and $\Pr(X_i = 0) = 1 - p_i$, and define the Poisson Binomial random variable X as the sum of all X_i: $X = \sum_{i=1}^{n} X_i$. The expected value $\mathbb{E}[X]$ of X is the expected support $\mathbb{E}[X] = \sum_{i=1}^{n} p_i$ of the trajectory $\mathcal{P}$ under the null hypothesis. Let $I_{\mathcal{P}} \subseteq [1,n]$ be the set of indices such that $i \in I_{\mathcal{P}}$ if and only if $p_i > 0$. The probability $\Pr(X \geq s_{\mathcal{P}})$ of observing $\mathcal{P}$ with a support equal or higher than $s_{\mathcal{P}}$, under the null hypothesis, is equal to the upper tail of a Poisson Binomial distribution:

$$\Pr(X \geq s_{\mathcal{P}}) = \sum_{k=s_{\mathcal{P}}}^{|I_{\mathcal{P}}|} \sum_{J\subseteq I_{\mathcal{P}}, |J|=k} \prod_{i\in J} p_i \prod_{j\notin J} (1-p_j).$$

The p-value defined by the formula above can be efficiently evaluated using a simple dynamic programming algorithm (Barlow and Heidtmann, 1984) (Supplementary Material Section 5.2.5).

MASTRO leverages *resampling* based methods to control false discoveries, which take into account the correlation structure of the trajectories, achieving higher statistical power than standard methods (e.g., Bonferroni (1936) or Benjamini and Hochberg (1995) corrections). To control the the Family-Wise Error Rate (FWER) (Bonferroni, 1936),

[1] `http://research.nii.ac.jp/~uno/codes.htm`

that is the probability of reporting one or more false discoveries in output, we make use of the Westfall-Young (WY) permutation testing procedure (Westfall and Young, 1993). We also leverage a resampling-based procedure (Storey and Tibshirani, 2003) to estimate the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995), the expected fraction of false discoveries that are reported as significant. We provide all details on the procedures used by MASTRO to control false discoveries in the Supplementary Material (Section 5.3).

3 Experiments

This section describes our experimental evaluation of MASTRO using simulated data (Section 3.1) and cancer data (Section 3.2) from 123 acute myeloid leukemia (AML) patients and from 99 non-small-cell lung cancer (NSCLC) patients.

Cancer data. The AML data includes 543 somatic mutations in 31 cancer-associated genes obtained by single-cell sequencing data from 123 patients; in accordance with previous works, we grouped mutations at the gene level, and analyzed phylogenetic trees generated by SCITE (Jahn *et al.*, 2016). The NSCLC data is from multi-region whole-genome sequencing data first described in Jamal-Hanjani *et al.* (2017). We obtained the data from Caravagna *et al.* (2018), which reports SNVs and focal copy number alterations in 79 putative driver genes, and using the phylogenetic trees reconstructed by CITUP (Malikic *et al.*, 2015). We grouped SNV and gene deletions alterations, but kept gene amplification as a distinct alteration type, in accordance with the analysis performed by CONETT (Hodzic *et al.*, 2020).

3.1 Results on simulated data

In the first simulation, we evaluate the statistical guarantees of MASTRO on false discoveries. We measured the empirical FWER incurred by MASTRO as the fraction of datasets, resampled under the null hypothesis (i.e., the order of all alterations is random, thus there is no significant trajectory in the trees), from which MASTRO reports at least one trajectory as significant (more details are in Supplementary Material Section 6.2.1). As expected, MASTRO reports significant trajectories in a fraction $\leq \alpha$ of the trials, thus correctly controlling the FWER below α. Furthermore, the estimated FWER is always very close to its nominal upper bound, showing that, by using the WY method, MASTRO does not overcorrect for multiple hypothesis testing but rather exploits existing correlations among trajectories. Moreover, using only 10^3 resampled datasets is typically sufficient to obtain accurate estimates. We performed an analogous evaluation of CONETT, in order to elucidate possible differences with our statistical test. We ran CONETT on the same resampled datasets, using analogous support treshold parameters of MASTRO (all details in Supplementary Material Section 6.2.1). Interestingly, we observed (Figure S3) that in all configurations CONETT reports large trees with low p-values (e.g., the majority below 10^{-2}), and that the sizes and p-values computed from the orignal (non-resampled) data are very similar to the ones obtained in our resampled datasets. These results remark the differences in the statistical tests employed by the two methods: CONETT shuffles alterations across the patients, assigning higher importance to their co-occurence in a set of patients rather than to their ordering. At the same time, CONETT does not distinguish anti-edges with directed edges, potentially inferring long trajectories not explicitely observed in the data.

In the second simulation we evaluate the capability of MASTRO in finding significant trajectories implanted in the data, representing a known ground truth. We defer all details to the Supplementary Material (Section 6.2.2), in which we show that MASTRO is very effective and sensible in recalling the ground truth, even in situations in which the implanted trajectory is affected by noise or imperfect inference of the ordering of alterations.

3.2 Results on cancer data

In this section we present the results of MASTRO on cancer data. For all cases, we find all frequent maximal trajectories with MASTRO observed in at least $\sigma = 2$ tumor trees, and evaluate the empirical estimate of the FDR of the top-k most significant results for different values of k (Figure S7). We discard trajectories with only 1 alteration, since their p-value is 1 (there is no specified ordering between any pair of alterations). MASTRO is always very fast: it finds all trajectories and computes all p-values in at most 5 seconds for the first two statistical tests, and 30 seconds for the third test. MASTRO corrects for false discoveries using 10^4 resampled datasets in less than 2 hours (using multithreading over 64 cores). We also run CONETT on the same datasets, using analogous parameters to compare it with our method: we use the same minimum support thresholds (parameter e for inserting alterations in the tree and t to select its root) of MASTRO, equal to $\sigma = 2$, and use default values for other parameters. We only report the edges of the optimal tree found by CONETT in at least σ tumor trees, using the settings described above without imposing additional constraints on the root (e.g., by specifying additional seeds). CONETT needs 15 seconds to find the optimal conserved tree and estimate its p-value on the AML trees and $\approx$ 4.5 hours on the NSCLC trees. We describe below the results obtained when the significance is assessed using the first null model described in Section 2.2. A detailed description of the analyses and additional results are in the Supplementary Material (Section 6.3).

Analysis of AML data. On AML data MASTRO finds 138 maximal trajectories with ≥ 2 alterations observed in at least 2 tumor trees, with an estimated $FDR = 0.2$ for the 40 most significant trajectories (Figure S7). Figure 3 shows a summary of the 40 most significant trajectories (Figure S5 shows all trajectories) into 4 types observed in different subsets of the patients. An interesting observation is that the most *frequent* trajectories are not necessarily the most significant. In fact, the set of 40 most frequent trajectories only contains 23 of the most significant trajectories.

The first two summaries (Figure 3 (a) and (b)) are the two major tumor progression patterns found in AML patients (Schuringa and Bonifer, 2020), as observed independently by Morita *et al.* (2020) and Miles *et al.* (2020): a mutation in an epigenetic factor (DNMT3A, IDH1, or IDH2) precedes mutations in nucleophosmin molecular pathway (NPM1), which are then followed by alterations of signalling genes (KRAS, NRAS, FLT3). MASTRO observes the latter to be almost always found in different branches of the trajectory, confirming their known clonal exclusivity in AML tumors. The third trajectory (Figure 3 (c)) describes an alteration of TET2 as the initiating event, in accordance with a progression pattern described by Miles *et al.* (2020) and the observation that TET2 can occur as both an initiating and a secondary event (Schuringa and Bonifer, 2020).

In addition to the known progression patterns above, MASTRO highlights a fourth trajectory (Figure 3 (d)), characterized by a mutation in NPM1, and then by mutations in RAS, FLT3, and PTPN11, which are mutually exclusive at the clonal level. Differently from the first two trajectory types, we observed that in almost all patients in which such progression pattern is observed, the mutation in NPM1 is not preceded by any other mutation (i.e., NPM1 is the first alteration following the root/germline cells). While NPM1 is an relevant gene for AML (Juliusson *et al.*, 2020; Falini *et al.*, 2020; Zarka *et al.*, 2020), this alternative progression pattern was not previously reported, and may describe a different modality of evolution characterizing a subset of patients not hit by an early alteration of an epigenetic factor.

Overall, the significant trajectories found by MASTRO are complex trajectories with multiple branches, denoting both clonally exclusive and co-occuring alterations. The sets of exclusive alterations are in accordance with the pairs identified by GeneAccord (Kuipers *et al.*, 2021) (for example, RAS with FLT3, and with PTPN11); however, MASTRO does not restrict to the exclusivity of alterations pairs, but extends to sets

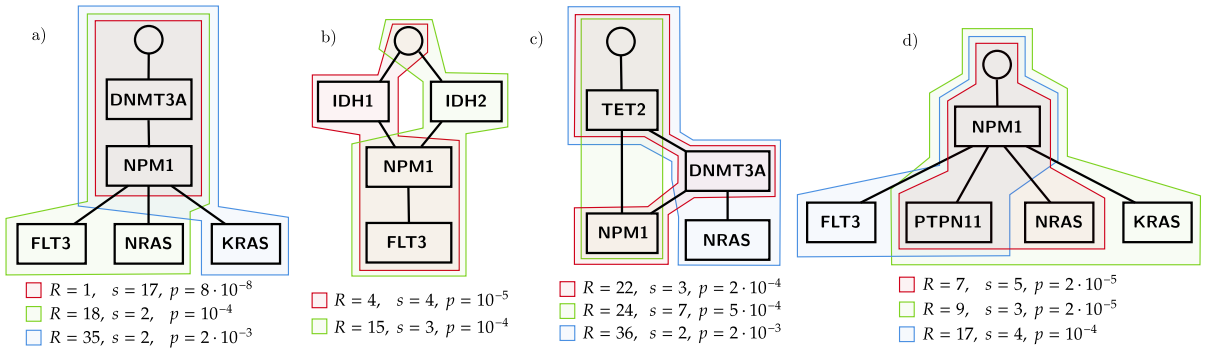

Fig. 3. Summary of the 40 most significant results of MASTRO on AML data (FDR ≈ 0.2) into 4 types observed in different subsets of patients. Each figure shows the composition of different trajectories found by MASTRO, highlighted with boxes of different colors. For each trajectory, we report its rank R (by increasing p-value), its support s, and its p-value p from the statistical test described in Supplementary Material Section 5.2.1.

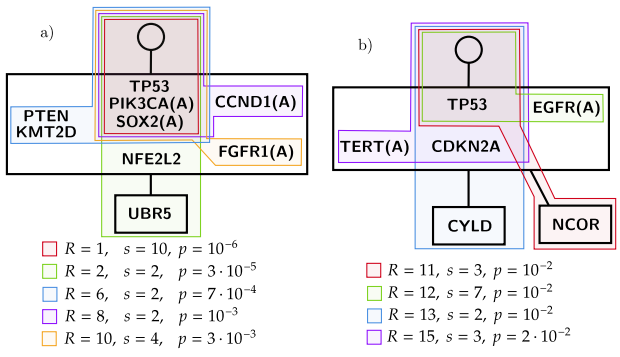

Fig. 4. Summary of the 15 most significant results of MASTRO on NSCLC data (FDR ≈ 0.3). Annotations and color coding as in Figure 3.

of alterations of higher cardinality (for example, the green trajectory of Figure 3(d) describes 3 exclusive subclonal alterations).

We then compared the output of MASTRO with the optimal consensus tree found by CONETT from the AML data (Figure S8). CONETT optimal tree is rooted at TET2 and it contains some of the paths described by the third summary found by MASTRO (and of other trajectories not shown in the summary, see Figure S5). CONETT identifies a tree similar to only one of the progression trajectories found by MASTRO, while including linear trajectories observed in a smaller number of tumor trees: the most frequent one is observed 4 times, while all others have support 3 and 2. Moreover, 7 tumor trees support TET2→NPM1 while 3 support the longer TET2→DNMT3A→NPM1; MASTRO finds both, while CONETT only identifies the latter. This highlights the fact that CONETT is specifically designed to find a consensus tree composed by a collection of linear trajectories that maximize the total path length. Instead, MASTRO simultaneously identifies multiple significantly conserved trajectories observed in different subsets of the patients, providing a more complete description of the conserved evolutionary trajectories.

Analysis of NSCLC data. In NSCLC, the 15 most significant trajectories from MASTRO (Figure S10) have an estimated FDR ≈ 0.3. Also in this case, we observed that the set of 15 most frequent trajectories contains only 6 of the 15 most significant trajectories. The trajectories are summarized in Figure 4. All trajectories are topologically simple, with 2 or 3 nodes in total and with multiple alterations within each node. This is not surprising, giving the topologies of the input trees, which contain few nodes with many alterations with unknown ordering (Figure S1). As a consequence, the data do not contain a signal reliably supporting complex or long trajectories, since there are very few known orderings between alterations. This highlights the fact that bulk sequencing, even if from multi-regional samples, may present intrinsic difficulties in reconstructing the temporal ordering of clonal alterations, compared to the much more informative phylogenetic trees that can be obtained from single-cell sequencing as shown by data from AML patients.

However, in some cases MASTRO is still capable of identifying interesting interaction patterns between alterations. In particular, the two summary trajectories in Figure 4(a) and (b) show that MASTRO identifies groups of alterations, with several genes (e.g., TP53, PIK3CA, CDKN2A, FGFR1, PTEN, CCND1, SOX2) known to be important in NSCLC (Jamal-Hanjani *et al.*, 2017; Jeong *et al.*, 2020), that are more frequently clonal, i.e., they occur more frequently together and in the highest nodes of the tree than expected by chance, in addition to trajectories involving alterations that are more subclonal than expected.

We compared the results of MASTRO with the optimal conserved tree computed by CONETT (Figure S12). CONETT tree is rooted at TP53, and contains several paths connecting various alterations. Almost all alterations belonging to the trajectories reported as significant by MASTRO (Figure 4) belong to the CONETT tree. However, we observe that the most frequent edges that are reported by CONETT are not actually conserved in the underlying tumor trees: this is because CONETT does not differentiate anti-edges (alterations without an ordering, in the same node of the tumor tree) and directed edges (alteration pairs with a known order, in different nodes of the tumor tree). For example, the most frequent edge reported by CONETT is TP53→SOX2(A) (where SOX2(A) denotes the amplification of SOX2), with 12 occurences. We note that, in all the 12 tumor trees containing both TP53 and SOX2(A), such alterations are *always* found in the same node of the tree, therefore there is no evidence of the ordering of such alterations in the tumor trees. This uncertainty in the ordering is supported by the Cancer Cell Fraction (CCF) values of the alterations, which are very close to 1 (and almost always 1) for both alterations in all 12 tumor trees. For other pairs of alterations (Table S1), the alterations involved in the most frequent edges rarely are present in different nodes (e.g., in 1 case over 6 tumor trees), with limited evidence supporting any ordering.

4 Conclusion

In this paper we introduced MASTRO a novel algorithm for the discovery of significant evolutionary trajectories in a set of phylogenetic tumor trees. MASTRO does not assume that the alterations in a trajectory arise *consecutively* in the tumors where they are observed, but only that the order of alterations is conserved. Our experimental analysis on simulated data shows that MASTRO properly controls for false discoveries while identifying conserved trajectories even of relatively low support. We showed MASTRO identifies significantly conserved trajectories in both

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

multi-region whole exome sequencing data from TRACERx non-small-cell lung cancer study (Jamal-Hanjani *et al.*, 2017) and in single-cell panel sequencing data from acute myeloid leukemia (Morita *et al.*, 2020). In both cases MASTRO identified a number of significant evolutionary trajectories recapitulating and extending the results reported in the original studies.

Funding

This work is supported, in part, by the Italian Ministry of Education, University and Research (MIUR), under PRIN Project n. 20174LF3T8 "AHeAD" and the initiative "Departments of Excellence" (Law 232/2016), and by University of Padova under project "SID 2020: RATED-X".

References

Barlow, R. E. and Heidtmann, K. D. (1984). Computing k-out-of-n system reliability. *IEEE Transactions on Reliability*, **33**(4), 322–323.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, **57**(1), 289–300.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubb. Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, **8**, 3–62.

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**(9), 707–714.

Christensen, S., Kim, J., Chia, N., Koyejo, O., and El-Kebir, M. (2020). Detecting evolutionary patterns of cancers using consensus trees. *Bioinformatics*, **36**(Supplement_2), i684–i691.

Deshwar, A. G., Vembu, S., Yung, C. K., Jang, G. H., Stein, L., and Morris, Q. (2015). Phylowgs: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome biology*, **16**(1), 1–20.

Diaz-Uriarte, R. and Vasallo, C. (2019). Every which way? on predicting tumor evolution using cancer progression models. *PLoS comp. bio.*, **15**(8), e1007246.

Eaton, J., Wang, J., and Schwartz, R. (2018). Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics*, **34**(13).

El-Kebir, M., Oesper, L., Acheson-Field, H., and Raphael, B. J. (2015). Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, **31**(12), i62–i70.

El-Kebir, M., Satas, G., Oesper, L., and Raphael, B. J. (2016). Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. *Cell systems*, **3**(1).

Falini, B., Brunetti, L., Sportoletti, P., and Martelli, M. P. (2020). Npm1-mutated acute myeloid leukemia: from bench to bedside. *Blood*, **136**(15), 1707–1721.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., *et al.* (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl j Med*, **366**, 883–892.

Govek, K., Sikes, C., and Oesper, L. (2018). A consensus approach to infer tumor evolutionary histories. In *Proceedings of the 2018 Acm international conference on bioinformatics, computational biology, and health informatics*, pages 63–72.

Han, J., Cheng, H., Xin, D., and Yan, X. (2007). Frequent pattern mining: current status and future directions. *Data mining and knowledge discovery*, **15**(1), 55–86.

Hodzic, E., Shrestha, R., Zhu, K., Cheng, K., Collins, C. C., and Cenk Sahinalp, S. (2019). Combinatorial detection of conserved alteration patterns for identifying cancer subnetworks. *GigaScience*, **8**(4), giz024.

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**(Supplement_1), i427–i435.

Hosseini, S.-R., Diaz-Uriarte, R., Markowetz, F., and Beerenwinkel, N. (2019). Estimating the predictability of cancer evolution. *Bioinformatics*, **35**(14).

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

Jamal-Hanjani, M., Wilson, G. A., McGranahan, N., Birkbak, N. J., Watkins, T. B., Veeriah, S., Shafi, S., Johnson, D. H., Mitter, R., Rosenthal, R., *et al.* (2017). Tracking the evolution of non–small-cell lung cancer. *New England Journal of Medicine*, **376**(22), 2109–2121.

Jeong, Y., Hellyer, J. A., Stehr, H., Hoang, N. T., Niu, X., Das, M., Padda, S. K., Ramchandran, K., Neal, J. W., Wakelee, H., *et al.* (2020). Role of keap1/nfe2l2 mutations in the chemotherapeutic response of patients with non–small cell lung cancer. *Clinical Cancer Research*, **26**(1), 274–281.

Juliusson, G., Jädersten, M., Deneberg, S., Lehmann, S., Möllgård, L., Wennström, L., Antunovic, P., Cammenga, J., Lorenz, F., Ölander, E., *et al.* (2020). The prognostic impact of flt3-itd and npm1 mutation in adult aml is age-dependent in the population-based setting. *Blood advances*, **4**(6), 1094–1101.

Kent, D. G. and Green, A. R. (2017). Order matters: the order of somatic mutations influences cancer evolution. *Cold Spring H. persp. in medicine*, **7**(4), a027060.

Khakabimamaghani, S., Malikic, S., Tang, J., Ding, D., Morin, R., Chindelevitch, L., and Ester, M. (2019). Collaborative intra-tumor heterogeneity detection. *Bioinformatics*, **35**(14), i379–i388.

Kuipers, J., Moore, A. L., Jahn, K., Schraml, P., Wang, F., Morita, K., Futreal, P. A., Takahashi, K., Beisel, C., Moch, H., *et al.* (2021). Statistical tests for intra-tumour clonal co-occurrence and exclusivity. *PLoS computational biology*, **17**(12).

Lawson, D. A., Kessenbrock, K., Davis, R. T., Pervolarakis, N., and Werb, Z. (2018). Tumour heterogeneity and metastasis at single-cell resolution. *Nat. cell biology*, **20**(12), 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**(1), 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**(1), 49–63.

Luo, X. G., Kuipers, J., and Beerenwinkel, N. (2021). Joint inference of repeated evolutionary trajectories and patterns of clonal exclusivity or co-occurrence from tumor mutation trees. *bioRxiv*.

Malikic, S., McPherson, A. W., Donmez, N., and Sahinalp, C. S. (2015). Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, **31**(9).

Malikic, S., Jahn, K., Kuipers, J., Sahinalp, S. C., and Beerenwinkel, N. (2019a). Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nature communications*, **10**(1), 1–12.

Malikic, S., Mehrabadi, F. R., Ciccolella, S., Rahman, M. K., Ricketts, C., Haghshenas, E., Seidman, D., Hach, F., Hajirasouliha, I., and Sahinalp, S. C. (2019b). Phiscs: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. *Genome research*, **29**(11), 1860–1877.

Marusyk, A., Janiszewska, M., and Polyak, K. (2020). Intratumor heterogeneity: the rosetta stone of therapy resistance. *Cancer cell*, **37**(4), 471–484.

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., *et al.* (2020). Single-cell mutation analysis of clonal evolution in myeloid malignancies. *Nature*, **587**(7834), 477–482.

Morita, K., Wang, F., Jahn, K., Hu, T., Tanaka, T., Sasaki, Y., Kuipers, J., Loghavi, S., Wang, S. A., Yan, Y., *et al.* (2020). Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nature communications*, **11**(1).

Navin, N. E. (2014). Cancer genomics: one cell at a time. *Genome biology*, **15**(8).

Nowell, P. C. (1976). The clonal evolution of tumor cell populations: Acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression. *Science*, **194**(4260), 23–28.

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., *et al.* (2015). Effect of mutation order on myeloproliferative neoplasms. *New England Journal of Medicine*, **372**(7), 601–612.

Popic, V., Salari, R., Hajirasouliha, I., Kashef-Haghighi, D., West, R. B., and Batzoglou, S. (2015). Fast and scalable inference of multi-sample cancer lineages. *Genome biology*, **16**(1), 1–17.

Ross, E. M. and Markowetz, F. (2016). Onconem: inferring tumor evolution from single-cell sequencing data. *Genome biology*, **17**(1), 1–14.

Schuringa, J. J. and Bonifer, C. (2020). Dissecting clonal heterogeneity in aml. *Cancer cell*, **38**(6), 782–784.

Schwartz, R. and Schäffer, A. A. (2017). The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, **18**(4), 213–229.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, **100**(16), 9440–9445.

Turajlic, S., Xu, H., Litchfield, K., Rowan, A., Chambers, T., Lopez, J. I., Nicol, D., O'Brien, T., Larkin, J., Horswell, S., *et al.* (2018). Tracking cancer evolution reveals constrained routes to metastases: Tracerx renal. *Cell*, **173**(3), 581–594.

Uno, T., Kiyomi, M., Arimura, H., *et al.* (2004). Lcm ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets. In *Fimi*, volume 126.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons.

Yates, L. R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., Aas, T., Alexandrov, L. B., Larsimont, D., Davies, H., *et al.* (2015). Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature medicine*, **21**(7), 751–759.

Zaccaria, S., El-Kebir, M., Klau, G. W., and Raphael, B. J. (2018). Phylogenetic copy-number factorization of multiple tumor samples. *Journal of Computational Biology*, **25**(7), 689–708.

Zafar, H., Navin, N., Chen, K., and Nakhleh, L. (2019). Siclonefit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome research*, **29**(11), 1847–1859.

Zarka, J., Short, N. J., Kanagal-Shamanna, R., and Issa, G. C. (2020). Nucleophosmin 1 mutations in acute myeloid leukemia. *Genes*, **11**(6), 649.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

5 Supplementary Material

5.1 Proofs

In this section we prove Theorem 1, that implies the NP-Hardness of the MFT problem. We restrict to the particular case $\sigma = n$ by defining the following simpler problem.

Definition 2. *(Maximal Trajectories (MT) problem)*
Instance: A multiset of $k \geq 1$ tumor trees $\{\mathcal{T}_1, \ldots, \mathcal{T}_k\}$.
Solution: All maximal trajectories observed in all tumor trees.

Theorem 2. *The MT problem is NP-Hard.*

Proof. Our proof is based on a reduction of the MT problem to the problem of computing all maximal cliques from an undirected graph (the *all-clique problem*), which is NP-Hard (Koch, 2001); we define it as follows.

Definition 3. *(All-clique problem, (Koch, 2001))*
Instance: An undirected graph $G = (V, E)$.
Solution: All maximal complete subgraphs of G.

Given an undirected graph $G = (V, E)$ with $V = \{1, \ldots, n\}$, denote its complement $G^c = (V, E^c)$ with edges $(v, w) \in E^c$ if and only if $(v, w) \notin E$. Let $E^c = \{e_1, \ldots, e_m\}$ with $e_i = (a_i, b_i)$ and $m = |E^c|$; define the set of $m + 1$ tumor trees $\{\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_m\}$ such that:

- for each i, $\mathcal{T}_i = (V \cup \{0\}, E_i)$;
- $E_0 = \{(0, j), j \in V\}$;
- $E_i = E_0 \setminus \{(0, b_i)\} \cup \{e_i\}$.

We now prove that a maximal trajectory $\mathcal{P} = (V_\mathcal{P}, E_\mathcal{P})$ observed in the $m + 1$ trees $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_m$ if and only if the subgraph induced by the set of nodes $V_\mathcal{P} \setminus \{0\}$ is a maximal clique of G.

Consider a clique $M = \{i_1, i_2, \ldots, i_k\} \subseteq V$ of $G = (V, E)$, and consider the maximal trajectory $\mathcal{T}_M$ with edges $\{(0, i_1), (0, i_2), \ldots, (0, i_k)\}$. Note that $\mathcal{T}_M$ is observed in $\mathcal{T}_0$. Now consider $\mathcal{T}_j$ with $j > 0$ and a vertex $i_\ell \in M$. Either $(0, i_\ell)$ is observed in $\mathcal{T}_j$, or there is vertex v such that $(0, v)$ and (v, i_ℓ) are observed in $\mathcal{T}_j$. Note that by construction $(v, i_\ell) \notin E$, that is, (v, i_ℓ) is not an edge of G, therefore $v \notin M$, which implies that $\mathcal{T}_M$ is observed in T_j. Since $\mathcal{T}_M$ is observed in $\mathcal{T}_0$ and in $\mathcal{T}_j$ for all $j > 0$, $\mathcal{T}_M$ is observed in all the $m + 1$ trees $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_m$. Note that the maximality of M implies the maximality of $\mathcal{T}_M$.

Now consider a trajectory $\mathcal{P} = (V_\mathcal{P}, E_\mathcal{P})$ observed in the $m+1$ trees $\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_m$. Note that since $\mathcal{P}$ is observed in $\mathcal{T}_0$, $E_\mathcal{P} = \{(0, i) : i \in V_\mathcal{P}\}$. Since $\mathcal{P}$ is observed in $\mathcal{T}_j$ for all $j > 0$, there is no pair $i, j \in V_\mathcal{P} \setminus \{0\}$ such that $(i, j) \in E^c$, that is, there is no pair $i, j \in V_\mathcal{P} \setminus \{0\}$ such that $(i, j) \notin E$. Therefore, the set $V_\mathcal{P} \setminus \{0\}$ is a clique of G. Note that the maximality of $\mathcal{P}$ implies the maximality of $V_\mathcal{P} \setminus \{0\}$. □

5.2 Details on MASTRO's statistical tests

In this Section we give the details for the three statistical tests used by MASTRO to assess the statistical significance of the trajectories.

Let $k = |V^G_\mathcal{P}| - 1$ be the number of alterations in $G_\mathcal{P}$, and $t = |V^G_\mathcal{T}| - 1$ be the number of alterations in $G_\mathcal{T}$ (both are equal to the number of nodes in the expanded tumor graph -1, ignoring the root empty node). Recall that a subgraph $H = (V_H, E_H)$ is isomorphic to $G_\mathcal{P}$ if there exists a bijection $f : V_H \to V^G_\mathcal{P}$ such that $(v, w) \in E_H$ if and only if $(f(v), f(w)) \in E^G_\mathcal{P}$. In this case, we denote $H \simeq G_\mathcal{P}$. Let the count $c(G_\mathcal{P}, G_\mathcal{T})$ of $G_\mathcal{P}$ in $G_\mathcal{T}$ be the number of subsets S of size $|V^G_\mathcal{P}|$ of the vertex set $V^G_\mathcal{T}$ of $G_\mathcal{T}$ whose induced subgraph $G_\mathcal{T}[S]$ is isomorphic[2] to $G_\mathcal{P}$:

$$c(G_\mathcal{P}, G_\mathcal{T}) = \left|\left\{S \subseteq V^G_\mathcal{T}, G_\mathcal{P} \simeq G_\mathcal{T}[S]\right\}\right|.$$

Additionally, define the number of automorphism $a(G_\mathcal{P})$ of $G_\mathcal{P}$ as the number of permutations π of the vertex set $V^G_\mathcal{P}$, such that every edge (a, b) belongs to $E^G_\mathcal{P}$ if and only if $(\pi(a), \pi(b)) \in E^G_\mathcal{P}$; in other words, a permutation σ defines a graph isomorphism from $G_\mathcal{P}$ to itself. We have

$$a(G_\mathcal{P}) = \left|\left\{\pi : (a, b) \in E^G_\mathcal{P} \iff (\pi(a), \pi(b)) \in E^G_\mathcal{P}\right\}\right|.$$

5.2.1 Indipendent assignment model

We now formally describe the first null model to assess the significance of frequent trajectories. Let $\mathcal{A}_\mathcal{T}$ be the set of alterations of $\mathcal{A}$ contained in the nodes of $\mathcal{T}$. We assume that alterations $\mathcal{A}_\mathcal{T}$ are placed independently and uniformly at random in the nodes of $\mathcal{T}$ (ignoring the root of $\mathcal{T}$). Note that, in this setting, some nodes of $\mathcal{T}$ may be empty: we take into account the possibility that some of the orderings among alterations are not always preserved. More formally, define the set W of all trees isomorphic to $\mathcal{T}$ such that, for each graph $T = (V, E) \in W$, it holds that each node $v \in V$ (with v different from the root of $\mathcal{T}$) contains a disjoint subset of $\mathcal{A}_\mathcal{T}$, and whose union corresponds to $\mathcal{A}_\mathcal{T}$: it holds $\{a \in v\} \cap \{a \in w\} = \emptyset, \forall v, w \in V, v \neq w$, and $\bigcup_{v \in V}\{a \in v\} = \mathcal{A}_\mathcal{T}$. We define the probability distribution $\mu^I_\mathcal{T}$ as the uniform distribution $U(W)$ on the set W. The probability that $\mathcal{P} \in \mathcal{T}$, assuming that $\mathcal{T}$ is a sample from $\mu^I_\mathcal{T}$, is

$$\Pr_{\mathcal{T} \sim \mu^I_\mathcal{T}}(\mathcal{P} \in \mathcal{T}) = \frac{c(G_\mathcal{P}, G_\mathcal{T})a(G_\mathcal{P})}{t^k}.$$

Note that the computation of the probability above requires to compute the number $c(G_\mathcal{P}, G_\mathcal{T})$ of subgraph isomorphisms, and the number of automorphisms $a(G_\mathcal{P})$. The subgraph isomorphism problem is computationally difficult in the worst case, and in MASTRO we use the efficient implementation of the vf2 algorithm (Cordella *et al.*, 2001) provided by `networkx`[3] to compute $c(G_\mathcal{P}, G_\mathcal{T})$. The computation of $a(G_\mathcal{P})$ is done by exhaustive enumeration of the permutations of the vertices of $G_\mathcal{P}$, which has been efficient in all our experiments.

5.2.2 Probabilities of simple trajectories

While the computations introduced in the previous section hold for general trajectories, we describe simplified expressions for some simpler cases.

Let a trajectory $\mathcal{P}$ composed of edges $E_\mathcal{P} = \{(r, a), (a, b)\}$, where r is the root node and $a, b \in \mathcal{A}$. The expanded tumor graph $G_\mathcal{P}$ of $\mathcal{P}$ contains the edges $E^G_\mathcal{P} = \{(r, a), (r, b), (a, b)\}$. It is simple to verify that the number of automorphisms $a(G_\mathcal{P})$ of $\mathcal{P}$ is 1, since there is only one permutation of the vertices labels that retains the same set of edges (the identity). Furthermore, we observe that, for any tumor graph $G_\mathcal{T}$ with $E^G_\mathcal{T} \supseteq E^G_\mathcal{P}$, the number of isomorphic induced subgraphs $c(G_\mathcal{P}, G_\mathcal{T})$ is simply given by the number of edges $(v, w) \in E^G_\mathcal{P}$ such that $r \notin v$, which is $|E^G_\mathcal{T}| - t$, where $t = |V^G_\mathcal{T}| - 1$. Therefore, the probability of trajectories with the same topology of $\mathcal{P}$ is $(|E^G_\mathcal{T}| - t)/t^2$.

We now consider a trajectory $\mathcal{P}$ composed of edges $E_\mathcal{P} = \{(r, a), (r, b)\}$; in this example, alterations a and b belong to different lineages and therefore show a potential pattern of clonal exclusivity. For this case, we observe that $a(G_\mathcal{P}) = 2$, since the permutation $\sigma(r) = r$, $\sigma(a) = b$, and $\sigma(b) = a$ yields an isomorphic graph (in addition to the

[2] Note that labelled undirected edges in both $E^G_\mathcal{T}$ or $E^G_\mathcal{P}$ can be replaced by two directed edges of opposite direction when checking for subgraph isomorphism.

[3] `https://networkx.org/`

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

identity). The number of isomorphic induced subgraphs $c(G_{\mathcal{P}}, G_{\mathcal{T}})$ in $G_{\mathcal{T}}$, for any tree $\mathcal{T}$, is equal to the number of (non-ordered) pairs (v, w) such that $v, w \in V^G_{\mathcal{T}}$ and v is not an ancestor of w, and viceversa, a quantity very easy to compute.

5.2.3 Permutation assignment

We now introduce the second null model used by MASTRO. For this statistical test, we randomly permute alterations, keeping the topology of the tree and the number of alterations in each node fixed. First, we define the set W as in Section 5.2.1, but we additionally require that the number of alterations in a node of $T \in W$ is equal to the number of alterations in the same node of $\mathcal{T}$. Equivalently, the expanded tumor graph G_T of T has the same topology of the expanded tumor graph $G_{\mathcal{T}}$ of $\mathcal{T}$, but the alterations are randomly permuted. We define the probability distribution μ_P as the uniform distribution $U(W)$ on the set W. We obtain that the probability of observing $\mathcal{P}$ in $\mathcal{T}$, assuming that $\mathcal{T}$ is a sample from μ_P, is

$$\Pr_{\mathcal{T}\sim\mu_P}(\mathcal{P} \in \mathcal{T}) = \frac{c(G_{\mathcal{P}}, G_{\mathcal{T}})a(G_{\mathcal{P}})(t-k)!}{t!}.$$

5.2.4 Independent assignment in random topology

In this third test, we assume that the topology of each tree $\mathcal{T}$ is not fixed, but sampled uniformly at random from the set of topologies of all tumor trees. This allows to take into account alternative topologies that are in accordance with the ones observed in the cohort. For a given topology, alterations are uniformly and independently assigned as described in Section 5.2.1. Combining these two sampling steps, we denote the probability distribution μ_T. It follows that the probability of observing a trajectory $\mathcal{P}$ within a tree $\mathcal{T}$ is

$$\Pr_{\mathcal{T}\sim\mu_T}(\mathcal{P} \in \mathcal{T}) = \frac{1}{n}\sum_{i=1}^{n} \frac{c(G_{\mathcal{P}}, G_{\mathcal{T}_i})a(G_{\mathcal{P}})}{\left(\left|V^G_{\mathcal{T}_i}\right| - 1\right)^k}.$$

5.2.5 Computation of the p-value

In this section we give the details on how to compute the p-value defined in Section 2.2. The idea is to use a dynamic programming approach, exploiting the following property:

$$\Pr\left(\sum_{i=1}^{j} X_i = k\right)$$
$$= \Pr\left(\sum_{i=1}^{j-1} X_i = k\right)(1-p_j) + \Pr\left(\sum_{i=i}^{j-1} X_i = k-1\right)p_j.$$

5.3 Details on Resampling Procedures to control False Discoveries

In this section we present the procedures we employ in MASTRO to provide guarantees on false discoveries by correcting for multiple hypothesis testing. In Section 5.3.1 we describe the Westfall-Young (WY) permutation testing procedure (Westfall and Young, 1993), which allows to control the Family-Wise Error Rate (FWER) (Bonferroni, 1936), that is the probability of reporting one or more false discoveries in output. In Section 5.3.2 we describe a resampling-based procedure to estimate the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995), the expected fraction of false discoveries that are reported as significant.

Let $\mathcal{D}^i$ be the i-th resampled dataset according to one of the statistical null distributions introduced in Section 2.2. In particular, considering the null model described in Section 5.2.1, each $\mathcal{D}^i = \{\mathcal{T}^i_1, \ldots, \mathcal{T}^i_n\}$ is obtained by sampling each $\mathcal{T}^i_j$ from $\mu^I_{\mathcal{T}}$, i.e., assigning the alterations observed in $\mathcal{T}_j$ to the nodes of $\mathcal{T}_j$ independently and uniformly at random. Define $p^i_{\mathcal{P}}$ as the p-value of the trajectory $\mathcal{P}$ computed from $\mathcal{D}^i$, and $p_{\mathcal{P}}$ as the p-value from $\mathcal{D}$.

5.3.1 Bounding the FWER

Let $FWER(\delta)$ be the FWER when using the significance threshold δ to report significant results (i.e., the set of trajectories with p-value $\leq \delta$). $FWER(\delta)$ can be empirically estimated by $\tilde{FWER}(\delta)$ as the fraction of minimum p-values $p^i_{\mathcal{P}}$ that are $\leq \delta$:

$$\tilde{FWER}(\delta) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left[\min_{\mathcal{P}\in MFT(\mathcal{D},\sigma)}\{p^i_{\mathcal{P}}\} \leq \delta\right].$$

To upper bound $\tilde{FWER}(\delta)$ below α while maximizing the set of reported results, the WY permutation testing procedure identifies the maximum δ such that $\tilde{FWER}(\delta) \leq \alpha$ (Westfall and Young, 1993), that is to identify $\hat{\delta} = \max\{\delta : \tilde{FWER}(\delta) \leq \alpha\}$. The WY permutation testing method is often very powerful (and asymptotically optimal (Meinshausen *et al.*, 2011)), as we will show in our experimental evaluation.

5.3.2 Empirical estimator of the FDR

In some cases controlling the FWER can be too restrictive; in many situations, in particular in exploratory analyses, one may tolerate to report a bounded *fraction* of false discoveries if the overall number of discoveries can be significantly increased, achieving more powerful statistical procedures. In this section we describe the procedure used by MASTRO to compute an estimate $\tilde{FDR}(\delta)$ of the $FDR(\delta)$, that is the expected fraction of false discoveries reported using the significance threshold δ.

Our approach is based on the procedure proposed by Storey and Tibshirani (2001, 2003), which yields a conservative estimate of the FDR under arbitrary dependance between the tested hypotheses. To obtain a more efficiently computable estimate, we simplify the procedure of Storey and Tibshirani (2001) by not estimating the proportion of true (null) hypotheses among the tested hypotheses. We instead use the upper bound 1 for such value instead, obtaining a slightly more conservative but much simpler procedure. In this way, $\tilde{FDR}(\delta)$ is defined as the average number of trajectories with p-value $\leq \delta$ computed on the resampled datasets $\{\mathcal{D}^i\}$ divided by the number of trajectoris on $\mathcal{D}$ with p-value $\leq \delta$:

$$\tilde{FDR}(\delta) = \frac{1}{m}\sum_{i=1}^{m} \frac{\left|\{\mathcal{P} \in MFT(\mathcal{D},\sigma) : p^i_{\mathcal{P}} \leq \delta\}\right|}{\max\left\{\left|\{\mathcal{P} \in MFT(\mathcal{D},\sigma) : p_{\mathcal{P}} \leq \delta\}\right|, 1\right\}}.$$

6 Additional experimental results

In this section we present in more details our experimental evaluation of MASTRO using both simulated and cancer data. In Section 3.1 we present a set of simulations with two goals: the first is to experimentally show that MASTRO controls false discoveries (Section 6.2.1); the second is to assess the effectives of MASTRO in discovering a known trajectory implanted on the data (Section 6.2.2). In Section 3.2 we use MASTRO to analyze data from 123 acute myeloid leukemia (AML) patients and from 99 non-small-cell lung cancer (NSCLC) patients.

6.1 Cancer data

The AML data includes 543 somatic mutations in 31 cancer-associated genes obtained by single-cell sequencing data from 123 patients; in accordance with previous works, we grouped mutations at the gene level, and analyzed phylogenetic trees generated by SCITE (Jahn *et al.*, 2016).

We obtained the NSCLC multi-region whole-genome sequencing data first described in Jamal-Hanjani *et al.* (2017). In particular, we obtained the data from (Caravagna *et al.*, 2018), which reports SNVs and focal copy number alterations in 79 putative driver genes, and using the phylogenetic reconstructed by CITUP (Malikic *et al.*, 2015) trees. In accordance with the analysis performed by CONETT (Hodzic *et al.*, 2020), we grouped SNV

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

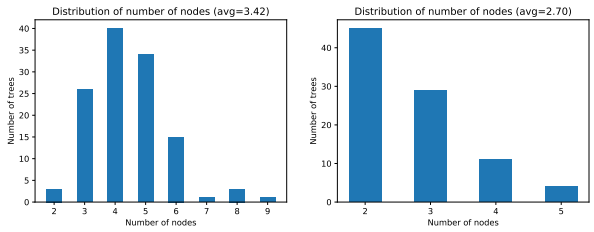

Fig. S1. Distribution of number of nodes in AML (left) and NSCLC (right) tumor trees. The average is annotated in the title.

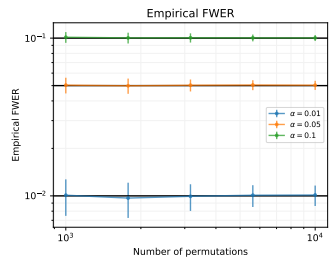

Fig. S2. Empirical FWER for MASTRO as function of the number m of random permutations (x axis), with $m \in [10^3, 10^4]$ and target FWER $\alpha \in \{0.01, 0.05, 0.1\}$ (lines shown in different colors). In a single trial, the FWER is estimated as the fraction of an independent set of 10^4 resampled datasets in which at least one significant result is found. For each m and α we show the average and standard deviation, computed over 10^3 independent trials.

and gene deletions alterations, but kept gene amplification as a distinct alteration type.

Figure S1 shows the distribution of the number of nodes in the trees of both datasets. While in AML the majority of the trees have > 3 nodes with an average of 3.42 nodes per tree, in NSCLC the trees are composed of at most 5 of nodes (including the germline root) with an average of 2.7 nodes per tree. This is due to most alterations being observed with very high abundance, and therefore not reliably ordered, in NSCLC.

6.2 Results on simulated data

6.2.1 Robustness on spurious discoveries

In this first set of experiments, we aim to evaluate the statistical guarantees of MASTRO on reporting false discoveries. To do so, we generated m resampled datasets as described in Section 5.3.1 and computed the corrected significance threshold $\hat{\delta}$ with the WY method (Section 5.3.1), bounding the FWER below α of frequent maximal trajectories with $\sigma = 2$. We considered $\alpha \in \{0.01, 0.05, 0.1\}$ and varied $m \in [10^3, 10^4]$. Note that, since the ordering of all mutations is random, no trajectory should be flagged as significant from such resampled datasets, that is, every reported trajectory is a false discovery. Therefore, we estimated the empirical FWER using an independent set of 10^4 resampled datasets as the fraction of datasets with at least one trajectory with p-value $\leq \hat{\delta}$. We repeat this estimate 10^3 times, and report in Figure S2 averages and standard deviations. As expected, MASTRO reports significant trajectories in a fraction $\leq \alpha$ of the trials, thus correctly controlling the FWER. Furthermore, the estimated FWER is always very close to its nominal upper bound α, showing that, by using the WY method, MASTRO does not overcorrect for multiple hypothesis testing but rather exploits existing correlations among trajectories. We observed that, while we use $m = 10^4$ for all our experiments on real data, using a number m of resamples in the order of 10^3 is typically enough to accurately control the FWER at the levels we considered. Note that we do not show results on the approximation quality of the FDR since, in this setting, all hypothesis are true nulls hypothesis. Therefore, controlling the FWER implies a control of the FDR and viceversa (since controlling the FDR implies a weak control of the FWER), so we would obtain analogous results.

We performed an analogous evaluation of CONETT, in order to elucidate possible differences with our statistical test. In fact, as stated in Section 1, the permutation test employed by CONETT does not preserve the set of alterations in each tumor, differently from our approach. We considered 10^4 random resampled dataset of AML and NSCLC datasets. We ran CONETT on such datasets, fixing the minimum support of the alterations to use in the root (its parameter t) to 10 (analogous to the values used in (Hodzic *et al.*, 2020); we observed similar results with other values of t), and considered different values of the minimum alteration frequencies e (alterations with support $< e$ are not inserted in the tree). Note that we considered only values of $e \geq 3$, since CONETT needs ≈ 4.5 hours to solve its ILP formulation for finding the optimal tree on

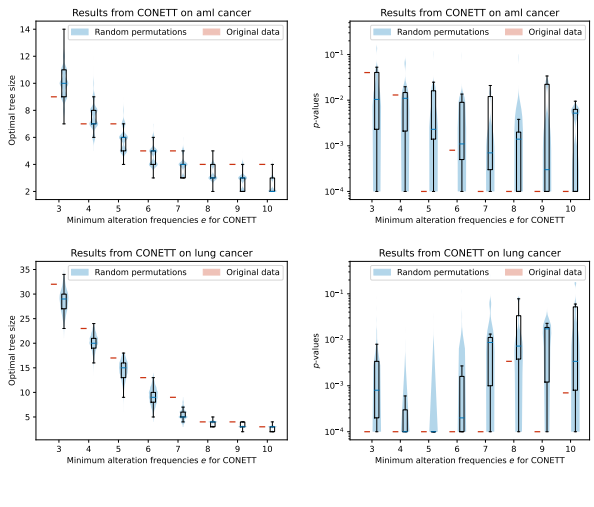

Fig. S3. Comparison of the sizes and empirical p-values of the optimal solutions of CONETT on original and randomly permuted datasets.

the NSCLC tumor trees for $e = 2$: it is not possible to run it on 10^4 resampled datasets in reasonable time. For each resampled dataset we computed the number of nodes of the reported optimal tree and the p-value computed by CONETT (using 10^4 iterations). Figures S3 shows the distribution of the size and p-value of the tree reported by CONETT for different values of e. We note that in all configurations CONETT reports a tree with p-value below 10^{-2} in most resampled datasets. Figures S3 also shows the comparison with the trees obtained on the original data (in red), whose sizes and p-values are very similar to the ones obtained in our resampled datasets. As discussed previously, these results are mostly due to the permutation strategy employed by CONETT, which shuffles alterations across the patients, assigning higher importance to their co-occurence in a set of patients rather than to their ordering. At the same time, CONETT does not distinguish anti-edges with directed edges, potentially inferring long trajectories not explicitly observed in the data. This is particularly relevant for the NSCLC data, in which there is a lower number of nodes per tree and a higher number of alterations in the same node.

6.2.2 Recovery of ground truth

In this section we describe our evaluation of the capability of MASTRO in finding significant trajectories implanted in the data, representing a known ground truth. We generated 10^4 pseudo-random datasets using the AML data as follows: for a given trajectory $\mathcal{P}$ (composed of unique alterations not present in the original data), a fixed value $N \geq 1$, and parameter

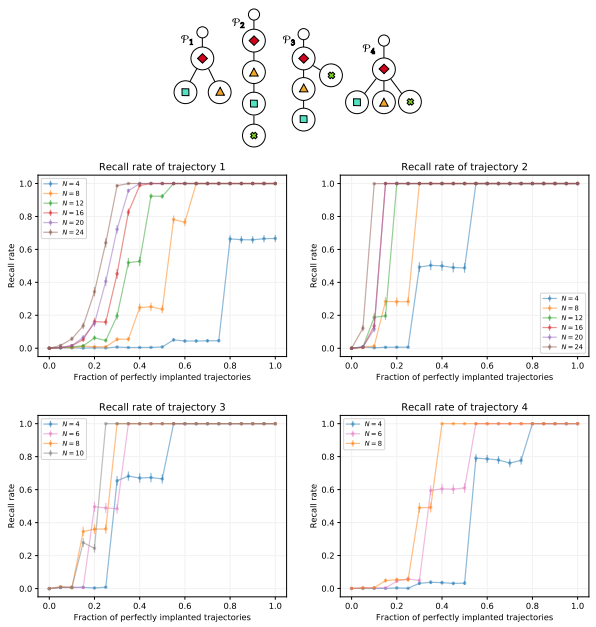

Fig. S4. Topologies (top) and fraction of trials in which the trajectory is found (recall rate, y axes) when the trajectory is implanted exactly in fN patients, and randomly in $(1-f)N$ patients (different values of N are different colors, f varies in the x axes).

$f \in [0,1]$, we implanted $\mathcal{P}$ in N patients, chosen uniformly at random from the set of patients with at least one subgraph isomorphic to $\mathcal{P}$ (i.e., in which it is possible to observe $\mathcal{P}$). However, we implant $\mathcal{P}$ perfectly, i.e., preserving the ordering among its alterations as in $\mathcal{P}$, only in fN of them: in the remaining $(1-f)N$ cases, we insert the alterations of $\mathcal{P}$ randomly. This allows to evaluate the statistical power of MASTRO in cases where the trajectory is not always perfectly observed in the data, for example when stochastic interferences, noise, or upstream errors in phylogenetic reconstruction affect some occurrences of the trajectories. We repeat this process for 4 trajectories with different number of nodes and topologies (shows at the top of Figure S4), and for every combination of N, f, and $\mathcal{P}$ we generated 10^3 datasets. Figure S4 shows the fraction of trials in which MASTRO reports the trajectory as significant with FWER at most 0.05. We can clearly see that, in most cases, a relatively small percentage f of occurrences is sufficient to report the trajectories, with a dependence on N. Naturally, a trajectory observed more frequently has higher potential to reach statistical significance, while an extremely rare trajectory may not. Interestingly, we observe that the topology of the trajectory has a sensible impact on the recall; a simpler topology may arise more easily just by chance, therefore needs to be observed more frequently; on the other hand, it is much less likely to observe more complex trajectories from random assignments of the alterations, therefore can be reported with high confidence from lower evidence. Furthermore, MASTRO achieves high statistical power while controlling the (quite restrictive) FWER. We may expect to retrieve significant trajectories with even higher power when controlling the more flexible FDR. Overall, these observations highlight the effectiveness of MASTRO in identifying significant trajectories from cancer phylogenies with characteristics similar to the ones obtained from current cancer datasets.

6.3 Results on cancer data

In this section we present the trajectories found by MASTRO on two collections of tumor trees computed from cancer data. As previously described, we find all frequent maximal trajectories with MASTRO that contains at least 2 alterations and are observed in at least $\sigma = 2$ tumor trees, and evaluate the empirical estimate of the FDR of the top-k most significant results for different values of k (Figure S7). We present the top-k most significant results for which the estimated FDR is low (e.g., ≤ 0.2). We also run CONETT on the same datasets, using analogous parameters to compare it with our method: we use the same minimum support thresholds (parameter e for inserting alterations in the tree and t to select its root) of MASTRO, equal to $\sigma = 2$, and use default values for other parameters. We only report the edges of the optimal tree found by CONETT in at least σ tumor trees, using the settings described above without imposing additional constraints on the root (e.g., by specifying additional seeds).

6.3.1 Analysis of AML tumor trees

In this section we present the results from AML data. MASTRO finds 138 maximal trajectories with at least 2 alterations and observed in at least 2 tumor trees, and estimates the FDR of the top-k results (sorted by increasing p-value) for various values of k (Figure S7). The p-values are computed with the independent assignment null model (Section 5.2.1). In Figure S6 we show statistics on the maximal trajectories found by MASTRO. We show the sorted p-values and corrected significance thresholds from the WY correction (Figure S6 top left), observing that 13 trajectories are significant with FWER ≤ 0.05, while 16 with FWER ≤ 0.1. We also show the support of the trajectories vs. the number of alterations (Figure S6 top right), the p-values of the trajectories vs. the number of alterations (Figure S6 bottom left), and the p-values of the trajectories vs. the support of alterations (Figure S6 bottom right). Overall, MASTRO finds significant results that are both frequent and rare in the data, with either a relatively small or higher number of alterations, taking into account the topology and individual occurrences of each alteration in the tumor trees. We observe that MASTRO estimates the FDR of the 40 most significant trajectories as 0.2, therefore we focus on these results, expecting most of them to be significantly more frequent than expected by chance. We present a summary of the 40 most significant trajectories discovered by MASTRO in Figure 3, while Figure S5 shows all such trajectories, including their support, p-value and set of tumor trees in which they are observed. Figure 3 shows that the 40 most significant trajectories can be summarised with 4 types of trajectories that are observed in different subsets of the patients. We obtained them combining multiple trajectories with common topologies and alterations (subcomponents highlighted with coloured boxes and surrounding the nodes) that belong to the set of the most frequent and significant trajectories. The first trajectory (a) is characterized by a mutation in DNMT3A, followed by a mutation in NPM1, and progressing with mutations in FLT3, NRAS, and KRAS, which are found in different branches as clonally exclusive. The core component of this trajectory (Germline→DNMT3A→NPM1, surrounded by the red box in the figure) is observed in 17 tumor trees, and it is the most significant result reported by MASTRO (rank $R = 1$, p-value $p = 8 \cdot 10^{-8}$). The progression towards RAS and/or FLT3 is supported by the most frequent and significant trajectories found by MASTRO (for simplicity, we show only two of them in Figure 3, delimited by blue and green lines, while others are shown in Figure S5). The second trajectory (b) is characterized by a mutation in IDH1 or IDH2 (red and blue boxes), both followed by alterations in NPM1 and FLT3. These trajectories are observed in a total of 7 patients, and are among the highest scoring results of MASTRO (ranks 4 and 15). Interestingly, Schuringa and Bonifer (2020) describe these two trajectories as the two major tumor progression patterns found in AML patients, as observed independently by Morita *et al.* (2020) and Miles *et al.* (2020): a mutation in an epigenetic factor (DNMT3A, IDH1, or IDH2) precedes mutations in nucleophosmin molecular pathway (NPM1), which are then followed by alterations of signalling genes (RAS

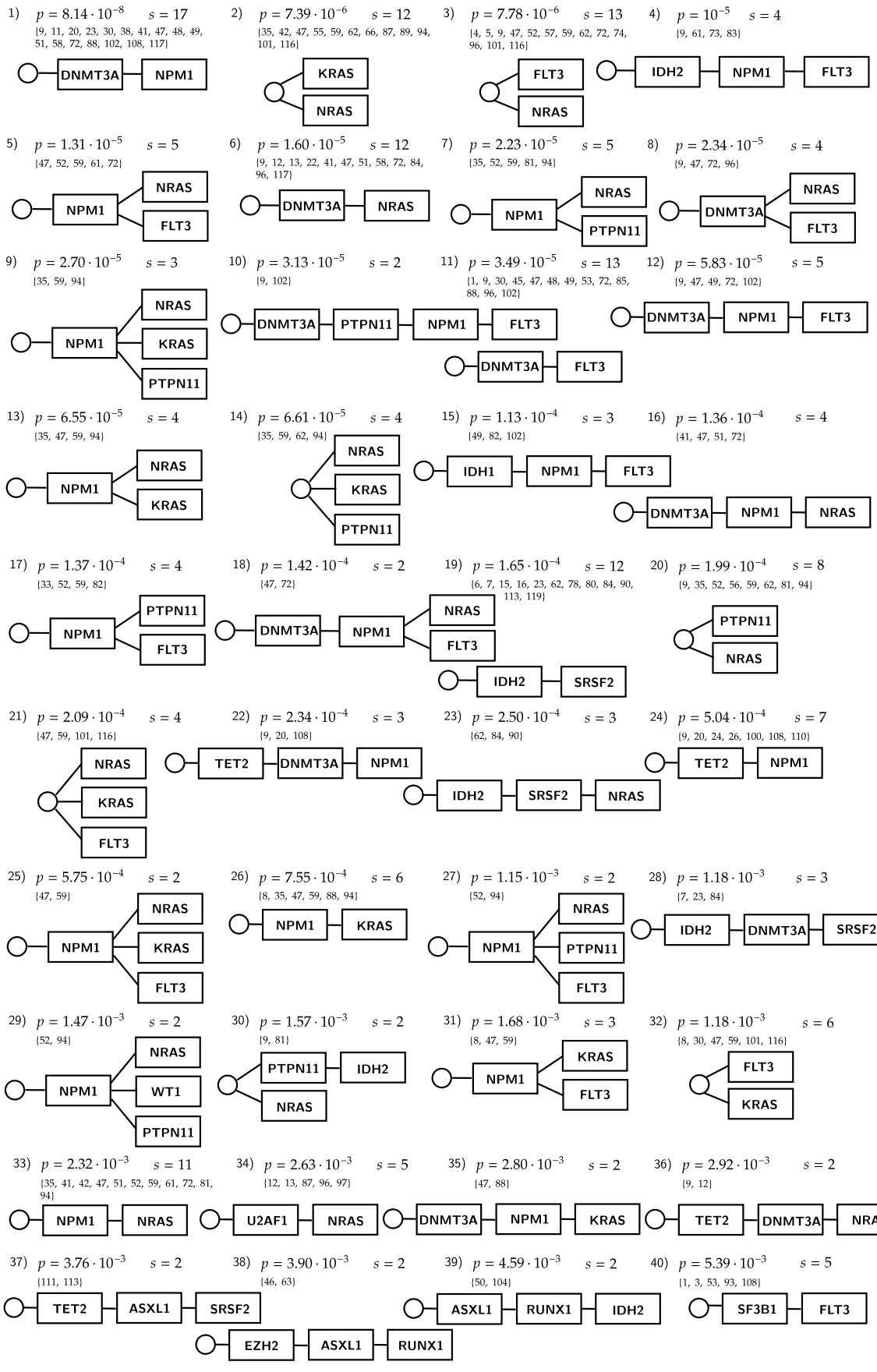

Fig. S5. The 40 most significant maximal trajectories found by MASTRO on AML data. For each trajectory we show its rank, its p-value p (from Section 5.2.1), its support s, and the set of indices of the tumor trees in which the transaction is observed.

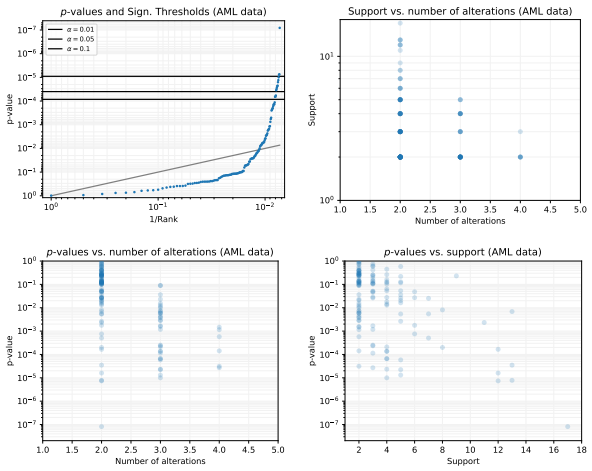

Fig. S6. Results from AML data. Sorted p-values (computed with the independent assignment statistical test, as described in Section 5.2.1) of the 138 trajectories with ≥ 2 alterations observed in at least two tumor trees. Black horizontal lines are significance threshold computed with WY permutation testing over 10^4 permutations. Number of nodes vs trajectories' support. Number of nodes vs p-values. Supports vs p-values.

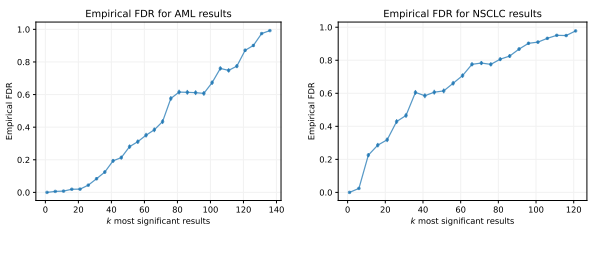

Fig. S7. Empirical estimates of the FDR of the k most significant results (k varies in the x axis).

and FLT3). Furthermore, MASTRO observes the latter to be almost always found in different branches of the trajectory, confirming their known exclusive relationship in AML tumors. The third type of trajectory discovered by MASTRO (trajectory (c)) describes an alteration of TET2 as the initiating event; in 7 tumor trees, TET2 is followed by a mutation in NPM1 (red trajectory), while in 3 of them the alteration of NPM1 is preceded by a mutation in DNMT3A (blue trajectory). In other 2 patients, DNMT3A is followed by a mutation in NRAS instead (green trajectory). Schuringa and Bonifer (2020) report that TET2 can occur as both an initiating and a secondary event, in accordance with a progression pattern described by Miles *et al.* (2020). In addition to these known progression patterns, MASTRO highlights a different trajectory (d): this trajectory is characterized by a mutation in NPM1, and then in mutations in RAS, FLT3, and PTPN11, which are mutually exclusive at the clonal level. Differently from the first two trajectory types, we observed that in almost all patients in which such different progression pattern is observed, the mutation in NPM1 is not preceded by any other mutation (i.e., NPM1 is the first alteration following the root/germline cells). While NPM1 is an relevant gene for AML (Juliusson *et al.*, 2020; Falini *et al.*, 2020; Zarka *et al.*, 2020), this alternative progression pattern was not previously reported; it may describe a different modality of evolution characterizing a subset of patients not hit by an early alteration of an epigenetic factor, and may suggest further investigations.

We remark that most of the trajectories identified by MASTRO are not linear, but describe rather complex trajectories with multiple

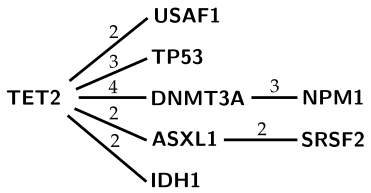

Fig. S8. Optimal conserved tree found by CONETT on AML data (root on the left, leaf nodes on the right). Numbers above an edge denote the number of tumor trees supporting the path from the root to the node on the left of the edge.

branches (e.g., some with $>$ 2 parallel branches). MASTRO identifies sets of alterations with both clonally exclusive and co-occuring alterations leveraging trajectories representing induced subgraphs (with a specified total ordering between all alterations). We observe that the sets of exclusive alterations are in accordance with the pairs identified by GeneAccord (Kuipers *et al.*, 2021) (for example, RAS with FLT3, and with PTPN11); however, MASTRO does not restrict to testing the exclusivity of alterations pairs, but extends the analysis to sets of alterations of higher cardinality (for example, the red trajectory of (d) in Figure 3 describes 3 exclusive subclonal alterations).

We now compare the output of MASTRO with the optimal consensus tree found by CONETT from the AML data, that we show in Figure S8. We observe that the optimal tree is rooted at TET2 and that it contains some of the paths described by the trajectory (c) identified by MASTRO (and of other trajectories not shown in the summary, see Figure S5). CONETT identifies a tree similar to only one of the progression trajectories described previously, while including linear trajectories observed in a smaller number of tumor trees (the most frequent is observed 4 times, all others have support 3 and 2; furthermore, 7 tumor trees support TET2→NPM1 while 3 support the longer TET2→DNMT3A→NPM1; MASTRO finds both, while CONETT only identifies the latter); this highlights the fact that CONETT is specifically designed to find a consensus tree composed by a collection of linear trajectories that maximize the total path length, otaining quite different results. Instead, MASTRO simultaneously identifies different significantly conserved trajectories observed in different subsets of the patients, providing a more complete description of the conserved evolutionary trajectories.

We also show in Figure S9 the p-values, corrected significance thresholds using WY, and estimated FDRs of the top-k results, using the other two statistical tests defined in Section 5.2. We observe that, overall, the reported p-values are mostly similar, with some differences for the smallest ones. While the values of the p-values are, in some cases, slightly different in magnitude, we observe the estimated FDR to be consistent for all three tests, showing that all tests agree on the fact that we expect the k most significant trajectories to be composed by a small fraction of false positives, for most values of k.

6.3.2 Analysis of NSCLC data

We now present the results identified by MASTRO from NSCLC tumor trees. We show the 15 most significant trajectories in Figure S10, that we seleted as their estimated FDR is not too large ($\approx$ 0.3, Figure S7), and summarize them with two trajectories in Figure 4. We show statistics for all trajectories in Figure S11. From Figure S10 we observe that all trajectories are topologically simple, composed by 2 or 3 nodes containing multiple alterations withing each node. This is not surprising, giving the topologies of the input trees, which contain few nodes with many alterations with unknown ordering. We note that MASTRO distinguishes alterations with a known and unknown order (connected by either a directed or undirected

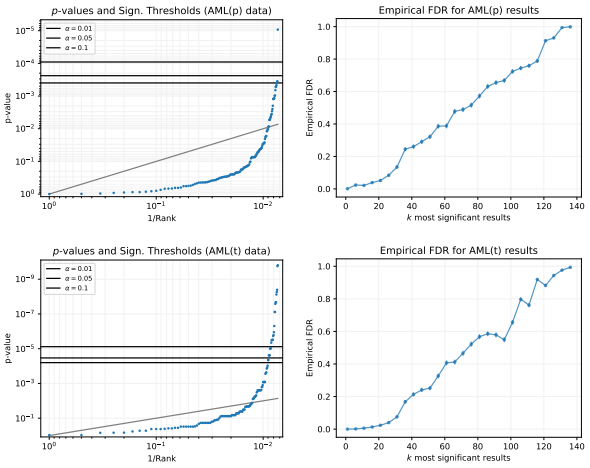

Fig. S9. Results from AML data using statistical tests based on permutations (Section 5.2.3, top plots with (p) label) and random topologies (Section 5.2.4, bottom plots with (t) label). Plots analogous to Figure S6 and S7.

anti-edge in the expanded tumor graphs, Figure 2). Relaxing this feature may lead to finding trajectories with spurious orderings not supported by the data: we argue that MASTRO discovers trajectories satisfying the available temporal information and that are better supported by the input trees. A consequence of this is that the signal that MASTRO evaluates to identify trajectories with significant ordering is much weaker, since there are very few known orderings between alterations. This highlights the fact that bulk sequencing, even if from multi-regional samples, may present intrinsic difficulties in reconstructing the temporal ordering of clonal alterations, compared to the much more informative phylogenetic trees that can be obtained from single-cell sequencing as shown by data from AML patients. However, in some cases MASTRO is still capable of identifying interesting interaction patterns between alterations, that we now describe. The first trajectory summarizing the most significant results of MASTRO, shown in Figure 4 (a), is composed by a core trajectory (surrounded by a red box) involving a mutation of TP53 and amplifications of PIK3CA and SOX2 whose order is not known (they belong to the same node). This trajectory is observed in 10 tumor trees, and it is the most significant of all results (rank $R = 1$, p-value $p = 10^{-6}$). This trajectory extends in 4 ways: in 4 tumor trees, it includes an amplification of FGFR1 (orange trajectory, $p = 3 \cdot 10^{-3}$); in 2 tumor trees, mutations in PTEN and KMT2D are also observed (purple trajectory, $p = 7 \cdot 10^{-4}$); 2 tumor trees also include an amplification of CCND1 (blue trajectory, $p = 10^{-3}$); in 2 tumor trees, it includes an alteration of NFE2L2, followed by a subclonal mutation of UBR5 (green trajectory, $p = 3 \cdot 10^{-5}$). The second trajectory shown in Figure 4 (b) is obtained composing 4 trajectories. All share a mutation of TP53, and extend with: a subsequent mutation of NCOR, observed in 3 tumor trees (red trajectory, $p = 10^{-2}$) an amplification of EGFR (blue trajectory, $p = 10^{-2}$) observed in 7 trees; a mutation of CDKN2A, followed by a mutation of CYLD, observed in 2 trees (green trajectory, $p = 10^{-2}$); an amplification of TERT and mutation of CDKN2A, found in 3 trees (purple trajectory, $p = 2 \cdot 10^{-2}$). These two trajectories summarize the fact that MASTRO identifies groups of alterations, known to be important in NSCLC (Jamal-Hanjani *et al.*, 2017; Jeong *et al.*, 2020), that are more frequently clonal, i.e., they occur more frequently together and in the highest nodes of the tree than expected by chance, in addition to trajectories involving alterations that are more subclonal than expected.

Alteration pairs	$\rightarrow$	$\leftarrow$	$\leftrightarrow$
SOX2(A), TP53	0	0	12
PIK3CA(A), TP53	0	0	10
PIK3CA(A), SOX2(A)	0	0	12
EGFR(A), TP53	1	0	5
KRAS, TP53	1	1	6
MGA, TP53	0	1	6
PIK3CA, TP53	0	1	4
TERT(A), TP53	0	2	5
EGFR, TP53	1	0	5
CDKN2A, TP53	1	0	9

Table S1. Number of times a pair of alteration X, Y (first column from the left) is observed with a known ordering (i.e., X is an ancestor of Y or viceversa, denoted by $X \rightarrow Y$ or $Y \rightarrow X$, second and third column) or with an unknown ordering (i.e., in the same node of the tumor tree, $X \leftrightarrow Y$, last column) in tumor trees inferred from NSCLC data.

We now compare the results of MASTRO with the optimal conserved tree computed by CONETT, shown in Figure S12. We observe that this tree is rooted at TP53, and contains several paths connecting various alterations. Almost all alterations belonging to the trajectories reported as significant by MASTRO (Figure 4) belong to the tree; however, we observe that the most frequent edges that are reported by CONETT are not actually conserved in the underlying tumor trees: this is because CONETT does not differentiate between anti-edges (alterations without an ordering, in the same node of the tumor tree) with directed edges (alteration pairs with a known order, in different nodes of the tumor tree). For example, the most frequent edge reported by CONETT is TP53→SOX2(A) (where SOX2(A) denotes the amplification of SOX2), with 12 occurences. We note that, in all the 12 tumor trees containing both TP53 and SOX2(A), such alterations are *always* found in the same node of the tree, therefore there is no evidence of the ordering of such alterations in the tumor trees. This uncertainty in the ordering is also confirmed by the reported Cancer Cell Fraction (CCF) values (reported by Caravagna *et al.* (2018)), which are always 1 (or very close to 1) in all patients with the alterations, confirming the fact that it is unclear how to distinguish the ordering of such clonal events. We observed similar relationships between TP53 and PIK3CA(A) (co-occurring in 10 trees, always in the same node), and PIK3CA(A) and SOX2(A) (co-occurring in 12 trees, always in the same node); for other pairs of alterations, as we show in Table S1, we observe that alterations involved in the most frequent edges rarely co-occur in different nodes (e.g., in 1 case over 6 tumor trees), bringing scarse evidence of their ordering.

References

Cordella, L. P., Foggia, P., Sansone, C., Vento, M. (2001). An improved algorithm for matching large graphs. *3rd IAPR-TC15 workshop on graph-based representations in pattern recognition*, 149–159.

Koch, I. (2001). Enumerating all connected maximal common subgraphs in two graphs. *Theoretical Computer Science*, **250**(1-2), 1–30.

Meinshausen, N., Maathuis, M. H., and Bühlmann, P. (2011). Asymptotic optimality of the westfall-young permutation procedure for multiple testing under dependence. *The Annals of Statistics*, pages 3369–3391.

Storey, J. D. and Tibshirani, R. (2001). Estimating false discovery rates under dependence, with applications to dna microarrays. Technical report, Technical Report 2001-28, Department of Statistics, Stanford University.

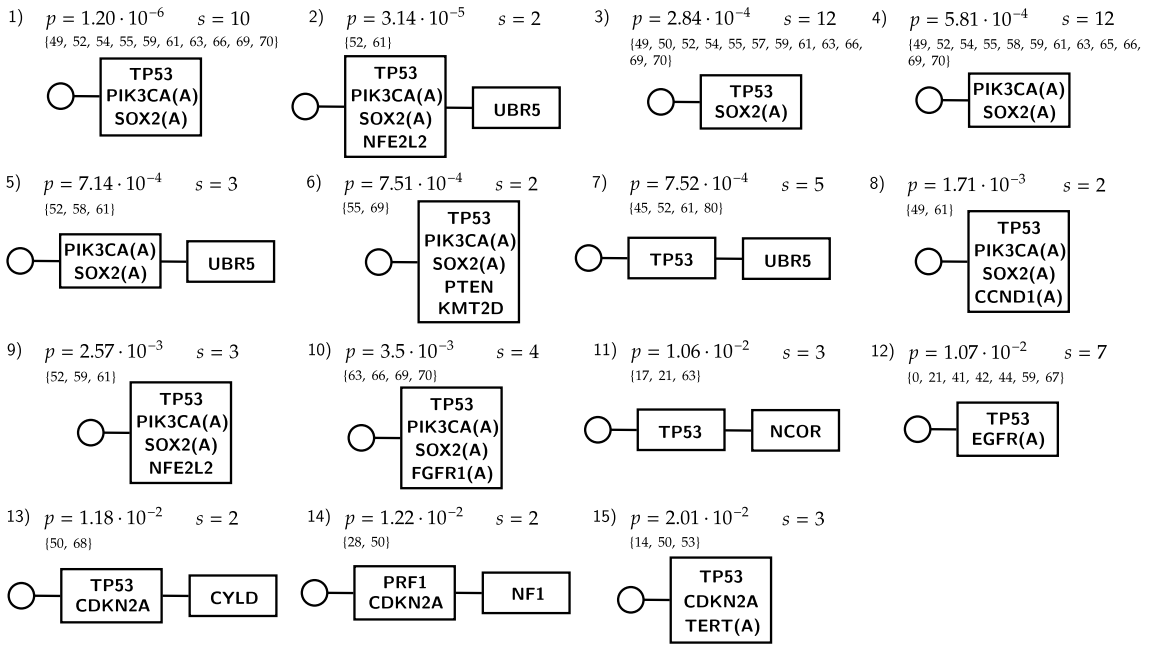

Fig. S10. The 15 most significant maximal trajectories found by MASTRO on NSCLC data. Results shown are analogous to Figure S5.

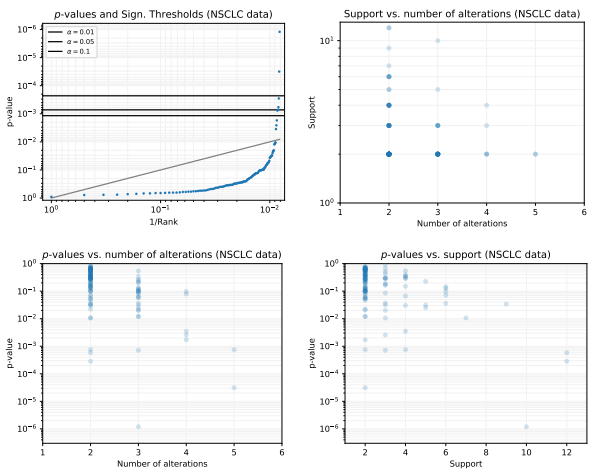

Fig. S11. Results from NSCLC data. p-values (computed as described in Section 5.2.1) of the 124 trajectories with at least two alterations and observed in at least two tumor trees. Black horizontal lines are significance threshold computed with WY permutation testing over 10^4 permutations.

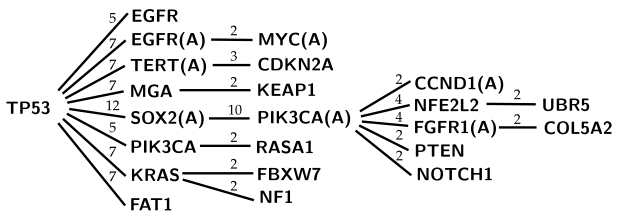

Fig. S12. Optimal conserved tree found by CONETT on NSCLC data (root on the left, leaf nodes on the right). Numbers above an edge denote the number of tumor trees supporting the path from the root to the node on the left of the edge according to CONETT.

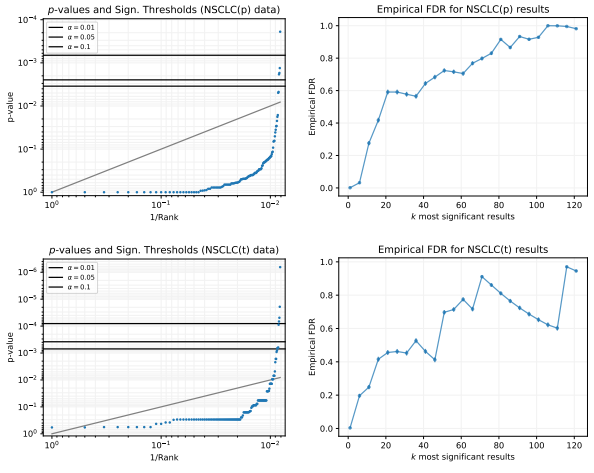

Fig. S13. Results from NSCLC data using statistical tests based on permutations (Section 5.2.3, top plots with (p) label) and random topologies (Section 5.2.4, bottom plots with (t) label). Plots analogous to Figure S11 and S7.