Identification of Conserved Evolutionary Trajectories in Tumors

Ermin Hodzic[1], Raunak Shrestha[2], Salem Malikic[3], Colin C. Collins[4,5], Kevin Litchfield[6], Samra Turajlic[6], and S. Cenk Sahinalp[7]

[1]School of Computing Science, Simon Fraser University, Burnaby, BC, Canada
[2]Department of Radiation Oncology, University of California San Francisco, CA, USA
[3]Department of Computer Science, Indiana University, Bloomington, IN 47408, USA
[4]Department of Urologic Sciences, University of British Columbia, Vancouver, BC, Canada
[5]Vancouver Prostate Centre, Vancouver, BC, Canada
[6]Cancer Evolution and Genome Instability Lab., Francis Crick Institute, London, UK
[7]Cancer Data Science Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA

Abstract

Motivation: As multi-region, time-series, and single cell sequencing data become more widely available, it is becoming clear that certain tumors share evolutionary characteristics with others. In the last few years, several computational methods have been developed with the goal of inferring the subclonal composition and evolutionary history of tumors from tumor biopsy sequencing data. However, the phylogenetic trees that they report differ significantly between tumors (even those with similar characteristics).
Results: In this paper, we present a novel combinatorial optimization method, CONETT, for detection of recurrent tumor evolution trajectories. Our method constructs a consensus tree of conserved evolutionary trajectories based on the information about temporal order of alteration events in a set of tumors. We apply our method to previously published datasets of 100 clear-cell renal cell carcinoma and 99 non-small-cell lung cancer patients and identify both conserved trajectories that were reported in the original studies, as well as new trajectories.
Availability: CONETT is implemented in C++ and available at https://github.com/ehodzic/CONETT.

1 Introduction

Cancer is a disease initiated by somatic genomic alterations. Accumulation of such alterations drives the progression of a cancer to an advanced form. As the tumor grows, new subpopulations of the tumor cells emerge with distinct genomic alteration profiles. After subsequent rounds of selection and expansion, these subpopulations give rise to substantial intra-tumor heterogeneity, which is arguably one of the main challenges in cancer management and treatment.

As multi-region, time-series and single cell sequencing data become more widely available, it is becoming clear that certain tumors share evolutionary characteristics with others. With new computational methods to identify recurrent cancer progression patterns from *multi-dimensional* tumor sequencing data, it may become possible to predict the likely course of evolution and perhaps an effective treatment strategy for certain cancer types [14, 21, 7]. E.g. novel approaches for targeted immuotherapy [5] could be developed into combinatorial strategies that simultaneously target multiple subclonal populations if a comprehensive characterization of recurrent subclonal expansion patterns could be established. Unfortunately, elevated levels of intra-tumor heterogeneity, with intricate subclonal mutational landscapes dominated by inconsequential passenger alterations, may make it difficult to discriminate predictable tumor progression *signals* from *noise* introduced by these statistically insignificant subpopulations.

In the last few years, several computational methods have been developed with the goal of inferring the subclonal composition and/or evolutionary history of tumors from tumor biopsy sequencing data including [9, 6, 27, 22, 10, 16, 28, 33, 8, 23, 24, 34]. Phylogenetic trees for representing tumor evolution processes reported by such methods provide (at least one of) the likely relative temporal order of genomic

alteration events. Unfortunately, the trees they report for many tumor samples differ significantly from other tumors with similar characteristics. In contrast, some recent studies have been able to recapitulate tumor evolution more accurately through long-term clinical followup [32, 1, 29]. These studies were able to find evidence of recurrent progression patterns involving two or more co-occurring genomic alterations. Since, until recently, no computational tool was purposefully developed to identify recurrent patterns of tumor evolution across multiple tumor samples, these studies typically relied on manual inspection of all possible sequences of alteration events, starting from the root of the tumor phylogeny (representing germline), all the way to leaf nodes [29]. We are aware of only two recent exceptions, which are both based on clustering tumor phylogenetic trees by the use of tree similarity/distance measures [25, 4]. In particular, the REVOLVER method [4] employs a maximum-likelihood learning strategy to construct a joint hidden tree model for all tumors in a cohort, which is then used to infer an individual tree depicting the temporal order of clonal alterations in each individual tumor. This is followed by a hierarchical clustering of the trees for the purpose of detecting edges shared within each cluster (and manually constructing likely evolutionary *trajectories* involving those edges).

In parallel to the effort summarized above, the problem of finding common sets of genomic alterations in cross-sectional tumor data has been investigated by several computational approaches. These approaches typically employ molecular interaction networks to identify *connected subnetworks* of genes that have been subjected to somatic alterations, that are *shared* across tumor samples in the sense that at least a fixed number (e.g. one) of the genes in the subnetwork is altered in each tumor sample [31, 20, 19, 30, 2, 3, 15]. Recently, cd-CAP [13] took a stricter approach, requiring preservation of every alteration in the subnetwork in at least a user defined number of tumor samples in a cohort, introducing the notion of a *conserved alteration pattern*. Through cd-CAP it has become possible to detect sizable networks of genes, altered in a similar way across tumor samples. However none of these approaches consider the *temporal order of alterations* to make inference about evolutionary processes common in certain tumor types.

In this paper, we present a novel computational method, CONETT (conserved evolutionary trajectories in tumors), for combinatorial detection of recurrent tumor evolution trajectories. CONETT considers as input a partial temporal ordering of alteration events for each tumor in a given set (of tumors), and aims to construct a "consensus" phylogeny tree whose topology captures the temporal ordering of the largest possible number of events without conflict with the partial temporal order in the individual tumors. In order to determine a consensus temporal ordering between as many pairs of alteration events as possible, CONETT maximizes the average node "depth" in such a phylogeny tree. A path from the root of that tree (which represents a specific *seed* alteration event) to any other node represents an evolutionary trajectory of alteration events which is conserved in a subset of tumors that agree with the temporal ordering of events along that trajectory.

In order to find the largest set of events that can be ordered into a consensus phylogeny, we first introduce the notion of an *evolutionarily conserved alteration set* [1] – a set of genomic alterations S which temporally follow a specific *seed* alteration s through evolutionary paths that consist *solely of* alteration events in S, in a user-specified number of tumors in a cohort. Based on that, we introduce the notion of a *conserved evolutionary trajectory*, as a path originating from the seed alteration s to a subset of alterations in S, with a specific (temporal) order, in a user-defined fraction of the tumors. We then describe a polynomial-time algorithm for the identification of maximum-size evolutionarily-conserved alteration sets from an arbitrary directed graph. Given an evolutionarily conserved alteration set, we finally describe an integer linear programming formulation to identify the *conserved evolutionary trajectory tree* – the consensus topology between the nodes of the set which determines the maximum number of temporally ordered pairs of nodes. Intuitively, the consensus tree represents the likely evolutionary trajectories that could be observed in a tumor in a single tree topology. We have applied CONETT to the TRACERx clear-cell renal cell carcinoma (ccRCC) data set [29] involving 100 tumors, as well as TRACERx non-small-cell lung cancer (NSCLC) data set [17] involving 99 tumors, and identified a number of significantly conserved evolutionary trajectories which involve sequence-altered genes and copy number alteration events – not reported in the original studies.

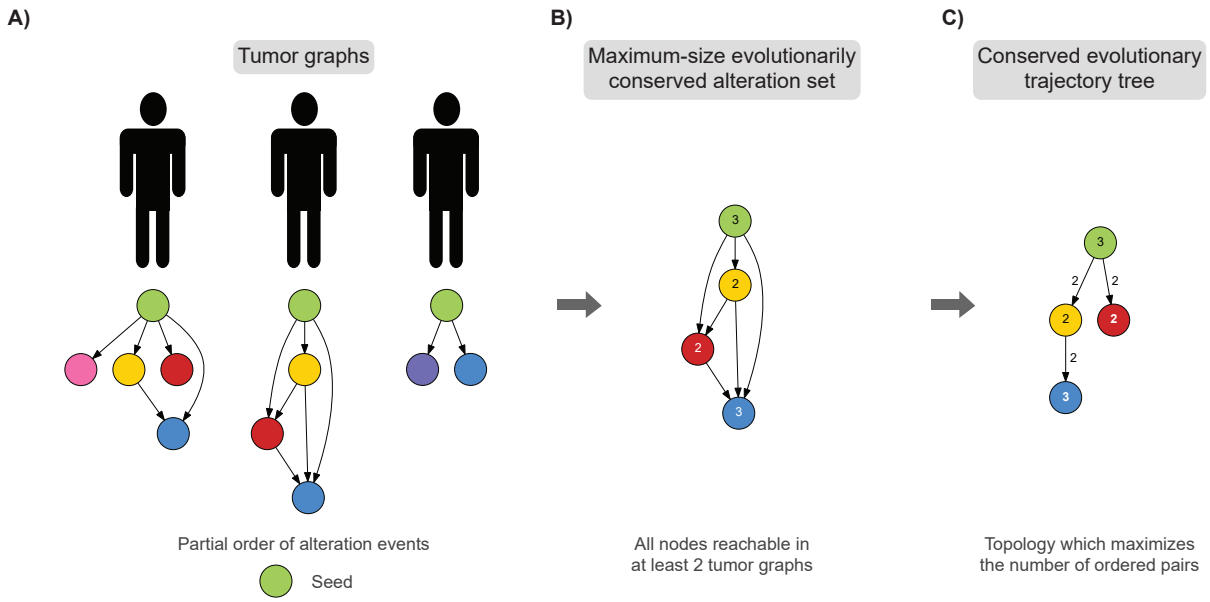

Figure 1: **Overview of CONETT's framework. A)** Tumor graphs represent partial temporal order of alterations in tumors. **B)** An evolutionarily conserved alteration set is identified such that there is an evolutionary path from the seed (green node) to every node in a specified number of tumors each (2 in the example). **C)** Based on the inferred temporal order of alterations in the three tumor graphs, a "consensus" phylogeny tree representing compatible conserved alteration trajectories is identified: here every path from the seed to any other node is conserved in 2 out of 3 tumors. The chosen topology contains the maximum number of node pairs for which there is a "consensus" temporal ordering.

2 Methods

Let $\mathcal{P}$ be a set of tumors and $\{G_p\}(p \in \mathcal{P})$ its corresponding collection of *tumor graphs* – a tumor graph G_p is a *directed graph* corresponding to tumor $p \in \mathcal{P}$ where each node corresponds to either *a somatically altered single gene* in the tumor, or *a copy number altered entire chromosomal arm* (labelled by the gene or chromosome and the type of alteration affecting it) [2] . A directed edge from one node to another indicates temporal precedence of the former alteration event to the latter, possibly inferred through the use of time series, multi-region, single cell or single molecule sequencing data (similar to what was considered by [4]). In case a tumor graph includes directed edges between a pair of nodes in both directions, the alteration events that correspond to the two nodes are considered to be *co-clonal*, i.e. it is not possible to temporally order the two events in this tumor, even though they are known to be on the same *lineage*. Thus, a tumor graph describes a partial temporal order of alteration events within a tumor. Given the above, we say that two nodes v_p and v_q in respective tumor graphs G_p and G_q are *identical* if they correspond to the same gene, or in case of copy number alterations, same chromosomal arm and the same alteration type (i.e. they have the same label) in respective tumors p and q. We also say that a subgraph G'_p of G_p and a subgraph G'_q of G_q are *identical* if their sets of nodes are identical, and these nodes are identically connected. If some subgraph G' has an identical subgraph in G_q, then we say that it is *shared by* G_q.

Consider a temporal ordering of a set of nodes, in the form of a directed path $v_1, v_2, \ldots, v_k$; we call this path a **conserved evolutionary trajectory**, if it is conserved in a significant number of tumors, i.e. if the partial ordering of nodes within a significant number of tumor graphs agrees with the node

[1] A notion of a conserved alteration pattern for detecting co-occurring alterations connected through molecular interactions was introduced earlier [13]. In contrast an evolutionarily conserved alteration set focuses on temporal ordering between alteration events.

[2] Note that CONETT allows the user to distinguish distinct types of somatic alterations such as single nucleotide alterations, short insertions, deletions, inversions and duplications within a gene, as well as copy number gains and losses within a chromosomal arm. However, the studies that published the data sets that CONETT was applied to in this paper, did not associate significant biological differences with these distinctions, and thus, for the sake of simplicity and the ability to compare our results, we similarly largely avoided employing such distinctions.

order along the path. A collection of conserved evolutionary trajectories that are *initiated by* the same founding driver event and have compatible (i.e. pairwise identical) temporal ordering of common nodes may be represented by a single, consensus phylogeny tree that includes each such trajectory as a (sub)path. Our goal is to identify such a tree of conserved evolutionary trajectories which (1) includes the largest possible number of alteration events found on conserved evolutionary trajectories and (2) determines the "consensus" temporal order between as many pairs of alteration events as possible (without conflicting with the tumor graphs that "support" the trajectories). We accomplish this through a two-stage approach where we first identify the largest set S of alteration events that follow a specific seed alteration event s in a given number of tumors (i.e. for which a consensus phylogeny could be established); and then we infer the tree topology T_S of these alteration events which features the largest number of pair-wise temporal orderings from the involved tumor samples by choosing the one with the largest average node depth (the two goals are equivalent since the number of node pairs with an ancestor-descendant relationship in the tree is equal to the sum of the path lengths between each node of the tree and the root). We briefly summarize these steps below.

We call a set of nodes S from $\{G_p\}(p \in \mathcal{P})$, with a designated seed node $s \in S$, an **evolutionarily conserved alteration set**, if for every node $v \in S$, there is a *sufficient* number of tumors $p \in \mathcal{P}$ such that G_p includes a directed path from s to v consisting only of nodes in S. Intuitively, an evolutionarily conserved alteration set S is a set of nodes that temporally follow a *founding driver alteration* s in an interdependent manner: i.e. there is a (temporal) path from s to each node v in a significant number of tumors that consists solely of nodes in S. In this paper we focus on identifying the *largest* node set S where each node $v \in S$ is connected to the seed node s through a path of nodes in S in at least a user-specified number of tumors. The set is identified through solving the *Maximum Path-Conserved Subgraph Identification (MPCSI)*, a combinatorial optimization problem, which is defined and solved by a polynomial time algorithm in Section 2.1.

The nodes in an evolutionarily conserved alteration set S may not be temporally ordered the same way across all tumors (with the exception of the seed node s which, by definition, precedes all others). Thus, we aim to order nodes of S into a consensus phylogeny tree T_S of compatible conserved evolutionary trajectories that include all nodes in S. In order to achieve the largest possible temporal ordering of nodes, T_S will be the consensus phylogeny tree which has the maximum total node depth – i.e. the total distance from s to every other node will be maximum possible. We call T_S the *conserved evolutionary trajectory tree* of S and the problem of computing T_S, the **Constrained Maximum Depth Spanning Tree problem**, for which we describe an integer linear programming (ILP) formulation in Section 2.2.[3] Note that T_S is constrained in the sense that each path (and thus each edge) in T_S must be conserved in a significant number of tumors. In the Supplementary Materials, we describe an exact, polynomial-time, algorithm to solve what we call the *Maximum Depth Spanning Tree problem* where the output tree T'_S does not need to satisfy constraints that we impose in the ILP formulation. Instead, T'_S must satisfy the relaxed constraint that each *edge* must be conserved in a significant number of tumors. When the average depth of T_S, or the number of nodes in S are small, the T'_S, the solution to the maximum depth spanning tree problem, is typically identical to T_S, indicating it is possible to use this polynomial-time algorithm to "compute" T_S efficiently.

2.1 Largest Evolutionarily Conserved Alteration Set

As mentioned in the previous section, CONETT identifies the largest evolutionarily conserved alteration set by solving the Maximum Path-Conserved Subgraph Identification problem (MPCSI). In this section, it is described how to solve the MPCSI problem on an set of arbitrary directed graphs. Formally, given a set of graphs $\{G_p\}(p \in \mathcal{P})$, let $\mathcal{G} = (\cup_{p\in\mathcal{P}} V(G_p), \cup_{p\in\mathcal{P}} E(G_p))$ be the smallest supergraph of all graphs, $G_p(p \in \mathcal{P})$ (so that each G_p is a subgraph of $\mathcal{G}$). The MPCSI problem is then defined as follows.

Maximum Path-Conserved Subgraph Identification problem (MPCSI): Given $\{G_p\}$ $(p \in \mathcal{P})$, their smallest supergraph $\mathcal{G}$, a seed node $s \in \mathcal{G}$, and a parameter t, find the largest set of nodes S of $\mathcal{G}$ such that for each node $v \in S$ there are at least t graphs G_p which include a path from s to v composed only of nodes in S.

[3] Even though it is possible to define a consensus evolutionary trajectory tree for the complete set of nodes across the tumor cohort, it may not be feasible to compute this tree since the number of such nodes could be prohibitively high. Focusing only on the nodes in the solution to the evolutionarily conserved alteration set problem (and pruning the vast majority of nodes in the tumor cohort), it becomes feasible to compute a consensus evolutionary trajectory tree.

To solve MPCSI problem for a given seed s, CONETT uses an iterative algorithm which starts with $S = \{s\}$. In each iteration, the algorithm considers adding to S a new node v to which there is a directed edge from any node in S in at least t graphs. Node v is only added to S if there are at least t graphs that include a path from s to v consisting only of nodes already in S. For finding out whether v satisfies this condition, CONETT constructs a data structure $P(v)$ for each node $v \in \mathcal{G}$, which maintains the set of graphs that have a directed edge from any node in S to v in the form of a bitmap. Naturally $P(s)$ is initially equal to the set of all graphs where s is present. For every other node v, $P(v)$ is initially equal to the set of tumor graphs that include edge (s, v). The algorithm terminates if there is no $v \in (\mathcal{G} - S)$ such that $|P_v| \geq t$.

Otherwise it identifies any node $u \in (\mathcal{G} - S)$ such that $|P_u| \geq t$ and adds it to S. Then it checks for each node $v \in S$ for which $(u, v) \in E(\mathcal{G})$, whether the addition of u to S adds any new graphs to P_v. If this is the case, the algorithm inserts the new graphs to P_v and recursively propagates the new graphs to any neighbour $w \in S$ of v that does not already have them in P_w. Figure S-1 illustrates the necessity of the recursive path propagation within S. Finally, the algorithm updates P_z for every node $z \in (\mathcal{G} - S)$ for which there is an edge $(v, z) \in E(\mathcal{G})(v \in S)$.

The correctness of the algorithm follows from the invariant that at each iteration, the algorithm considers paths from s to nodes that are either in S or are neighbours of nodes in S, thereby ensuring the existence of at least t paths (each in a separate graph). We also note that for any given $v \in S$, the total number of times that P_v is updated by the algorithm can not be more than the number of graphs. Furthermore the total number of times the algorithm can visit v is no more than the number of edges in $\mathcal{G}$. As a result, the running time of the algorithm is $O(|V(\mathcal{G})| \cdot (|\mathcal{P}| + |E(\mathcal{G})|))$.

2.2 Ordering nodes into a Conserved Evolutionary Trajectory Tree

The MPCSI formulation described in Section 2.1 produces the maximum-size evolutionarily-conserved alteration set S that consists of all alterations for each of which there is a path (composed of nodes in S) from the seed alteration s, in at least t of the tumor graphs. However, the formulation does not require these paths to be the same in all involved tumor graphs; there could be differences among the tumor graphs with respect to the temporal order of their shared alterations. On the other hand, it is possible to identify evolutionary trajectories, i.e. directed paths $s, v_1, v_2, \ldots, v_k$ formed by nodes in S, each originating from s and conserved in at least a user defined number of tumor graphs. In case that two or more such conserved evolutionary trajectories are consistent with respect to the temporal order of every pair of nodes, one can come up with a single consensus evolutionary tree where each such trajectory is included as a path. Our goal here is to construct a consensus evolutionary tree T_S of consistent trajectories which together include *all* nodes in S (thus being a spanning tree) and produce the maximum number of ordered node pairs, based on partial information from each involved tumor graph; we call this tree the *conserved evolutionary trajectory tree* of S. The conserved evolutionary trajectory tree T_S has to have the maximum possible total depth (the distance between s and every other node in S) among all consensus evolutionary trees that include all nodes in S, since the number of ancestor-descendant pairs is equal to the sum of node depths; thus being a *maximum-depth* spanning tree. This definition also agrees well with the recent analysis on clear cell Renal Cell Carcinomas, the main cancer type we focus on in this paper, from the Pan Cancer Analysis of Whole Genomes (PCAWG) project, which revealed that the time required for sub-clonal diversification is much longer compared to that for many cancer types [11].

To ensure that the ordering of nodes in the resulting maximum-depth spanning tree T_S agrees with the involved tumor graphs of S, CONETT imposes constraints on possible ancestor-descendant relationship of all pairs of nodes. For that, it computes, in a preprocessing step, a *weight* function for each pair of nodes in S as a measure of confidence for their temporal ordering as follows. Let $\mathcal{T}(S)$ be the set of tumor graphs that include the node s, and let $E(S)$ be the set of edges $\{(u, v) \mid u, v \in S\}$, which are present in at least one tumor graph in $\mathcal{T}(S)$. For each directed edge $(u, v) \in E(S)$, let $\mathcal{T}(S)_{uv}$ represent the subsets of tumor graphs in $\mathcal{T}(S)$ which contain paths from s to both u and v, and let $P_{u,v} \subseteq \mathcal{T}(S)_{uv}$ be the subset of tumor graphs in $\mathcal{T}(S)_{uv}$ that include the edge (u, v). Likewise, let $R_{u,v} = P_{v,u}/P_{u,v}$ be the subset of tumor graphs in $\mathcal{T}(S)_{uv}$ that include the directed edge (v, u) but not (u, v). Then the weight of (u, v) is set to $W_{u,v} = \frac{|P_{u,v}| - |R_{u,v}|}{|\mathcal{T}(S)_{uv}|}$. By allowing only edges whose weight is at least a specific threshold δ, we obtain an edge set $E(S, \delta)$ of high-confidence possible temporal orderings between pairs of nodes in S (setting $\delta = 1.00$ would ensure that not a single tumor graph in $\mathcal{T}(S)$ disagrees with any ancestor-descendant pair in the tree). Additionally, CONETT offers the user the ability to use the additional constraint that each path from s to any node v (given by the tree T_S) must be included in at least a user defined fraction of

Maximize: $\sum_{v \in S} D_v$

s.t.

(1)	$s:$	$D_s = 0$
(2)	$\forall (v \in S):$	$X_{v,s} = 0$
(3)	$\forall (u,v):$	$PD_{u,v} \leq D_u$
(4)		$PD_{u,v} \leq X_{u,v} \cdot \lvert S \rvert$
(5)	$\forall (v \neq s):$	$D_v = 1 + \sum_{(u,v) \in E(S,\delta)} PD_{u,v}$
(6)	$\forall (v \neq s):$	$\sum_{(u,v) \in E(S)} X_{u,v} = 1$
(7)	$s:$	$\sum_{(s,u) \in E(S)} f(e_{s,u}) = \lvert S \rvert - 1$
(8)	$\forall (v \neq s):$	$\sum_{(u,v) \in E(S)} f(e_{u,v}) =$ $1 + \sum_{(v,u) \in E(S)} f(e_{v,u})$
(9)	$\forall (u,v):$	$\lvert S \rvert \cdot X_{u,v} \geq f(e_{u,v})$
(10)		$X_{u,v} \leq f(e_{u,v})$
(11)	$\forall (u,v) \notin E(S,\delta):$	$X_{u,v} = 0$
(12)	$\forall (u,v,w):$	$PA_{u,v,w} \leq A_{u,v}$
(13)		$PA_{u,v,w} \leq A_{v,w}$
(14)		$PA_{u,v,w} \geq A_{u,v} + A_{v,w} - 1$
(15)	$\forall (u,v):$	$A_{u,v} \geq X_{u,v}$
(16)		$\lvert S \rvert \cdot A_{u,v} \geq \sum_{w \in S} PA_{u,v,w}$
(17)		$A_{u,v} \leq X_{u,v} + \sum_{w \in S} PA_{u,v,w}$
(18)	$\forall (u,v) \notin E(S,\delta):$	$A_{u,v} = 0$
(19)	$\forall (p \in \mathcal{T}_s):$	$PTH_s^p = 1$
(20)	$\forall (p \notin \mathcal{T}_s):$	$PTH_s^p = 0$
(21)	$\forall (u,v,p):$	$PTHA_{u,v}^p \leq PTH_u^p$
(22)		$PTHA_{u,v}^p \leq X_{u,v}$
(23)		$PTHA_{u,v}^p \geq PTH_u^p + X_{u,v} - 1$
(24)	$\forall (v,p):$	$PTH_v^p = \sum_{(u,v)} \alpha_{u,v} \cdot PTHA_{u,v}^p$
(25)	$\forall (v \neq s):$	$\sum_{p \in \mathcal{T}_s} PTH_v^p \geq \frac{1}{3} \cdot \lvert P_{s,v} \rvert$

Figure 2: The ILP formulation for inferring the conserved evolutionary trajectory tree

tumor graphs in $\mathcal{T}(S)$ that include the node v (we have used 1/3 in our experiments).

CONETT's ILP formulation in Figure 2 specifies the *constrained maximum-depth spanning tree* T_S through binary variables $X_{u,v}$ defined for each edge $(u,v) \in E(S)$ which will be set to 1 if (u,v) is in T_S (is a tree edge), and set to 0 otherwise. For each pair of nodes u and v in S there is a variable $A_{u,v}$, which will be set to 1 if u is an ancestor of v in T_S, and 0 otherwise. For each tumor p and each node $v \in S$, there is a variable PTH_v^p which will be set to 1 if the exact path between s to v in T_S is also present in tumor graph G_p. Lastly, for each node $v \in S$ there is a variable D_v representing the distance between s and v in T_S. The objective of the ILP is to maximize the total depth of nodes in S (and since $|S|$ is fixed, the average depth).

Now we go through the constraints of the ILP formulation. First, the seed node s is set to have depth 0 by constraint (1), and it cannot be a child of any other node due to constraint (2). Every other node v has depth that is 1 more than the depth of its parent node u; the edge to node v from its parent u is determined by auxiliary variables $PD_{u,v}$: $PD_{u,v}$ is set to 0 if u is not the parent of v – by constraint (4); it is set to D_u if u is the parent of v - because of constraints (3) and (5).

To ensure that T_S is a tree, we require that each node can have only one parent; this is ensured by constraint (6). We enforce that the nodes of T_S form a connected component of size $|S|$ by considering a fictitious *network flow* originating at seed node s of $|S| - 1$ units – by constraint (7). The flow loses 1 unit at each node - by constraint (8). If there is a positive amount of flow through the edge (u,v), then it is an edge in T_S – by constraint (9). If there is no flow through the edge (u,v) then it is not in T_S – by constraint (10). Finally, constraint (11) does not allow a tree edge between a pair of nodes below a

certain user-defined high-confidence treshold δ.

The set of constraints (12–18) ensures that all ancestor-descendant pairs of nodes in T_S represent high-confidence temporal orderings. Specifically, constraint (18) eliminates from consideration all such pairs with insufficient support from tumor graphs. We introduce auxiliary variables $PA_{u,v,w}$, set to 1 if u is an ancestor of w and w is an ancestor of v and set to 0 otherwise – by constraints (12), (13) and (14). A parent of a node is its ancestor by default, by constraint (15). Constraints (16) and (17) impose transitivity on ancestorial relationships: if u is an ancestor of v and v is an ancestor of w then u must be an ancestor of w.

The set of constraints (19–25) ensures that the "consensus" topology of T_S has sufficient support from the tumor graphs in form of sharing complete paths towards nodes in the tree. Constraint (25) requires each trajectory (or path) from the root s to a node v in the tree to be conserved in at least a user specified fraction (we have used 1/3) of tumor graphs $\mathcal{T}(S)$ that include node v. Similar to the use of auxiliary variables $PD_{u,v}$ and $PA_{u,v}$ above, we introduce auxiliary variables $PTHA^p_{u,v}$, set to 1 if and only if there is a path from s to u in tumor graph G_p, and (u, v) is a tree edge of T_S – by constraints (21), (22) and (23). A tree path from s to v exists in a tumor graph G_p if G_p both includes the path from s to u as well as the tree edge from u to v - by constraint (24). Here, the indicator constant $\alpha_{u,v}$ takes the value 1 if edge $(u, v) \in G_p$, and 0 otherwise. By default, the seed node s has a path in tumor graph G_p if it is present in it – by constraints (19) and (20).

Note that only $X_{u,v}$ variables are required to be binary. All other variables can be continuous (with the addition of constraints that previously-binary variables can not be larger than 1), which reduces the complexity of the model and the running time and space required to solve it.

2.3 Assessment of the statistical significance of MPCSI solutions

In order to assess the statistical significance of a node set S computed by CONETT as a solution to the MPCSI problem, we perform a variation of the standard permutation test on the input data. Let $N_{a,p}$ represent the number of nodes (genes and chromosome arms) altered by alteration type a in tumor graph G_p, and let $F_{v,a}$ represent the number of tumors in which node v occurs altered by a. [4] For each tumor $p \in \mathcal{P}$, and each alteration type a, we randomly draw (without repetitions) $N_{a,p}$ nodes, and mark them as altered by a, in case they were altered by a in at least one tumor graph in the original input. The nodes are randomly drawn from a distribution based on their recurrence frequency, i.e. each node v is drawn with the probability $\frac{F_{v,a}}{\sum_{u \in V(\mathcal{G})} F_{u,a}}$. That way, the average recurrence frequency of each node and alteration type in the permuted data is equal to the original recurrence frequency, preserving hidden inter-dependencies among somatic mutations and copy number alterations that were the cause of the observed recurrence frequencies and evolutionarily conserved alteration sets. Additionally, consistent with observations on the TRACERx ccRCC data set we use in this paper, special care was devoted to ensuring that a copy number event affecting a whole chromosome does not appear together with a copy number event altering either of its arms within the same tumor. Lastly, the randomly chosen altered nodes replace the original nodes of the input tumor graph G_p after being subject to a random permutation.

The above process is repeated 1000 times. In order to estimate the statistical significance of a subgraph S of size k that was obtained by running CONETT with parameter t, we run the method on each generated permutation of the input data using the same value of t, and record the number of times that the resulting subgraph has a size of k or greater. That number, divided by the total number of experiments (1000 for our case), gives us the empirical p-value estimate for S.

3 Results

3.1 Data and pre-processing

3.1.1 Clear-cell renal cell carcinoma (ccRCC)

We obtained ccRCC data from the TRACERx renal study [29], which gathered multi-region sequencing data from 100 tumors. The study collected non-silent mutation data in about 110 genes that are deemed to

[4] As mentioned earlier, CONETT can be set to distinguish alteration types on individual genes or chromosomal arms. We have not treated distinct alteration types differentially in the algorithms described above, however in case certain genes/chromosomal arms included in S are not strongly associated with a specific alteration type, (across the tumors) our empirical test can be set to differentiate alteration types as described here. Otherwise a can be set to the same value for all nodes.

be high-confidence ccRCC driver genes, obtained through a sequencing panel from which single nucleotide variants (SNVs), dinucleotide variants (DNVs) and small insertion and deletions (INDELs) were inferred. The study also detected somatic copy number alterations (SCNAs), which were reported only on the level of chromosome arms or whole chromosome – provided at least 50% of the chromosomal arm had a copy number alteration. Then, in each tumor, the study clustered all the alterations and reconstructed a phylogenetic tree of clonal hierarchy.

For every somatically altered gene and copy number altered chromosomal arm that is found in a particular tumor's phylogenetic tree, we created a node labeled by the name of the gene (or chromosome) and the type of alteration that affects it. Since, in this dataset, SNVs, DNVs and small insertion and deletion events result in inactivation of the genes that they affect, we treated them all as a single "inactivating" alteration type. Then, we build a directed graph on this set of nodes for a particular tumor, such that a directed edge is drawn from node u to node v if the clonal node in the phylogenetic tree which carries the alteration affecting node u is an ancestor of the clonal node in the phylogenetic tree which carries the alteration affecting node v. In the case of alteration events A and B that belong to the same clone, we draw a bi-directed edge (two parallel directed edges, in opposite directions) between the nodes that are associated with A and B. In 5 tumor trees, there was immediate branching right after the germline node. To ensure that there are no cross links in our tumor graphs (which can happen if a gene is present in both branches), leading to false precedence relationship between ancestors of the cross link source and descendants of the cross link destination, we split such trees at the root and treat them as two separate tumor phylogenetic trees.

As a result, we obtained tumor graphs from 105 "tumors" (after splitting the 5 trees with branching at germline node), containing 242 different nodes in total, representing driver genes affected by inactivating sequence alterations, and copy number gains and losses on chromosomes – both on whole chromosome and arm level. Figure S-2 shows recurrence frequencies of the 242 events; the y axis shows the recurrence frequency (in log scale), and events are sorted along the x axis based on their y values for clarity purposes. The figure reveals the sparsity of the data – more than a third of genes that carry sequence alterations each occur in only a single tumor, and more than a half occur in at most 2. Among more recurrently altered genes and chromosome arms (those that are found altered by the same alteration type in at least 10 tumors), there are 19 (35%) copy number gain-altered, 26 (50%) copy number loss-altered and 12 (9%) sequence-altered nodes.

3.1.2 Non-small-cell lung cancer (NSCLC)

We obtained data from 99 NSCLC patients from [4]. Originating from whole-exome sequencing of multiple spatially separated regions from the TRACERx lung cancer study [17], [4] provide cancer cell fraction (CCF) values, as well as clustering information, for single nucleotide variations and focal copy number alterations in 79 putative driver genes of NSCLC (we note that there was no case of a gene being affected by alterations of different types in any patient). Using CITUP [22], we inferred phylogenetic trees with the clustering information provided by [4], and for each patient we selected the tree with the minimum error score (which was always unique). We then construct tumor graphs from the phylogenetic trees similarly to the procedure used to generate tumor graphs in the ccRCC data (Section 3.1.1).

Contrary to the ccRCC data, there was no provided information about copy number alterations on a chromosome level in this dataset. Since in [4], the analysis was performed on the gene level, ignoring the specific alteration type (or having just one universal alteration type), we took a similar approach to make our results comparable by merging SNVs and gene deletions into a single "inactivating" alteration type. However, we kept gene amplifications as separate alteration type since the downstream effect significantly differs from SNVs and deletions.

3.2 Conserved evolutionary trajectories in ccRCC

We first used CONETT to identify statistically significant and highly-conserved alteration sets and associated trajectories in the TRACERx renal study (see Section 3.1.1), by considering as a seed node each gene and chromosomal arm that are found to be altered in at least 10 tumors. For each seed, in this experiment we set the parameter t in the MPCSI formulation to a high fraction of the seed's recurrence frequency ($\geq 48\%$). From the evolutionarily conserved alteration sets, we constructed conserved evolutionary trajectory trees, using an edge score threshold of $\delta = 0.85$ (see Section 2.2) and empirically assessed the statistical significance of these trajectory trees (using methodology described in Section 2.3). Below we report on the 10 most significant evolutionarily conserved alteration sets.

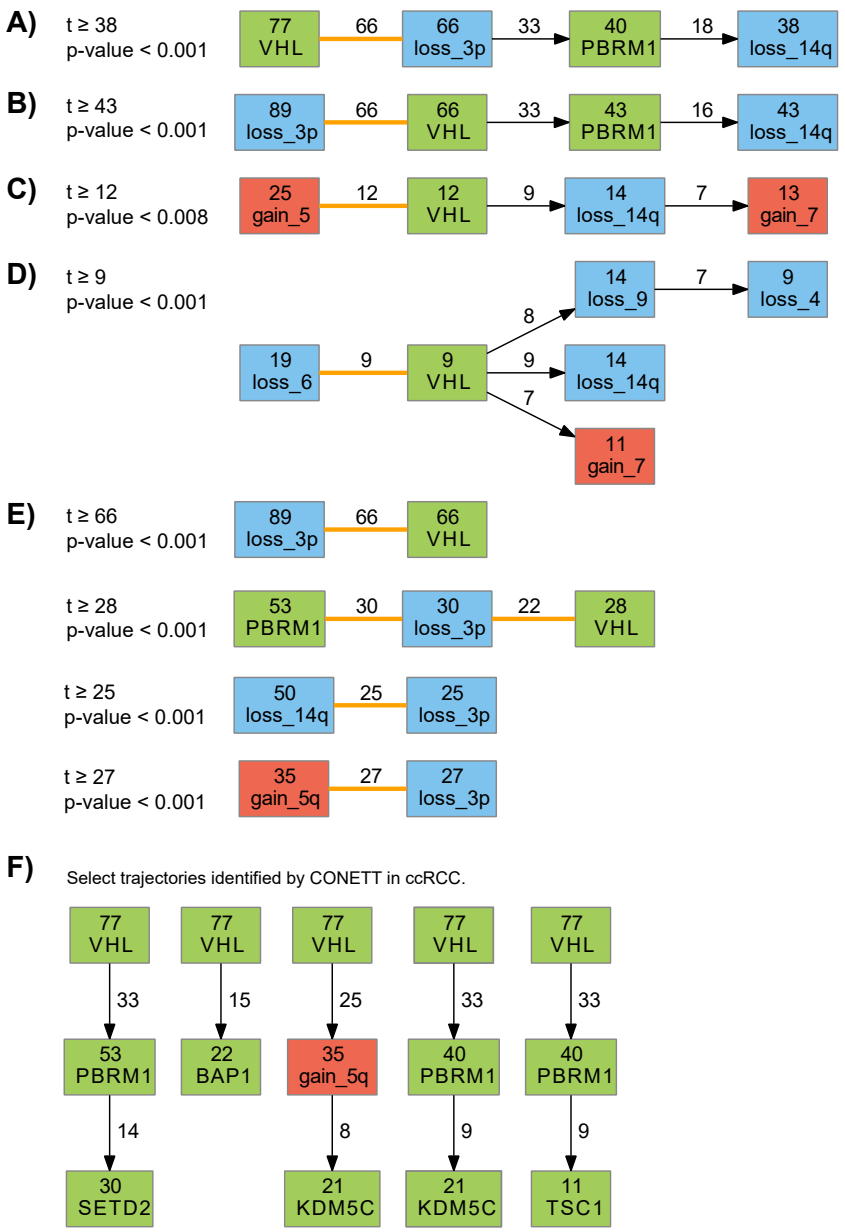

Figure 3: **Conserved evolutionary trajectories identified through CONETT in renal cancer.** The numerical value in each node is the total number of tumor graphs that include any directed path from the seed node to that node involving only the nodes in the conserved alteration set. The value on each edge is the number of tumor graphs that include the directed path in the tree, starting from the seed node and ending with that edge. Each undirected edge (orange) marks a pair of co-clonal alterations that can not be temporally ordered. **A, B, C, D)** Conserved evolutionary trajectory trees constructed for highly-conserved alteration sets. **E)** Statistically significant highly-conserved, co-clonal sets, from which it is not possible to reconstruct evolutionary trajectories. **F)** Select trajectories from conserved evolutionary trajectory trees seeded by the germline node, *VHL* and loss on chromosome 3p.

Four of these sets contain 2 or 3 nodes, involving genes and copy number altered chromosome arms that are co-clonal (Figure 3E): these include sequence alterations on driver gene VHL and copy number loss on chromosome 3p, which are known to be founder alterations in ccRCC. Conserved evolutionary trajectories seeded either by sequence alterations on VHL or copy number loss on 3p (Figure 3A-B) both include

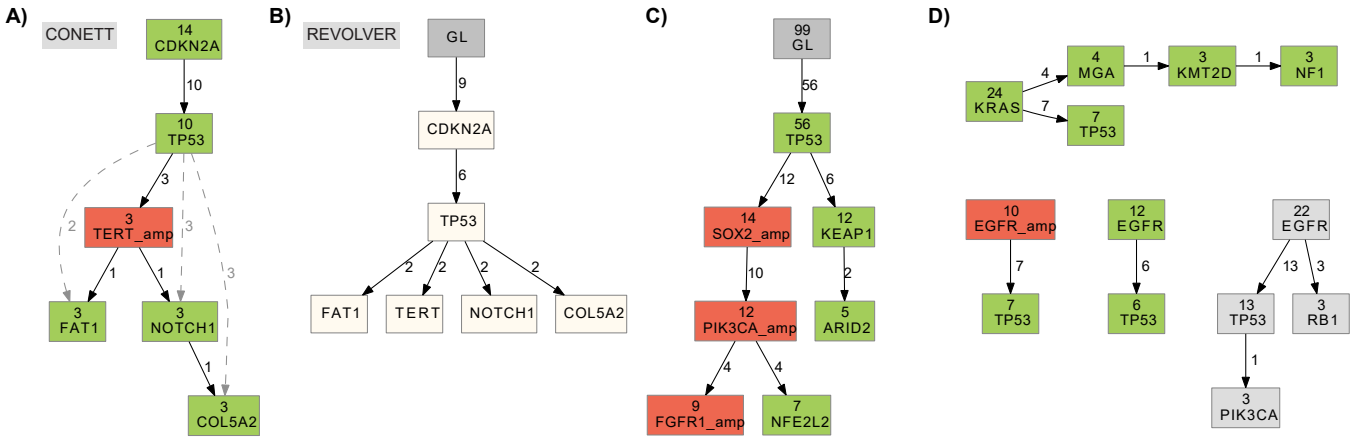

Figure 4: **Conserved evolutionary trajectories identified through CONETT in lung cancer. A** CONETT-constructed tree of conserved evolutionary trajectories seeded by an SNV on the gene *CDKN2A* ($t = 3$). It captures all the genes shown in the likely tree of cluster C5, selected by REVOLVER (shown in B). It captures more occurences of the genes and the edges and it shows additional ordering of the alterations. **B)** A likely evolutionary tree drawn from the edges in cluster C5, identified by REVOLVER. Edge weights show the number of occurrences of the edge within the cluster. **C)** Select trajectories identified by CONETT which have a *TP53* SNV as the clonal event. **D)** Trajectories seeded by recurrent NSCLC drivers *KRAS* and *EGFR*. Rightmost tree has EGFR amplifications and SNVs merged together.

sequence alterations in PBRM1 and copy number loss on chromosome 14q as highly-conserved follow-up events (p-value < 0.001), with sequence alteration in PBRM1 typically occurring before copy number loss on 14q. Copy number gain on chromosome 5 (occurring in about a quarter of all tumors) is accompanied by co-clonal sequence alteration in VHL, and is followed by copy number loss on chromosome 14q and copy number gain on chromosome 7, each observed in about half of the tumors (p-value < 0.008, Figure 3C). The remaining three conserved trajectories are all seeded by the copy number loss on chromosome 6, occuring in 19 tumors (Figure 3D shows the largest trajectory tree). It is followed by losses on 14q and 9 at a high conservation rate of 70% (p-value < 0.001), gains on chromosome 7 at 55% (p-value < 0.010) and sequence alterations of VHL and loss on chromosome 4 at 50% (p-value < 0.001). Loss on chromosome 4 follows loss of chromosome 9 in more than a third of the tumors.

Next, we looked into trees of trajectories seeded by the most recurrent events in the data set. As sequence alterations in VHL (found in 77% tumors) and copy number loss on chromosome 3p (found in 89% tumors) play a key role in initiation of ccRCC, we examined the two evolutionarily conserved alteration sets and conserved evolutionary trajectories that are seeded with either of these events and contain nodes that are altered in at least 10 tumors each ($t = 10$, both of the sets have p-value < 0.001). We also examined conserved evolutionary trajectories that start with the germline node, with same value of t, as it is present in the whole data set. The resulting trees of trajectories are shown on Figures S-4, S-5, and S-6; while Figure 3F shows a number of select trajectories from those trees.

Consistent with the trees reported in the original TRACERx renal study, the trajectories obtained by CONETT include a series of sequence alteration events *VHL* $\rightarrow$ *PBRM1* $\rightarrow$ *SETD2* observed in 14 tumors, as well as *VHL* $\rightarrow$ *BAP1* observed in 15 tumors. Additionally, CONETT also detected two conserved evolutionary trajectories terminating with a sequence alteration in gene KDM5C that collectively cover 80% of tumors where it is present: *VHL* & *loss_3p* $\rightarrow$ *gain_5q* $\rightarrow$ *KDM5C* in 8 tumors and *VHL* $\rightarrow$ *PBRM1* $\rightarrow$ *KDM5C* in 9 tumors; these trajectories were not reported in the original study. Note that KDM5C is a chromatin modifier, similar to BAP1; its dysregulation alters the activity of many other genes [12]. Additionally, *VHL* $\rightarrow$ *PBRM1* $\rightarrow$ *TSC1* is found in 9 tumors, representing 75% tumors where *TSC1* is present. CONETT also identified a number of less frequently altered genes that do not display high confidence temporal ordering. These alteration events in genes such as *MTOR* (in 12/17 tumors), *OBSCN* (in 7/12 tumors), and *MUC16* (in 10/16 tumors) all seem to branch out from an alteration in *VHL*.

Inactivations of both *PBRM1* and *BAP1* are known to rarely co-occur in ccRCC [26, 29], and trees of conserved evolutionary trajectories on Figures S-4, S-5 and S-6 always place them into different branches.

However, most of the time when they do co-occur, *PBRM1* alteration precedes *BAP1* alteration. We thus ran CONETT using *PBRM1* alteration as the seed event, to find out whether BAP1 sequence alteration appears as a conserved folow-up event, using $t = 10$ (p-value < 0.001). The resulting tree of conserved evolutionary trajectories is shown on Figure S-7. Alteration on *BAP1* was not identified as a downstream event, strengthening the hypothesis that their co-occurrence is most likely due to upstream events such as loss on 3p and inactivation of *VHL*.

Loss on chromosome 14q is the second most recurrent large-scale copy number loss event (after loss on 3p), occurring in 50 tumors. However, contrary to loss on 3p which regularly appears as the earliest alteration event in tumors, loss on 14q is always placed later, after both VHL and loss on 3p. It is often preceded by inactivation of *PBRM1*, although not in every single tumor (Figure S-7 shows that loss on 14q follows inactivation of *PBRM1* in 22 tumors). There is also strong evidence of precedence of gains on chromosomes 5 and 5q (which are themselves mutually exclusive) to loss on chromosome 14q – trajectories seeded by loss on 14q do not detect gains on chromosome 5 or 5q as downstream events (Figure S-8), but trajectories seeded by gains on 5 and 5q do detect loss on 14q as a downstream event (Figure S-9). Additionally, Figure S-9A shows presence of mutual exclusivity of losses on chromosome 3p and 14q in tumors that contain gain on chromosome 5, whereas that relationship is not noticed in tumors that contain gain on 5q (Figure S-9B). This pattern of mutual exclusivity was not reported in the original ccRCC TRACERx study.

3.3 Conserved evolutionary trajectories in NSCLC

We used CONETT to identify conserved evolutionary trajectories in the TRACERx non-small-cell lung cancer study (see Section 3.1.2). This data was analyzed by [4], allowing us to compare the CONETT-identified trajectories with the edges in clusters identified by their method.

In [4], the authors highlighted evolutionary trajectories constructed from edges in cluster C5 (Figure 4B). Using CONETT, with the SNV-altered gene *CDKN2A* as the seed, we were able to recover the full topological order of the genes in cluster C5 (Figure 4A), as well as additional ancestor-descendant relationship between amplification on *TERT* and SNVs on *FAT1* and *NOTCH1*. The CONETT-identified tree captures more occurrences of tree edges *CDKN2A* $\rightarrow$ *TP53* and *TP53* $\rightarrow$ *TERT*, as well as edges from *TP53* towards *NOTCH1* and *COL5A2*, which were tree edges in cluster C5 but are just forward edges in CONETT (dashed gray lines); due to CONETT trajectories capturing additional ancestor-descendant relationship between amplification of *TERT* and SNVs on sibling nodes *FAT1* and *NOTCH1*.

Additionally, we used CONETT to construct trees of trajectories with known early drivers (as well as the germline node) as seeds (Figures S-10, S-11), using the edge threshold value $\delta = 1.00$. Figure 4C shows select conserved trajectories which are supported by both trees. The trajectories capture *TP53* $\rightarrow$ *SOX2* & *PIK3CA* amplifications $\rightarrow$ *NFE2L2* (in 4/9 tumors), as well as *TP53* $\rightarrow$ *KEAP1* (in 6/12 tumors) – inactivations of *NFE2L2* and *KEAP1* may be linked to increased chemoresistance [18]. All *TP53*-seeded evolutionarily conserved alteration sets were found to be statistically significant (p-value < 0.001). Interestingly, trajectories seeded by SNV and amplification of *EGFR* did not pick up any node other than *TP53*, which they are often found in the same cluster with (Figure 4D). Even when we merge amplifications and SNVs of *EGFR*, *RB1* and *PIK3CA* just barely emerge as downstream altered genes. *KRAS*-seeded trajectories capture *MGA* and *KMT2D*, consistent with the germline-seeded tree (Figure S-10).

4 Discussion

We present CONETT, a novel computational method for combinatorial detection of conserved evolutionary trajectories for tumors. CONETT is the first such method that goes beyond just identifying clusters of common edges and directly produces a consensus phylogeny which aims to maximize the number of temporally-ordered pairs of alteration events. We introduce the notion of an evolutionarily conserved alteration set and give a polynomial time algorithm that computes it by solving the maximum path-conserved subgraph identification problem. Given a conserved alteration set, we then show how to construct a conserved evolutionary trajectory tree via an ILP formulation. We applied our method to a recently published datasets of 100 ccRCC tumors and 99 NSCLC tumors, and identified conserved evolutionary trajectories of sequence-altered genes and copy number altered chromosomal arms. Some of these conserved evolutionary trajectories involving sequence altered genes are confirmed by the original studies; CONETT also reported several additional trajectories which were not reported earlier.

As per the original studies of TRACERx renal and lung cancer data, we have not differentiated sequence-alteration types and have treated all sequence-altered genes identically; this allowed us to compare CONETT results with those of the original studies. However CONETT's framework is sufficiently general to offer the user the ability to differentiate somatic alteration types in the trajectories it identifies so as to perform in depth analysis of larger and richer data sets to be published in the near future.

We acknowledge that CONETT's ILP formulation currently constructs a consensus tree from the evolutionarily conserved alteration set, selecting a single tree topology in which each node occurs only once. That forces the method to select only a single trajectory from a given seed towards each node, even though there might be multiple different trajectories that each occur in a significant fraction of tumors (as evidenced by the two different trajectories towards *KDM5C* discovered by CONETT and shown on Figures S-4 and S-6, together covering 80% of tumors in which the gene is sequence-altered). In future, it would be useful to devise a way to simultaneously identify different conserved trajectories in tumor evolution towards the same node.

We also note that the "transitivity" of the ancestor-descendant relationship in tumor graphs that we introduced in the paper and used in the analysis simplifies the process of identification of maximum-sized evolutionarily conserved alteration sets (since all nodes that are reachable become outgoing neighbours of the seed). However, the polynomial-time algorithm that we present in this paper can be used to compute the maximum-sized evolutionarily conserved alteration set in a collection of arbitrary directed graphs. As larger and more precise data sets emerge, as well as depending on the motivation for and the specific application of the analysis, it might become desirable to use an alternative version of tumor graphs which do not exhibit complete transitivity in the connectivity of the nodes, but instead provide possible edges to be considered as consensus tree edges. Our robust formulation could then be directly applied to such cases.

References

[1] Christopher Abbosh, Nicolai J. Birkbak, Gareth A. Wilson, Mariam Jamal-Hanjani, Tudor Constantin, Raheleh Salari, John Le Quesne, and et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature*, 545(7655):446–451, 2017.

[2] Nicolas Alcaraz, Tobias Friedrich, Timo Kötzing, Anton Krohmer, Joachim Müller, Josch Pauling, and Jan Baumbach. Efficient key pathway mining: combining networks and omics data. *Integrative Biology*, 4(7):756–764, 2012.

[3] Anna Bomersbach, Marco Chiarandini, and Fabio Vandin. An efficient branch and cut algorithm to find frequently mutated subnetworks in cancer. In *International Workshop on Algorithms in Bioinformatics*, pages 27–39. Springer, 2016.

[4] Giulio Caravagna, Ylenia Giarratano, Daniele Ramazzotti, Ian Tomlinson, Trevor A. Graham, Guido Sanguinetti, and Andrea Sottoriva. Detecting repeated cancer evolution from multi-region tumor sequencing data. *Nature Methods*, 15(9):707–714, August 2018.

[5] Michael D. Crowther, Garry Dolton, Mateusz Legut, Marine E. Caillaud, Angharad Lloyd, Meriem Attaf, Sarah A. E. Galloway, Cristina Rius, Colin P. Farrell, Barbara Szomolay, Ann Ager, Alan L. Parker, Anna Fuller, Marco Donia, James McCluskey, Jamie Rossjohn, Inge Marie Svane, and John D. Phillips & Andrew K. Sewell. Genome-wide crisprcas9 screening reveals ubiquitous t cell cancer targeting via the monomorphic mhc class i-related protein mr1. *Nature Immunology*, January 2020.

[6] Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun Ho Jang, Lincoln Stein, and Quaid Morris. Phylowgs: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome biology*, 16(1):35, 2015.

[7] Ramon Diaz-Uriarte and Claudia Vasallo. Every which way? On predicting tumor evolution using cancer progression models. *PLoS computational biology*, 15(8):e1007246, aug 2019.

[8] Jesse Eaton, Jingyi Wang, and Russell Schwartz. Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics*, 34(13):i357–i365, 2018.

[9] Mohammed El-Kebir, Layla Oesper, Hannah Acheson-Field, and Benjamin J Raphael. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70, 2015.

[10] Mohammed El-Kebir, Gryte Satas, Layla Oesper, and Benjamin J Raphael. Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. *Cell systems*, 3(1):43–53, 2016.

[11] Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner, Stefan C. Dentro, Santiago Gonzalez, Daniel Rosebrock, Thomas J. Mitchell, Yulia Rubanova, Pavana Anur, Kaixian Yu, Maxime Tarabichi, Amit Deshwar, Jeff Wintersinger, Kortine Kleinheinz, Ignacio Vázquez-García, Kerstin Haase, Lara Jerman, Subhajit Sengupta, Geoff Macintyre, Salem Malikic, Nilgun Donmez, Dimitri G. Livitz, Marek Cmero, Jonas Demeulemeester, Steven Schumacher, Yu Fan, Xiaotong Yao, Juhee Lee, Matthias Schlesner, Paul C. Boutros, David D. Bowtell, Hongtu Zhu, Gad Getz, Marcin Imielinski, Rameen Beroukhim, S. Cenk Sahinalp, Yuan Ji, Martin Peifer, Florian Markowetz, Ville Mustonen, Ke Yuan, Wenyi Wang, Quaid D. Morris, Paul T. Spellman, David C. Wedge, Peter Van Loo, , and . The evolutionary history of 2,658 cancers. *bioRxiv*, 2018.

[12] Xiaoqiang Guo and Qiaoxia Zhang. The Emerging Role of Histone Demethylases in Renal Cell Carcinoma. *Journal of kidney cancer and VHL*, 4(2):1–5, 2017.

[13] Ermin Hodzic, Raunak Shrestha, Kaiyuan Zhu, Kuoyuan Cheng, Colin C Collins, and S Cenk Sahinalp. Combinatorial detection of conserved alteration patterns for identifying cancer subnetworks. *GigaScience*, 8(4), April 2019.

[14] Sayed-Rzgar Hosseini, Ramon Diaz-Uriarte, Florian Markowetz, and Niko Beerenwinkel. Estimating the predictability of cancer evolution. *Bioinformatics*, 35(14):i389–i397, July 2019.

[15] Borislav H. Hristov and Mona Singh. Network-based coverage of mutational profiles reveals cancer genes. *Cell Systems*, 5(3):221–229.e4, 2017.

[16] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome biology*, 17(1):86, 2016.

[17] Mariam Jamal-Hanjani, Gareth A. Wilson, Nicholas McGranahan, Nicolai J. Birkbak, Thomas B.K. Watkins, Selvaraju Veeriah, Seema Shafi, Diana H. Johnson, Richard Mitter, Rachel Rosenthal, Max Salm, Stuart Horswell, Mickael Escudero, Nik Matthews, Andrew Rowan, Tim Chambers, David A. Moore, Samra Turajlic, Hang Xu, Siow-Ming Lee, Martin D. Forster, Tanya Ahmad, Crispin T. Hiley, Christopher Abbosh, Mary Falzon, Elaine Borg, Teresa Marafioti, David Lawrence, Martin Hayward, Shyam Kolvekar, Nikolaos Panagiotopoulos, Sam M. Janes, Ricky Thakrar, Asia Ahmed, Fiona Blackhall, Yvonne Summers, Rajesh Shah, Leena Joseph, Anne M. Quinn, Phil A. Crosbie, Babu Naidu, Gary Middleton, Gerald Langman, Simon Trotter, Marianne Nicolson, Hardy Remmen, Keith Kerr, Mahendran Chetty, Lesley Gomersall, Dean A. Fennell, Apostolos Nakas, Sridhar Rathinam, Girija Anand, Sajid Khan, Peter Russell, Veni Ezhil, Babikir Ismail, Melanie Irvin-Sellers, Vineet Prakash, Jason F. Lester, Malgorzata Kornaszewska, Richard Attanoos, Haydn Adams, Helen Davies, Stefan Dentro, Philippe Taniere, Brendan O'Sullivan, Helen L. Lowe, John A. Hartley, Natasha Iles, Harriet Bell, Yenting Ngai, Jacqui A. Shaw, Javier Herrero, Zoltan Szallasi, Roland F. Schwarz, Aengus Stewart, Sergio A. Quezada, John Le Quesne, Peter Van Loo, Caroline Dive, Allan Hackshaw, and Charles Swanton. Tracking the evolution of non–small-cell lung cancer. *New England Journal of Medicine*, 376(22):2109–2121, June 2017.

[18] Youngtae Jeong, Jessica A. Hellyer, Henning Stehr, Ngoc T. Hoang, Xiaomin Niu, Millie Das, Sukhmani K. Padda, Kavitha Ramchandran, Joel W. Neal, Heather Wakelee, and Maximilian Diehn. Role of KEAP1/NFE2l2 mutations in the chemotherapeutic response of patients with non–small cell lung cancer. *Clinical Cancer Research*, 26(1):274–281, September 2019.

[19] Yoo-Ah Kim, Raheleh Salari, Stefan Wuchty, and Teresa M Przytycka. Module cover–a new approach to genotype-phenotype studies. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 135. NIH Public Access, 2013.

[20] Mark D M Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, Jonathan V Eldridge, Jacob L Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, Michael S Lawrence, Abel Gonzalez-Perez, David Tamborero, Yuwei Cheng, Gregory A Ryslik, Nuria Lopez-Bigas, Gad Getz, Li Ding, and Benjamin J Raphael. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, 47(2):106–114, 2014.

[21] Kamil A. Lipinski, Louise J. Barber, Matthew N. Davies, Matthew Ashenden, Andrea Sottoriva, and Marco Gerlinger. Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. *Trends in cancer*, 2(1):49–63, jan 2016.

[22] S. Malikic, a. W. McPherson, N. Donmez, and C. S. Sahinalp. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31(January):1349–1356, may 2015.

[23] Salem Malikic, Katharina Jahn, Jack Kuipers, S Cenk Sahinalp, and Niko Beerenwinkel. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nature communications*, 10(1):2750, 2019.

[24] Salem Malikic, Farid Rashidi Mehrabadi, Simone Ciccolella, Md Khaledur Rahman, Camir Ricketts, Ehsan Haghshenas, Daniel Seidman, Faraz Hach, Iman Hajirasouliha, and S Cenk Sahinalp. Phiscs: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. *Genome Research*, 2019.

[25] Yusuke Matsui, Atsushi Niida, Ryutaro Uchi, Koshi Mimori, Satoru Miyano, and Teppei Shimamura. phyC: Clustering cancer evolutionary trees. *PLOS Computational Biology*, 13(5):e1005509, May 2017.

[26] Samuel Peña-Llopis, Silvia Vega-Rubín de Celis, Arnold Liao, Nan Leng, Andrea Pavía-Jiménez, Shanshan Wang, Toshinari Yamasaki, Leah Zhrebker, Sharanya Sivanand, Patrick Spence, Lisa Kinch, Tina Hambuch, Suneer Jain, Yair Lotan, Vitaly Margulis, Arthur I Sagalowsky, Pia Banerji Summerour, Wareef Kabbani, S W Wendy Wong, Nick Grishin, Marc Laurent, Xian-Jin Xie, Christian D Haudenschild, Mark T Ross, David R Bentley, Payal Kapur, and James Brugarolas. BAP1 loss defines a new class of renal cell carcinoma. *Nature Genetics*, 44(7):751–759, June 2012.

[27] Victoria Popic, Raheleh Salari, Iman Hajirasouliha, Dorna Kashef-Haghighi, Robert B West, and Serafim Batzoglou. Fast and scalable inference of multi-sample cancer lineages. *Genome biology*, 16(1):91, 2015.

[28] Edith M Ross and Florian Markowetz. Onconem: inferring tumor evolution from single-cell sequencing data. *Genome biology*, 17(1):1–14, 2016.

[29] Samra Turajlic, Hang Xu, Kevin Litchfield, Andrew Rowan, Stuart Horswell, Tim Chambers, Tim O'Brien, Jose I. Lopez, Thomas B.K. Watkins, David Nicol, Mark Stares, Ben Challacombe, Steve Hazell, Ashish Chandra, Thomas J. Mitchell, Lewis Au, Claudia Eichler-Jonsson, Faiz Jabbar, Aspasia Soultati, Simon Chowdhury, Sarah Rudman, Joanna Lynch, Archana Fernando, Gordon Stamp, Emma Nye, Aengus Stewart, Wei Xing, Jonathan C. Smith, Mickael Escudero, Adam Huffman, Nik Matthews, Greg Elgar, Ben Phillimore, Marta Costa, Sharmin Begum, Sophia Ward, Max Salm, Stefan Boeing, Rosalie Fisher, Lavinia Spain, Carolina Navas, Eva Grönroos, Sebastijan Hobor, Sarkhara Sharma, Ismaeel Aurangzeb, Sharanpreet Lall, Alexander Polson, Mary Varia, Catherine Horsfield, Nicos Fotiadis, Lisa Pickering, Roland F. Schwarz, Bruno Silva, Javier Herrero, Nick M. Luscombe, Mariam Jamal-Hanjani, Rachel Rosenthal, Nicolai J. Birkbak, Gareth A. Wilson, Orsolya Pipek, Dezso Ribli, Marcin Krzystanek, Istvan Csabai, Zoltan Szallasi, Martin Gore, Nicholas McGranahan, Peter Van Loo, Peter Campbell, James Larkin, and Charles Swanton. Deterministic evolutionary trajectories influence primary tumor growth: TRACERx renal. *Cell*, 173(3):595–610.e11, April 2018.

[30] Igor Ulitsky, Akshay Krishnamurthy, Richard M. Karp, and Ron Shamir. DEGAS: de novo discovery of dysregulated pathways in human diseases. *PloS one*, 5(10):e13367, oct 2010.

[31] Fabio Vandin, Eli Upfal, and Benjamin J Raphael. Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*, 18(3):507–522, 2011.

[32] Lucy R. Yates, Stian Knappskog, David Wedge, James H.R. Farmery, Santiago Gonzalez, Inigo Martincorena, Ludmil B. Alexandrov, Peter Van Loo, Hans Kristian Haugland, Peer Kaare Lilleng, Gunes Gundem, Moritz Gerstung, Elli Pappaemmanuil, Patrycja Gazinska, Shriram G. Bhosle, David Jones, Keiran Raine, Laura Mudie, Calli Latimer, Elinor Sawyer, Christine Desmedt, Christos Sotiriou, Michael R. Stratton, Anieta M. Sieuwerts, Andy G. Lynch, John W. Martens, Andrea L. Richardson, Andrew Tutt, Per Eystein Lønning, and Peter J. Campbell. Genomic Evolution of Breast Cancer Metastasis and Relapse. *Cancer Cell*, 32(2):169–184.e7, 2017.

[33] Simone Zaccaria, Mohammed El-Kebir, Gunnar W Klau, and Benjamin J Raphael. Phylogenetic copy-number factorization of multiple tumor samples. *Journal of Computational Biology*, 25(7):689–708, 2018.

[34] Hamim Zafar, Nicholas Navin, Ken Chen, and Luay Nakhleh. Siclonefit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome research*, 29(11):1847–1859, 2019.

Supplementary Materials

1 Supplementary Methods

1.1 Polynomial-Time Algorithm for the Maximum Depth Spanning Tree Problem

Maximum depth spanning tree (MDST): Given an unweighted, directed graph S and a root node $s \in S$, find a spanning tree with the maximum total node depth.

We solve the MDST problem via an incremental algorithm which starts with the node s and then expands the tree, one node at a time. We use a modified version of depth-first search (DFS) to add new nodes that do not already belong to the tree, and then check whether depth of any nodes that are already in the tree can be increased via the new node (algorithms 1 and 2).

Lemma 1. *Algorithm 1 maximizes the total node depth.*

Proof. Let u be a newly added node at the current iteration of the DFS recursion. First, we show that in the current iteration node u has no better choice for a parent in the tree than the one that added it. If it did, then there would be an edge from another node in the tree towards u, and by nature of DFS, it would have already been added to the tree before. Hence this must be the first time that we have encountered node u and the current placement is the only possible one.

Next, we show that edge choices that the algorithm picks are optimal by proving the correctness of the update procedure. The proof follows via the DFS tree created by the order in which nodes are visited. Let (u, v) be an edge considered by the algorithm. Note that forward and back edges can be safely ignored. If (u, v) is a a forward edge, then v has already been processed by the recursion and $D_v \geq D_u + 1$, so no change is necessary. If (u, v) is a back edge, then v is an ancestor of u, and including this edge would create a cycle. In order to break the cycle, some node w on path $v \to ... \to u$ would have to be unlinked from its current parent and become child of an ancestor of v in order to keep the graph connected. Since any ancestor of v has depth that is smaller than depth of v by at least 1, this edge swap at best keeps the depth of all nodes in the cycle the same as before. Let (u, v) be a cross edge. If $D_v \geq D_u + 1$, then the depth of node v and its descendants would not increase by changing its parent to u. If $D_v < D_u + 1$, then by unlinking v from its current parent and placing it as a child of u, algorithm 2 would increase depths of v and all its descendants. However, since there might be an ancestor of v that u also has an edge towards, we do not want to connect u to v, but to the ancestor. Since this is true for every such ancestor, it follows that updating nodes in topologically sorted order results in optimal depth.

From the previous two paragraphs, it follows that at every iteration of the algorithm, the extension step correctly places node u in the tree, and correctly updates all remaining nodes' parent edges. Therefore, at the last iteration, we will have an optimal tree. □

Lemma 2. *Running time of algorithm 1 on a connected graph $G = (V, E)$ is $O(|V||E|)$.*

Proof. The main DFS skeleton of the code traverses the graph in a depth-first manner, adding each node that is already not in the tree in total time $O(|E| + f(n) \cdot |V|)$, where $f(n)$ is the time needed to update the tree after each newly added node. In the worst case, the new node can cause depth update for each node in the tree. Each update needs to be pushed from the updated node down to its whole subtree. In order to avoid updating same nodes multiple times, we topologically sort the nodes and push updates starting with nodes that are closest to the root and doing the same for every subsequent node whose depth isn't already updated by one of the previous nodes. Topological sort is carried out in time $O(|V|)$ via counting sort, since we already have depth values of the nodes and they are integers smaller than $|V|$. Since each node gets visited and updated at most once via normal BFS, total time that the update step takes is $O(|V| + |E|)$. Since our graph is always connected, i.e. $|E| = \Omega(|V|)$, the total running time of the algorithm is $O(|V||E|)$. □

Algorithm 1: Maximum Depth Spanning Tree

Input: Unweighted graph S, root node s
Output: Parent array P, depth array D
Result: Constructs a spanning tree which maximizes the depth of each node

```
for i = 1, 2, ..., |S| do
    P_i = -1
    D_i = -1
    A_i = false ;                    // Flag for node being in the recursion stack
end
P_s = s
D_s = 0
DFS(s, P, D, A)
return P, D
```

Algorithm 2: DFS(u, S, P, D, A)

Input: Unweighted graph S, node u, parent array P, depth array D, ancestor flag array A
Result: Constructs a spanning tree rooted in u which maximizes the depth of each node

```
A_u = true ;                         // Ancestor of any new node that it adds
/* First, update depths of nodes already in the tree                          */
U = ∅
for (u, v) ∈ E(S) do
    if (P_v > -1) and (A_v == false) and (D_v < D_u + 1) then
        U = U ∪ {v}
    end
end
sort nodes in U by their depth, in increasing order
for i = 1, 2, ..., |U| do
    if (D_{U_i} < D_u + 1) then
        updateDepthsInSubtree(U_i, S, P, D) ;                    // Simple BFS
    end
end
/* Next, discover new nodes                                                   */
for (u, v) ∈ E(S) do
    if (P_v == -1) then
        D_v = D_u + 1
        P_v = u
        A_v = true
        DFS(s, P, D, A)
    end
end
A_u = false ;                                            // No longer ancestor
return
```

2 Supplementary Figures

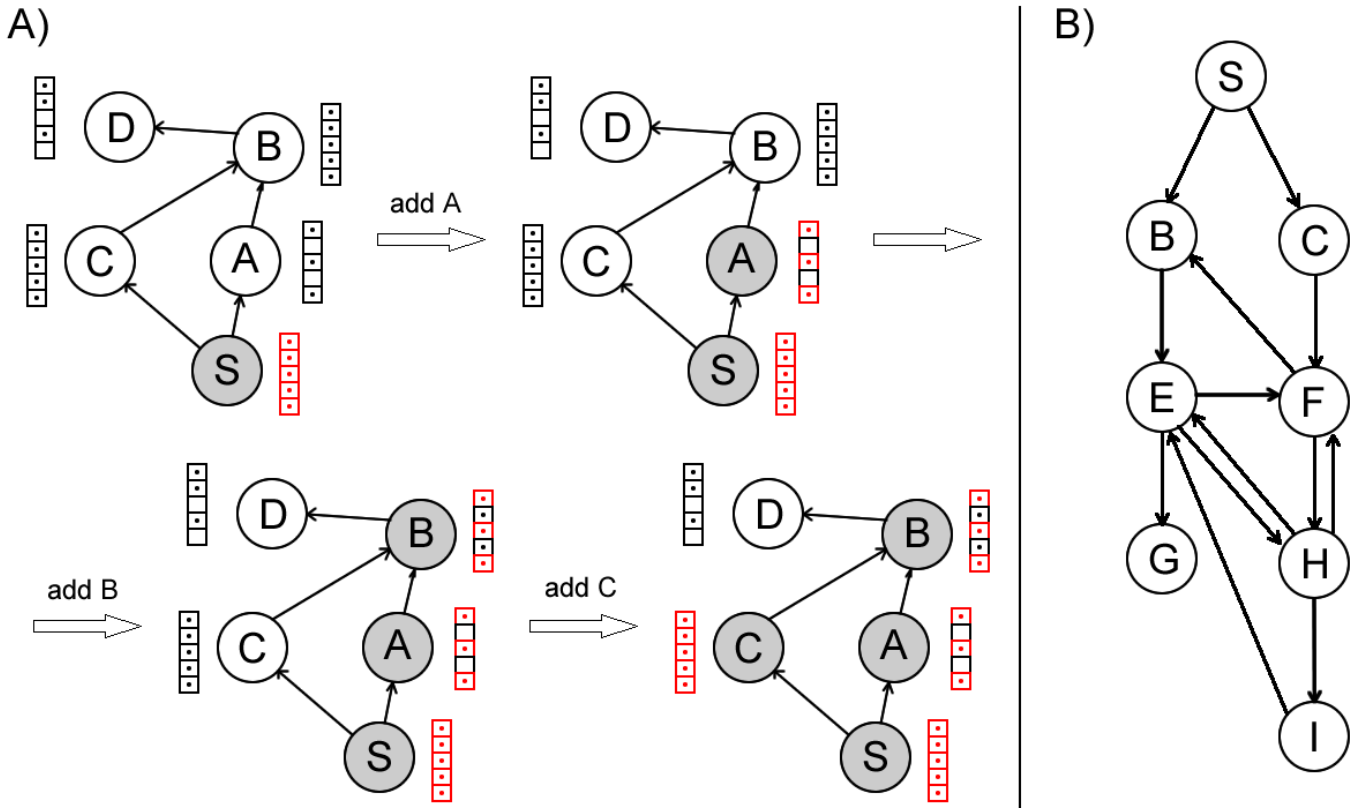

Figure S-1: **Illustration of the importance of propagation of new paths throughout the subgraph during execution of the MPCSI-solving algorithm.** Boxes next to each node represent tumors (total of 5), and there is a dot in a box if the corresponding tumor contains the alteration in the node. **A)** Let $t = 3$. If we add nodes to S in order $S \to A \to B \to C$, then D cannot be added unless C updates paths of B. Since the graph has a topological ordering, adding nodes in topological order would avoid the need for updating subgraph nodes' paths, e.g. $S \to A \to C \to B \to D$. **B)** This graph contains multiple cycles and does not have a topological ordering of its nodes. Each newly added node has to update paths of every other subgraph node before they are pushed to outside neighbours of the subgraph.

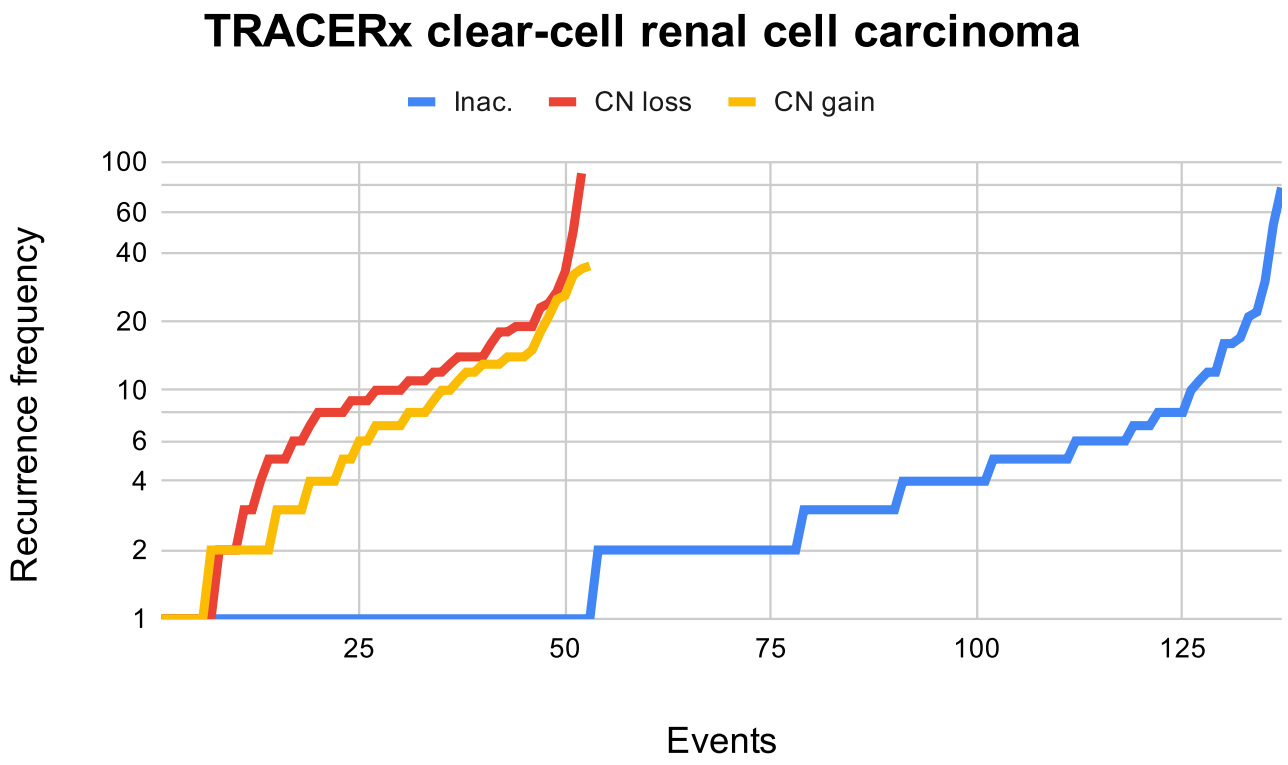

Figure S-2: **Recurrence frequency of altered genes and chromosome arms in TRACERx renal cohort.** The plot shows recurrence frequency of 137 inactivating sequence alteration (blue), 52 copy number loss (red) and 53 copy number gain (yellow) events. Events are sorted by their y value along the x axis for clarity. The y axis is in log scale.

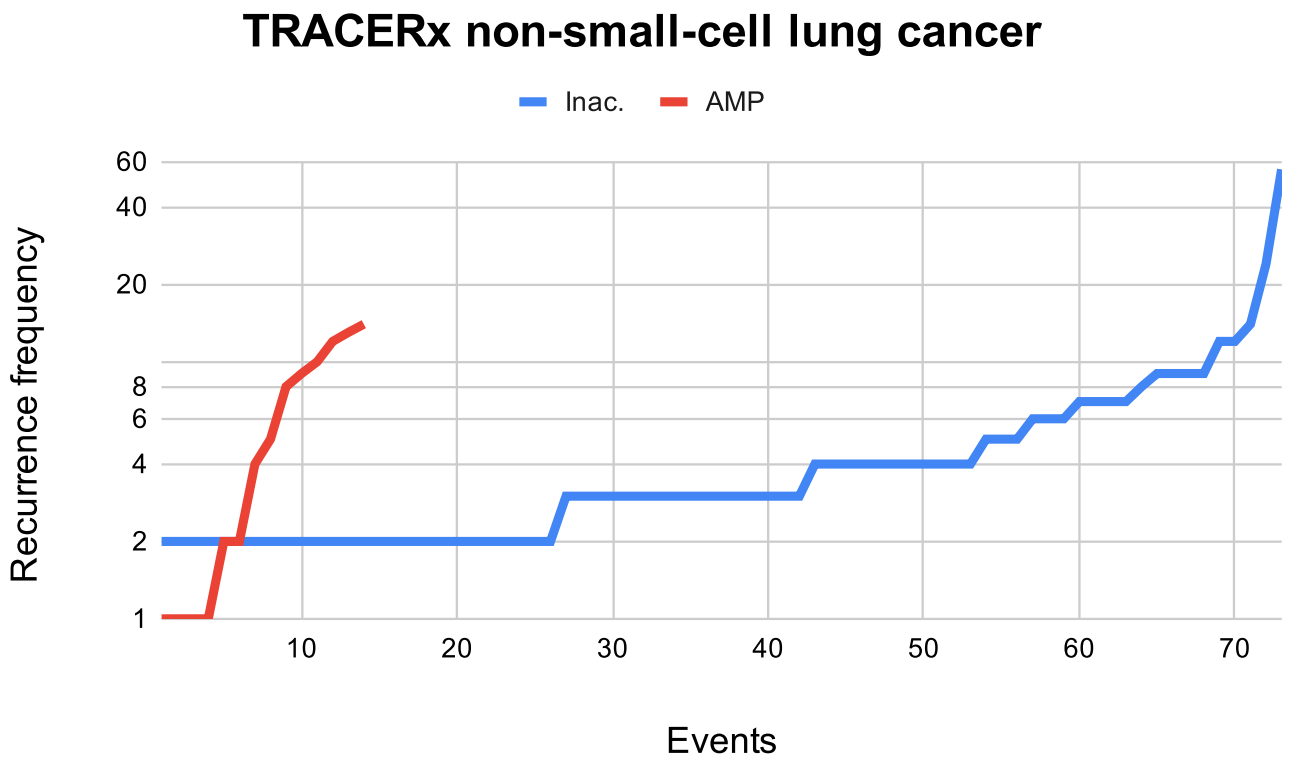

Figure S-3: **Recurrence frequency of altered genes and chromosome arms in TRACERx lung cohort.** The plot shows recurrence frequency of 73 inactivating sequence alteration (blue) and 14 gene amplification (red) events. Events are sorted by their y value along the x axis for clarity. The y axis is in log scale.

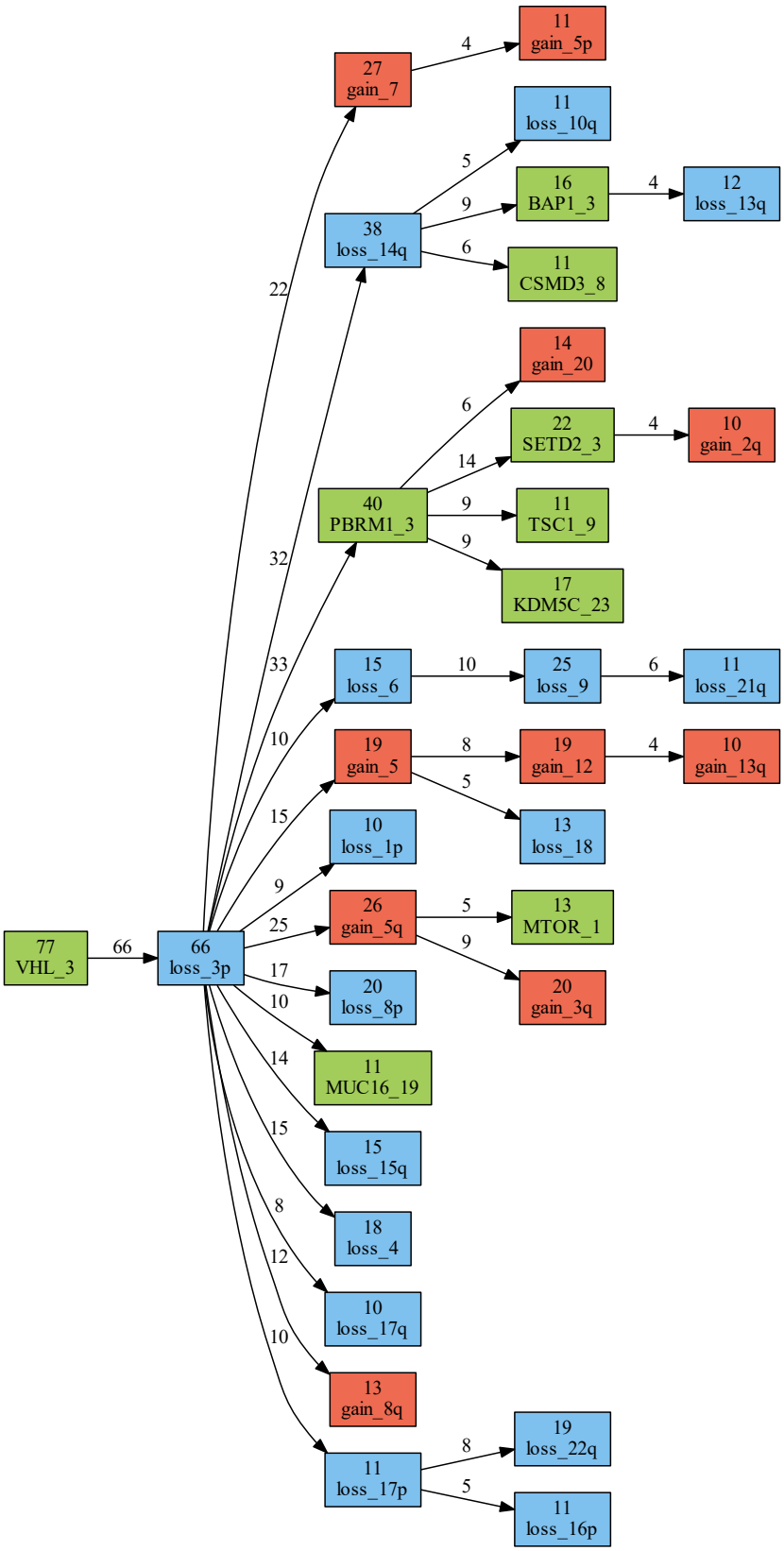

Figure S-4: **VHL-seeded tree of trajectories with $t = 10$ in TRACERx ccRCC.**

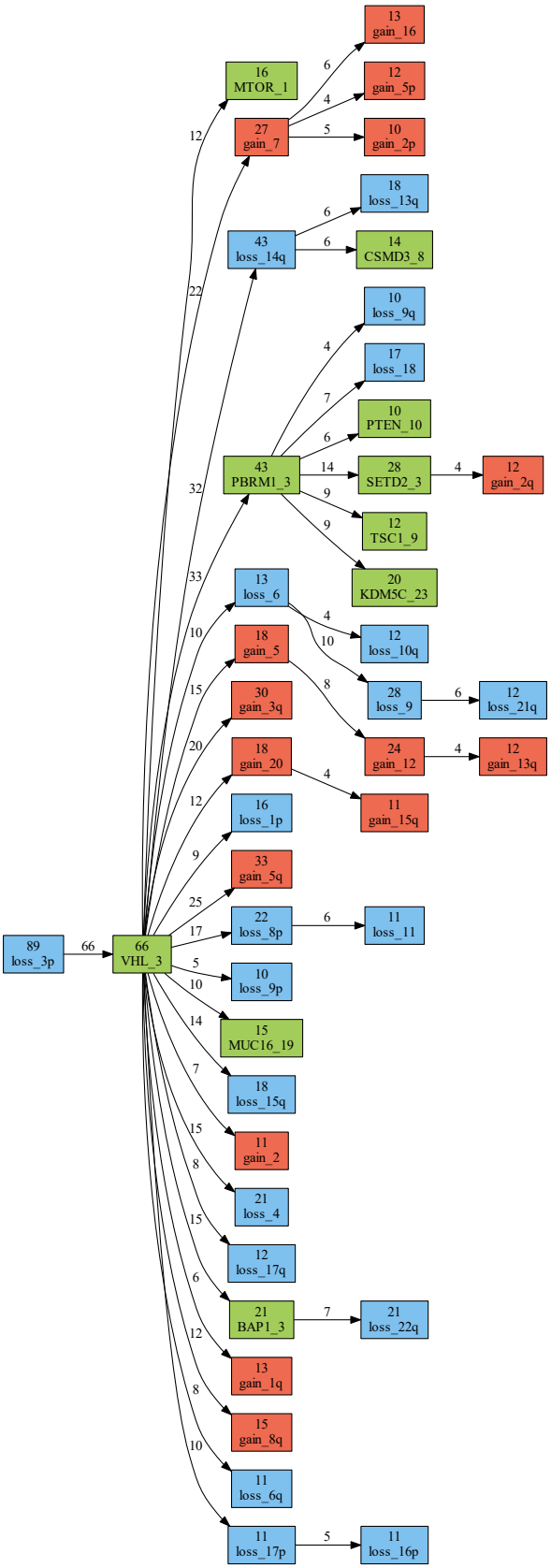

Figure S-5: **loss3p-seeded tree of trajectories with $t = 10$ in TRACERx ccRCC.**

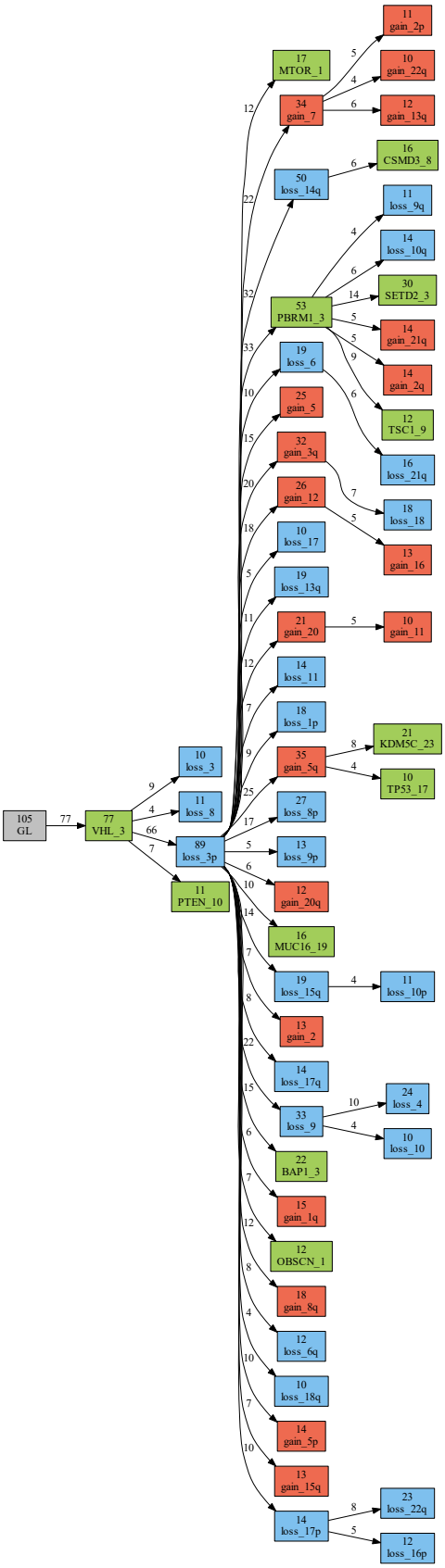

Figure S-6: **Germline-seeded tree of trajectories with $t = 10$ in TRACERx ccRCC.**

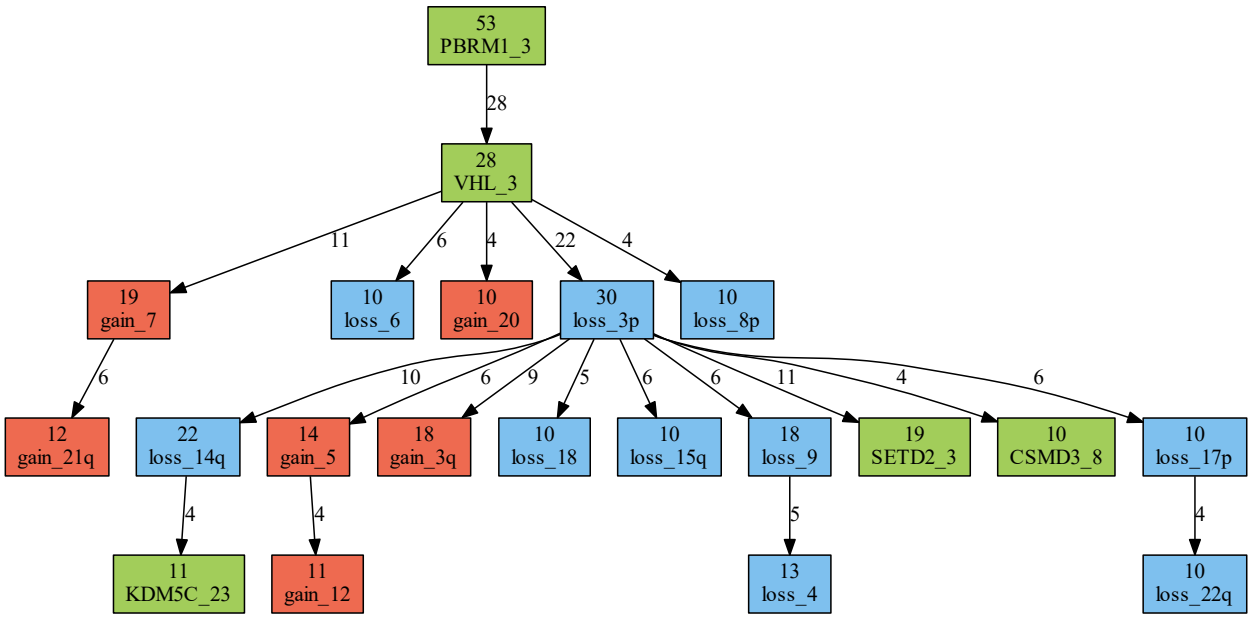

Figure S-7: **PBRM1-seeded tree of trajectories with $t = 10$ in TRACERx ccRCC.**

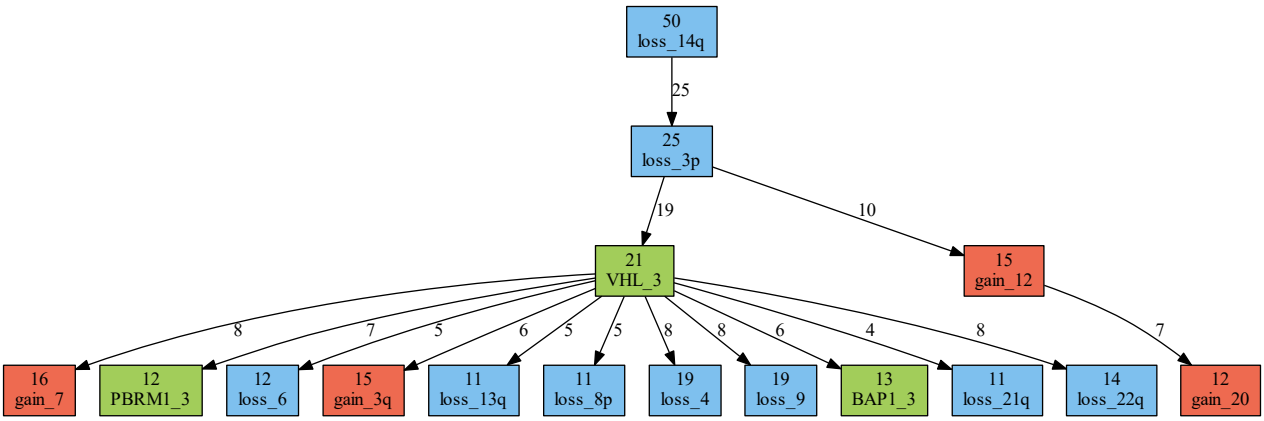

Figure S-8: **loss14q-seeded tree of trajectories with $t = 11$ in TRACERx ccRCC.**

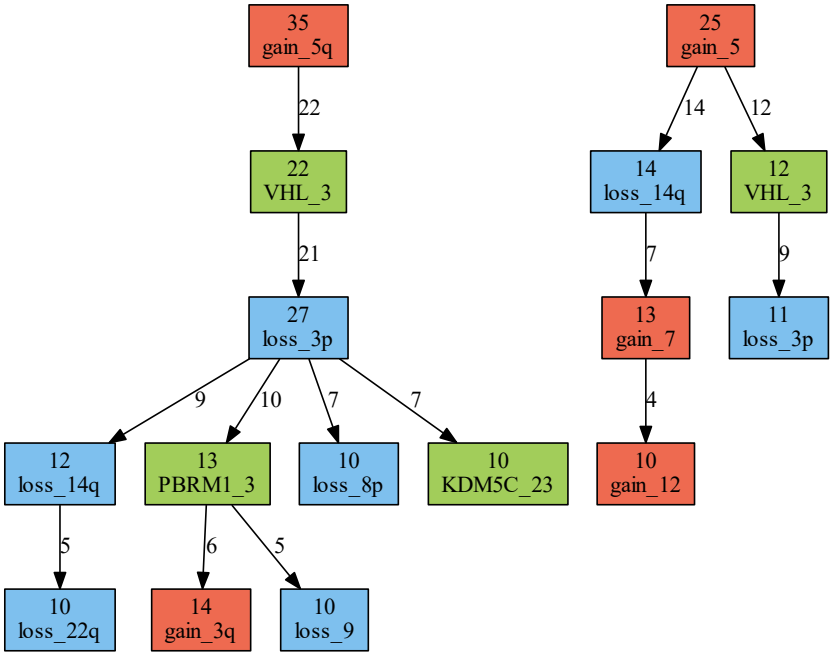

Figure S-9: **gain5 and gain5q seeded trees of trajectories with $t = 10$ in TRACERx ccRCC.**

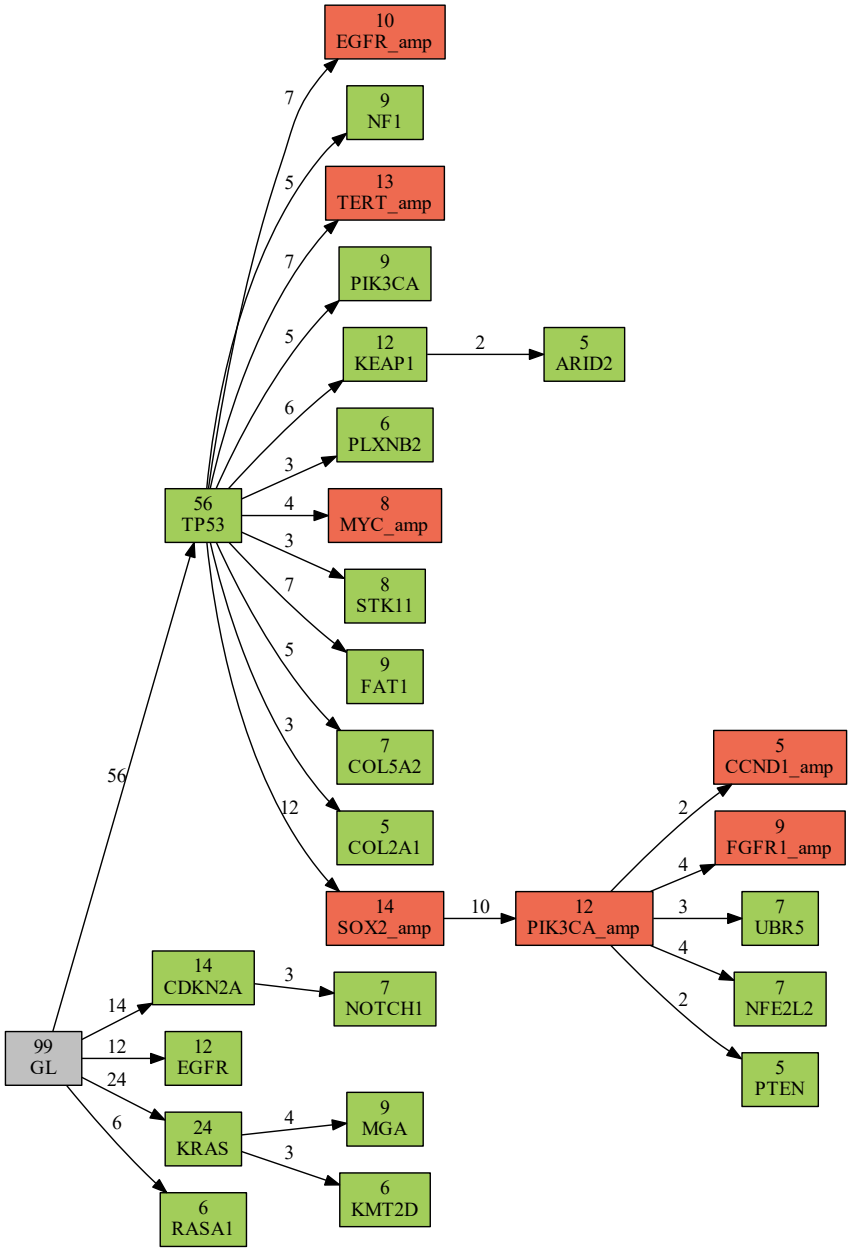

Figure S-10: **Germline seeded tree of trajectories with $t = 5$ in TRACERx NSCLC.**

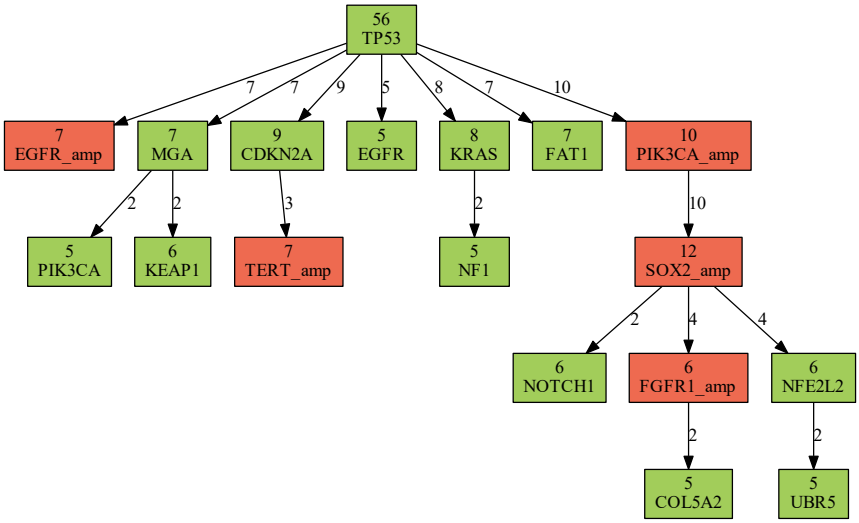

Figure S-11: **TP53 seeded tree of trajectories with $t = 5$ in TRACERx NSCLC.**