Learning mutational graphs of individual tumor evolution from multi-sample sequencing data

Daniele Ramazzotti[1], Alex Graudenzi*[2], Luca De Sano[2], Marco Antoniotti[2], and Giulio Caravagna[3]

[1]Department of Pathology, Stanford University, California 94305, USA
[2]Department of Informatics, Systems and Communication, University of Milan-Bicocca, 20126 Milan, Italy
[3]School of Informatics, University of Edinburgh, Edinburgh EH8 9YL, United Kingdom

Abstract

Phylogenetic techniques quantify intra-tumor heterogeneity by deconvolving either clonal or mutational trees from multi-sample sequencing data of individual tumors. Most of these methods rely on the well-known infinite sites assumption, and are limited to process either multi-region or single-cell sequencing data. Here, we improve over those methods with TRaIT, a unified statistical framework for the inference of the accumulation order of multiple types of genomic alterations driving tumor development. TRaIT supports both multi-region and single-cell sequencing data, and output mutational graphs accounting for violations of the infinite sites assumption due to convergent evolution, and other complex phenomena that cannot be detected with phylogenetic tools. Our method displays better accuracy, performance and robustness to noise and small sample size than state-of-the-art phylogenetic methods. We show with single-cell data from breast cancer and multi-region data from colorectal cancer that TRaIT can quantify the extent of intra-tumor heterogeneity and generate new testable experimental hypotheses.

1 Introduction

Intra-tumor heterogeneity (ITH) is the final product of the complex interplay arising from competition, selection and neutral evolution of cancer cell subpopulations, and currently represents a major hurdle in the development of effective diagnostic and therapeutic strategies for most cancer (sub)types [1–7]. For this reason, in the last years an impressive number of computational approaches to reconstruct the evolutionary history of tumors have been devised (see [8, 9] for

*To whom correspondence should be addressed.

recent reviews), taking advantage of the ever increasing resolution and availability of genomic data, as provided especially by *multi-region* and *single-cell sequencing* (SCS) experiments.

Such multi-sample datasets allow to overcome some limitations of single-sample *bulk* sequencing, which returns a noisy mixture of signals from the tumor subpopulations detected in the sequenced biopsy (e.g., the TCGA data [10]). In particular, the analysis of multiple spatially-separated regions of a tumor and its metastases is recently producing a better and clearer picture of ITH in various tumor types [11–16]. Accordingly, a number of algorithms to infer *tumor phylogenies* from allele frequencies, which were originally ideated for single-sample datasets, have been extended to multi-sample data [17–24].

Among multi-sample datasets, SCS data are likely to become predominant in the next few years, as they provide the highest possible resolution [25]. Yet, this technology still suffers from several technical challenges in *cell isolation* and *genome amplification* of single cells, which produce a broad range of data-specific errors, e.g., *allelic dropouts*, *false alleles*, *missing data*, *non-uniform coverage* and *doublets* [26–32]. This state of affairs prevents straightforward applications of perfect phylogeny algorithms to SCS data [33], and a growing number of cancer-specific probabilistic approaches to infer phylogenies from SCS data have being proposed [34–41]. Some of these approaches estimate the *temporal ordering of accumulation of genomic alterations* via *mutational trees* [36–38], whereas others *deconvolve clones and their evolutionary relations* [39–41].

Most of these methods rely on the *Infinite Sites Assumption* (ISA), according to which each mutation occurs at most once during the evolutionary history of a tumor, and is never lost. Unfortunately, violations of the ISA are more frequent than originally expected, due to chromosomal deletions and loss of heterozygosity, which could lead to *back mutations*, and to *convergent evolution*, in which the same mutation is observed in independent clones or lineages (i.e., *parallel mutations*) [34]. More in general, such methods usually depend on a large number of ad-hoc technical assumptions and parameters, e.g., noise model, search schemes, etc., which need to be opportunely tuned, often requiring computationally demanding automated procedures and/ or prior biological knowledge about the underlying phenomenon. Similarly, in many cases arbitrary heuristics are needed to disambiguate equivalently optimal solutions, e.g., when seeking for a maximum parsimony phylogenetic tree.

Here we introduce TRaIT (Temporal oRder of Individual Tumors), a computational framework to infer the order of accumulation of mutations in single tumors that unifies several distinct approaches.

- TRaIT is the first method that explicitly supports both multi-region and SCS data within a unique statistical framework, with remarkable performances with both data types without the need for a model of noise specific to them;

- TRaIT extends mutational trees by accounting for violations of the ISA due to convergent evolution, for certain values of the observed probabilities. In Figure 5-B the analysis of a multi-region colorectal cancer dataset via TRaIT is shown, in which parallel mutations of a driver gene hit independent lineages. As the general validity of the ISA is debated [34], this represents a major advantage of TRaIT with respect to tree-based phylogenetic techniques.

- TRaIT can detect other complex phenomena underlying ITH such as the presence of *multiple independent trajectories* in the same tumor, or *confounding factors* annotated in the data. In the former case, these could be due either to *multiple cells of origin* [42], or to tumour initiation triggered by epigenetic states not annotated in the data (e.g., methylations). The latter relates to the annotation, in the data, of events that are unrelated to the progression.
- TRaIT can process any kind of genomic lesion, e.g., somatic mutations, copy number alterations, fusions, etc., allowing for an unprecedented information integration among data types;
- TRaIT outperforms techniques specifically tailored for multi-region or SCS data, in terms of accuracy and robustness to data-specific errors and small sample size.
- TRaIT displays a significant improvement in terms of computational time and scalability, which represents another key advantage in anticipation of the increasing availability of large-scale studies/ datasets on single tumor evolution.

TRaIT's models are assembled by combining simple "building blocks": if a model contains edge $x \rightarrow y$, then x^+ mutants are ancestral to y^+ ones, and y^+ mutants are *statistically associated* to x^+. Such conditions describe the *underlying clock* among x and y, and are estimated via simple inequalities from data; their confidence is assessed via several statistical approaches, such as testing, bootstrap and cross-validation.

The simplicity of the underlying theoretical framework has several computational advantages, which we exploited to implement a suite of efficient algorithms that can model complex temporal structures, up to *direct acyclic graphs* (DAGs) with disconnected components, hence capturing different key aspects of tumor evolution and allowing for ISA violations (Figure 1-D). On the overall, despite supporting a wider range of data and models compared to standard phylogenies, our methods have state-of-the-art accuracy and more stable performance with small sample size and different data types, as well as lower computational complexity and higher scalability, as extensive tests on synthetic data suggest.

TRaIT can be used to infer the order of accumulating mutations in individual tumors, but not to deconvolve tumor clones' signatures. Thus, we show in this paper how to couple TRaIT to methods for clones detection, so to draw a new and all-encompassing pictures of tumor evolution and ITH, where one can infer *which clones were present in the input data (signatures), and how mutations accumulated within each clone.*

2 Materials and methods

TRaIT includes 4 optimal polynomial-time algorithms (Figure 2) that process a binary matrix D with n columns and m rows [37]. D stores n variables (mutations, CNAs, etc.) detected across m samples (single cells or multi-region samples). If an entry in D is 1, then the associated variable is detected in the sample. Missing data in SCS are handled by a standard EM procedure with multiple imputations [43]. A priori estimates of false positives/ negatives rates

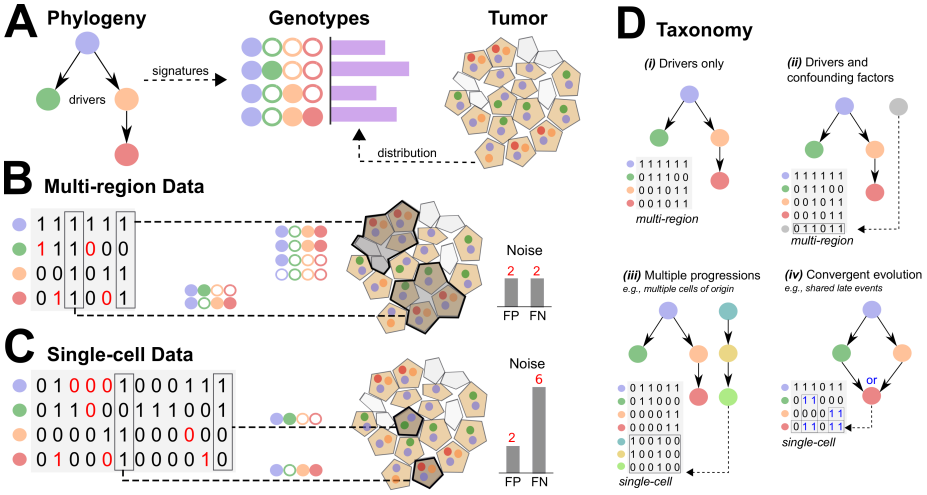

Figure 1: **A.** A phylogenetic model of tumor progression describes the order of accumulation of somatic mutations, CNAs, etc. The model describes a set of possible genotype signatures, which are observed with an unknown spatial and density distribution in a real tumor. **B.** Multi-region bulk sequencing processes a signal mixed from different tumor subpopulations, with potential contamination of non-tumor cells. Thus, a sample will be likely annotated with lesions from different tumor lineages (green, red), creating spurious correlations in the data. In this case, we expect the rate of false positives and negatives in the calling to be symmetric. **C.** If we sequence genomes of single cells we can, in principle, have a precise signal from each subpopulation. However, the inference with these data is made harder by high levels of asymmetric noise, and errors in the calling. **D.** We are interested in studying temporal models of cancer progression in 4 possible scenarios. (i) when all annotated mutations are related to the progression, (ii) when data harbours confounding factors, (iii) when a tumor might have multiple cells of origin and, accordingly, multiple independent progressions and (iv) when independent evolutionary trajectories converge toward a certain mutation (i.e., a confluence). Case (iv) is when the Infinite Sites Assumption (ISA) is violated, as red mutations appear in two distinct evolutionary trajectories.

$\epsilon_+, \epsilon_- \geq 0$ in D can be provided to each algorithm. All the algorithms are implemented in R, in the TRONCO tool for Translational Oncology [44, 45].

At their core, TRaIT's algorithms exploit a bootstrap procedure to compute 2 p-values per edge – one for temporal direction, one for association's strength, according to Suppes' theory of *probabilistic causation* [47] (see Methods). Two algorithms infer *mutational trees* (Edm, Gbw) and two DAGs (ChL, PRIM) (see Methods and Supplementary Information) – for this reason our framework supports *mutational graphs*. All algorithms can return a model with separate components, suggesting that data lacks statistically significant associations, *or* harbours multiple progressions which can be inputed to tumour triggers not annotated in the data (e.g., epigenetic lesions).

For these reasons, TRaIT can be used (Figure 1-D):

(i) when all D's variables are actually involved in the progression (i.e., all the

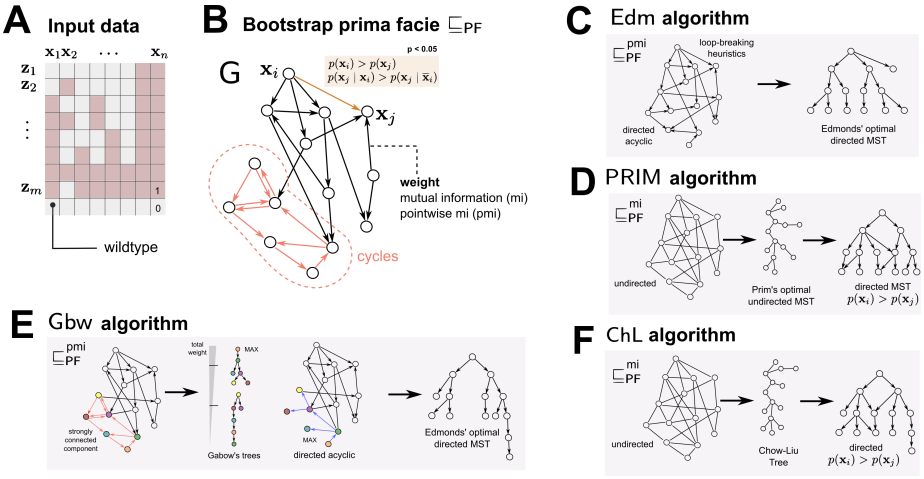

Figure 2: **A**. TRaIT's input data is a binary matrix that stores the presence/absence of a variable in a sample (e.g., a mutation, a CNA, or a persistent epigenomic lesion). Some of these observations harbour false positives and negatives (noise). **B**. We estimate via bootstrap the *prima facie* ordering relation $\sqsubseteq_{\mathsf{PF}}$ that satisfies Suppes' conditions for probabilistic causation, here used as estimators of temporal orderings. This, in turn, induces a graph G over variables $\mathbf{x}_i$. G can be weighted by information-theoretic measures for variables' association. **C.** we weight G with pointwise mutual information (pmi), and we use an heuristic that makes G acyclic by removing less confident edges (see [46]); then we use Edmonds' optimal solution for minimum directed spanning trees. **D.** if we weight G with mutual information (mi) and disregard edges' orientation, we can use Prim's optimal solution for minimum undirected spanning trees. To find a direction for each edge, we use mutations' frequencies. With this algorithm, we can accomodate violations of the ISA due to confluent trajectories when Suppes' conditions are still valid among the parent/ child nodes (see Methods). **E.** if we use Gabow's optimal solution for path traversals of cyclic component, we can detect the best tree that makes G acyclic. Then, we can again use Edmonds' algorithm for spanning trees. **F.** A Bayesian optimal mode-selection strategy can compute the Chow-Liu tree that induces the distribution with minimum divergence from the true one. In this case, we process G as in PRIM to detect edges' direction. As for PRIM algorithm, these models can accomodate violations of the ISA due to confluent trajectories according to the parameter values (see Methods).

events are *drivers*);

(*ii*) when some of D's variables are *confounding factors* (i.e., miscalled mutations);

(*iii*) when D contains multiple independent progressions (i.e., *multiple cells of tumor origin*);

(*iv*) when distinct evolutionary trajectories *converge* to the same variable (i.e., certain drivers are shared by distinct (sub)clonal histories).

Edm and Gbw infer models for $(i, \ldots, iii)$, whereas only ChL and PRIM can explicitly account for (iv); the latter case happens when the ISA is violated. One can choose which TRaIT's algorithm to use according to, e.g., research goals or prior knowledge on the evolutionary process. A rule of thumb might be the application of all TRaIT's algorithms, followed by a comparative analysis of their output models – as we show in the case studies. The creation of a *consensus* model could be also effective to this end. In the Results section we present results from simulations of different experimental conditions and data; these tests allowed to assess the performance of the algorithms in the four scenarios, thus providing indications to an appropriate choice according to the specific case.

The detailed mathematical description of all TRaIT's algorithm is included in the Supplementary Information, whereas in the following some details on the overall theoretical framework are provided.

Suppes' probabilistic causation [47]

Let $p(\cdot)$ be multinomial estimates of the probabilities in D. For every pair of variables x and y in D, x is a plausible cause of y if

$$p(x) > p(y) \quad \textit{and} \quad p(y \mid x) > p(y \mid \neg x)\,. \tag{1}$$

The former condition acts as an *Infinite Sites Assumption* (ISA), as we are assuming that *lesions are persistent*; which can be weakened in certain situations (see below). So, we estimate temporal precedence by marginal probabilities. The latter implies condition statistical dependence: $p(x, y) \neq p(x)p(y)$ [48]. For an edge to be part of our models, *both* conditions must be satisfied. When this is not the case, an edge is included and a non-significant p-value returned (see below). By iterating this approach, we can create models.

This tool is the core ingredient of successful causal approaches for cancer evolutionary inference [46]. It represents a necessary but not sufficient estimator of selective advantage, and combined with a statistical frameworks to disentangle true from spurious associations, can detect selection [49]. With data from a single patient, we limit its power to predict just temporal orderings (see the Supplementary Information).

Working scenarios (Figure 1-D)

Theres is a huge deal of variability in cancer data types, in cancer ITH, as well as in our ability to call mutations etc. Besides, several aspects of cancer evolution are yet undeciphered, so we considered four working scenarios representative of different biologically and technologically-motivated assumptions, and defined corresponding algorithms.

The simplest setting is (i) when all D's variables are *involved in the progression*. Then, we generalize this to when (ii) some of D's variables are annotated in D, but irrelevant to tumor progression (e.g., calling uncertainty or other *confounding factors*). Besides, we account for a (iii) a tumor with multiple cells of origin, and we aim at identifying multiple independent models from a unique dataset. The fourth case is that of a tumor that shows (iv) selective pressures that converge towards variable x. It seems reasonable to consider (i, ii) more common than (iii, iv).

Algorithms (Figure 2)

TRaIT's algorithms use a *non-parametric bootstrap strategy* to assess Suppes' conditions among variables pairs, and include them in a *direct graph* G. Then, four different strategies can compute, from G, a model. Output models can be interpreted as *Suppes-Bayes Causal Networks* [50–52], an extension of Bayesian Networks, with *maximum likelihood estimates* of the parameters θ [53] – bearing in mind that usually such models are used in truly causal approaches. The output is *the Maximum A Posteriori probabilistic model* that best explains D.

The algorithms are inspired by $(i - iv)$, which require us to infer (i) a mutational tree T, (ii) T and some detached nodes for the variables identified as confounding, (iii) a set of trees, $\{T_i\}$, usually called a *forest* and (iv) a *direct acyclic graph* G (since confluent trajectories lead to a node with multiple incoming edges).

TRaIT implements two algorithms to infer trees: Edm (Edmonds) (Figure 2-C), Gbw (Gabow) (Figure 2-E), based on *weighted directed minimum spanning tree* reconstruction. These algorithms scan G to identify the T that maximizes the edges' weights, which are computed via information-theoretic measures of the degree of association of variables – e.g., (pointwise) *mutual information*. Edm and Gbw differ from the way they order *strongly connected components* [54, 55] that appear in G because of finite sample bias. Computationally, Gbw is more expensive and general than Edm.

Two additional algorithms, ChL (Chow-Liu) (Figure 2-F), PRIM (Figure 2-D), are available to infer direct acyclic graphs. ChL is a Bayesian model-selection method to factorize a joint distribution over the input variables [56]. PRIM is the equivalent to Edm for undirected structures, is applied by rendering G undirected, and weighting it with mutual information (which is symmetric). By assigning *a posteriori* an ordering to the undirected spanning trees that is consistent with the variables' frequency, we can retrieve confluent relations that capture violations of the ISA. In TRaIT these cases can be detected under certain conditions of the parameters: if – in the true progression – x and y converge to z, we will not detect two confluent trajectories if $p(x) \leq p(z)$ or $p(y) \leq p(z)$; see the Supplementary Information.

Detection of k independent progressions is a feature available for all algorithms, as it is enforced by the bootstrap when G has k disconnected components. Each group describes evolutions as triggered by multiple initiation cells. Thus, by G's estimation and by picking the proper TRaIT's algorithm, one can easily cover scenarios $(i, \ldots, iv)$.

Complexity

We observe that *all* TRaIT*'s algorithms are optimal polynomial-time algorithmic solutions to each of their corresponding combinatorial problems.* Thus, they *scale well with sample size*, a problem sometimes observed with Bayesian approaches that cannot compute a full posterior and need to sample. TRaIT's algorithms, however, do not have a rich description of uncertainty since they return a single model, but can be however paired with a posteriori forecasts' assessment strategies (e.g., cross-validation/ bootstrap) rather easily [49].

Data types

TRaIT's algorithms work with both SCS and multi-region data. We expect D to contain *noisy observations* of the *unknown true genotypes* (Figure 1-A-C). The algorithms can be informed of the usual *false positives and negatives* rates $\epsilon_+ \geq 0$ and $\epsilon_- \geq 0$, respectively. This adds no overhead to the computation, but prevents to learn noise rates from D, as it is instead possible with techniques as SCITE [37]. Since the algorithms show stable performance for slight variations in the input noise rates, avoidance of complex noise-estimation schemas seems a plus, especially when reasonable estimates of ϵ_+ and ϵ_- are known a priori. This strategy is also used in OncoNEM [40].

For SCS with missing data we use a standard Expectation Maximization approach to input missing values; we repeat it n times, and then perform inference and select the MAP best model out of the n trials.

Code availability

All the algorithms included in TRaIT are implemented in R, and are available in the TRONCO tool for TRanslational ONCOlogy [44,45]. TRONCO is available under a GPL3 license at its webpage: https://sites.google.com/site/troncopackage or at Bioconductor: https://bioconductor.org.

Data availability

All data used in this paper are available from the supplementary material of [30] and [57]. To allow the reviewers to replicate our case studies we provide the source code as well as the input data at https://goo.gl/Ku13MM.

Additional file 1 — Supplementary Information

The detailed description of TRaIT's algorithmic framework and the results of extensive tests both on simulated data and several real datasets are provided in the Supplementary Information.

3 Results

Simulations

We assessed the performance of TRaIT's algorithms with simulated single cell and multi-region data.

In particular, we generated multiple batches of independent synthetic datasets from random phylogenies (generative models), with $5 \leq n \leq 20$ nodes and different levels of topological complexity (Figure 1-D). SCS datasets with $10 \leq m \leq 100$ cells and multi-region datasets with $5 \leq m \leq 50$ regions (accounting for sampling bias) were created. To test the robustness against imperfect data, false positives, false negatives (highly asymmetric for SCS) and/or missing data were introduced in the true genotypes, consistently with previous studies [37]. Multiple configurations of parameters were scanned, and we measured the ability to infer true edges (*sensitivity*), and discriminate false ones (*specificity*); further details on data generation are available as Supplementary Information.

We compared our methods to SCITE, the state-of-the-art for phylogenetic inference of *mutational trees* from SCS data [37]. In the test, we also included previously developed approaches to causal inference from single-sample data (CAPRESE [48] and CAPRI [46]).

Full results are in Supplementary Figures S3 and S5–S15. Here we show four simulations in Figure 3; these settings are consistent with the results across all tests. Figure 3 displays the results for TRaIT and SCITE[1] in canonical settings of noise and sample size, for case (*i*) (SCS and multi-region data), for case (*ii*) (multi-region data), and for case (*iii*) (SCS data).

All the techniques achieve high sensitivity and specificity scores from SCS generated by phylogenies with drivers only – Edm and Gbw highlighting the best results (medians approx. 0.8 and 1). When we sampled multi-region data from the same topology, performances worsened for all methods likely due to the smaller sample size and the mixed bulk signal. The introduction of confounding factors (2 out of $n = 13$ variables), does not to impact the performance significantly, and all algorithms mostly discriminate the true generative model. Finally, the inference of tumors with multiple independent progressions proves to be a harder task, as sensitivity decreases and the performance of all methods are similar. Notice that SCITE, in all tests, achieves the lowest specificity; this might point at a mild-tendency to overfitting, probably due to the combination of its search scheme and noise-learning model (see also Conclusion).

General conclusions can be drawn from the whole set of tests that we carried out. As expected, performances improve with lower noise and larger datasets. In particular, with SCS data Gbw, Edm and SCITE seem the best algorithms; they generally achieve very similar sensitivity, even though the latter presents (on average) lower specificity. For SCS data, all the tested algorithms seem very efficient up to 20/30% of missing data, with SCITE showing a slightly greater robustness (Supplementary Figure S11).

Results on multi-region data display similar trends, with Gbw and Edm showing the overall best performance. In this case, however, SCITE is less effective in retrieving both the true and the false relations, especially with small datasets and/or low noise levels.

Interestingly, by systematically analyzing the impact of a variation of the input ϵ_+ and ϵ_- with respect to the true noise values, we discovered that the performance is rather stable for all algorithms (in Figure 3-D we show Gbw algorithm); this supports our choice – in line with other tools [40]. – of not implementing sophisticate noise-learning strategies in TRaIT.

Finally, a computation time assessment allowed to record a 3× speedup of all the algorithms included in TRaIT with respect to SCITE, on standard CPUs (Supplementary Table 10).

SCS data: Triple-Negative Breast Cancer

We applied TRaIT to a SCS dataset of Triple-Negative Breast Cancer (patient `TNBC` in [30]). The input data consists of single-nucleus exome sequencing of 32 cells: 8 *aneuploid* (`A`) cells, 8 *hypodiploid* (`H`) cells and 16 normal cells (`N`) (Figure 4-A).

[1] SCITE can return a posterior with many equivalent-scoring mutational trees; in those cases, to compute its error, we selected the first of those models.

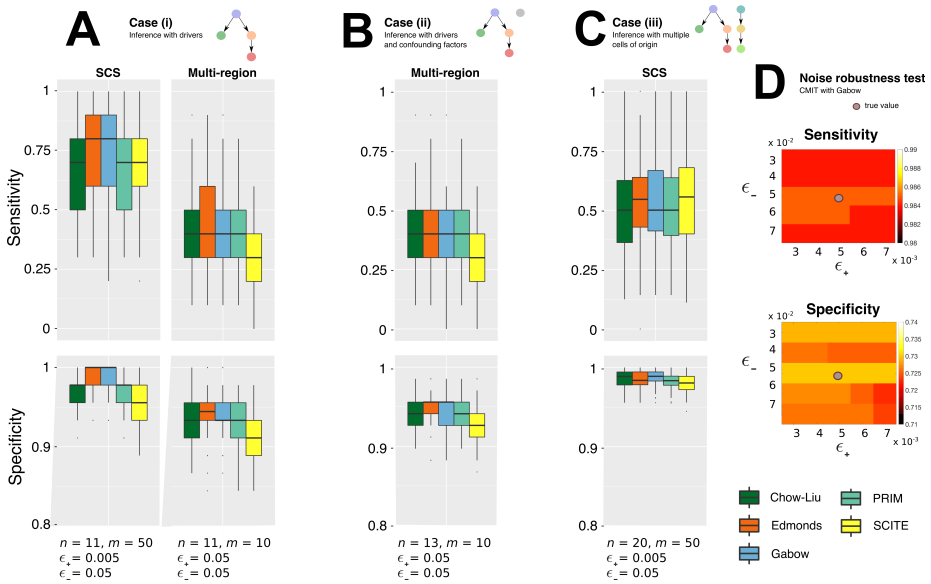

Figure 3: We estimate from simulations the rate of detection of true positives (*sensitivity*) and negatives (*specificity*), visualized as *box-plots* from 100 independent points. We compare TRaIT's algorithms and SCITE, the state-of-the-art for mutational trees inference. For each data type, here we show here a mild-noise setting with canonical sample size: in SCS data noise is $\epsilon_+ = 5 \times 10^{-3}; \epsilon_- = 5 \times 10^{-2}$, in multi-region $\epsilon_- = 5 \times 10^{-2}$. Extensive results for different models, data type, noise and sample size are in Supplementary Figures 3–14. **A.** Here we use a generative model from [39] (Supplementary Figure 5). (left) SCS datasets with $m = 50$ single cells, for a tumor with $n = 11$ mutations. (right) Multi-region datasets with $m = 10$ spatially separated regions, for a tumor with $n = 11$ mutations. **B.** We augment the setting in A-right with 2 random variables (with random marginal probabilty) to model confounding factors, and generated SCS data. **C.** We generated multi-region data from a tumor with $n = 21$ mutations, and a random number of 2 or 3 distinct cells of origin to model independent progressions. **D.** Spectrum of average sensitivity and specificity for Gbw algorithm estimated from 100 independent SCS datasets sampled from the generative model in Supplementary Figure 5 ($m = 75$, $n = 11$). The true noise rates are $\epsilon_+ = 5 \times 10^{-3}; \epsilon_- = 5 \times 10^{-2}$; we scan input ϵ_+ and ϵ_- in the ranges: $\epsilon_+ = (3, 4, 5, 6, 7) \times 10^{-3}$ and $\epsilon_- = (3, 4, 5, 6, 7) \times 10^{-2}$.

In [30], with a bulk sequencing control, mutations detected both in the bulk and in the majority of the cancer cells were annotated as *clonal*, whereas those undetected in the bulk as *subclonal*. The authors then manually curated a qualitative phylogenetic tree (Figure 4-B). We here run TRaIT with the mutational profiles of each single cell describing the presence/ absence of *nonsynonymous point mutations* of the 22 genes selected in [30]. The rate of missing values in this dataset is very low (around 1%); as suggested in [30], we used 9.73×10^{-2} for allelic dropout rate and 1.24×10^{-6} for false positive rate.

For these data, all TRaIT's algorithms return trees, suggesting consistency with the ISA (Supplementary Figures S16-S17); here, we discuss Edm's one since

that algorithm achieved the best performance in the simulations with drivers and confounding factors (Figure 4-C). To improve the analysis, from the same dataset we also deconvolve the signatures of putative clones with OncoNEM, and compare the predictions (Figure 4-D).

TRaIT allows to characterize the qualitative phylogeny provided in [30] by identifying the gradual accumulation of point mutations, expectedly due to defects in DNA repair or replication machineries, both in the clonal and subclonal histories of the tumor.

On the one hand, Edm's model displays high-confidence branched evolution consistent with subclone $\mathtt{A}_1$ (PPP2R1A, SYNE2 and AURKA mutations), $\mathtt{A}_2$ (ECM2, CHRM5 and TGFB2 mutations), and H (NRRK1, AFF4, ECM1, CBX4 mutations) [30]. On the other hand, TRaIT provides a notably higher resolution in the description of the mutations annotated as clonal in [30], e.g., PTEN, TBX3 and NOTCH2, are suggested to trigger tumor initiation. These results are also consistent with the presence of different molecular clocks operating at different stages of tumour growth [30]. TRaIT allows to formulate new hypotheses about undetected subclones, possibly characterized by private mutations in AKAP9, or in JAK1, SETBP1 and CDH6, which however would require further experimental confirmations.

Clones' analysis via OncoNEM detects 10 clones, their lineages and temporal relations, thus refining the qualitative analysis of [30]. Remarkably, such results are mostly consistent with ours, as the mutational ordering predicted by OncoNEM (obtained by estimating the assignment of mutations to clones, as suggested in [40]) largely overlaps with that inferred via TRaIT. This is particularly evident for early events, and for most of the late subclonal ones; exception made for subclone H, which is not detected by OncoNEM. As mutations in ARAF, AKAP9, NOTCH3 and JAK1 have the same marginal probability, their temporal ordering can not univocally determined from these data. TRaIT, in fact, provides a p-value $p > 0.05$ for the direction of those edges, suggesting that any permutation of their ordering would be possible. For this reason, unless more sequenced cells were available, we can not univocally match the clonal signatures obtained with OncoNEM and the temporal orderings identified by TRaIT for these temporally-intermediate events.

This result proves that the concerted use of techniques for the inference of mutational ordering, together with clonal deconvolution approaches, can provide a picture of tumor evolution and ITH at an unprecedented resolution and accuracy.

Multiple-biopsy data: MSI-high Colorectal Cancer

We applied TRaIT to a moderately-differentiated MSI-High colon cancer characterized by a primary tumor and a right hepatic lobe metastasis, with no prior treatments (patient "P3" in [57]). For this patient, targeted DNA resequencing of three regions of the primary tumor (P3-1, P3-2, and P3-3) and of two metastatic regions (L-1 and L-2), allowed to identify 47 *nonsynonymous point mutations* and 11 *indels* [57] (Figure 5-A).

To process this dataset with TRaIT, we first grouped the mutations with the same signature across the five regions, hence obtaining: (a) a clonal group, including the 34 mutations detected in all the samples, (b) a subclonal group, including the 3 mutations detected only in the L regions, and (c) 8 mutations

with different mutational profiles. Our methods will not resolve their ordering since their signals are statistically undistinguishable. However, we will be able to order the groups against the mutations.

With these data both PRIM and ChL predict confluent evolutionary trajectories (Supplementary Figures S19), suggesting a violation of the ISA. In Figure 5-B we show PRIM's direct acyclic graph and Edm's tree. Both models predict *branched tumor evolution* and high ITH among the subclonal populations, consistently with the phylogenetic analysis carried out in [57].

First, the clonal lesions – the clonal root – trigger the first expansions of this tumor, with mutations in the key colorectal drivers APC, KRAS, PIK3CA and TP53 [49]. These biomarkers are ubiquitous, and could not be used to disentangle the mutational spectrum of the primary tumor from the metastatic lesions, in accordance with [58].

Second, the models identify distinct branches outgoing from the trunk, which discriminate the different subclonal evolutions. In both models one subclonal trajectory is initiated by a *stopgain SNV* in the DNA damage repair gene ATM. Edm, in particular, characterizes region P3-1 by a subsequent accumulation of INHBA and CDKN2A nonsynonimous mutations, whereas P3-3 by SMAD4 (stopgain SNV) and KMT2C (frameshift). Conversely, PRIM infers a more complex model, in which two confluent trajectories anticipate common late mutations in different regions: mutations of either INHBA *or* of TGFBR2 may precede mutations of CDKN2A in region P3-1, whereas in region P3-3 alterations of either INHBA *or* of SMAD4 might precede alterations of KMT2C. Interestingly, the alterations of CDNK2A might point to a *cell cycle arrest hallmark* for this tumor. Notice that the model inferred via PRIM exactly fits in scenario (*iv*), which could not be identified with canonical phylogenetic approaches.

Third, in both models the *subclonal metastatic* expansion is originated by a stopgain SNV in GNAQ, anticipating mutations in SMAD4, SETD2, AR (i.e., the subclonal group) and PPP2R1A. The models suggest canonical convergent evolution (i.e., parallel) towards SMAD4 (a stopgain SNV in the primary tumor, and a nonsynonimous mutation in the metastasis). The transducer of transforming growth factor-β superfamily signaling SMAD4 regulates cell proliferation, differentiation and apoptosis [59], and its loss is usually correlated with colorectal metastases [60]. AR is a transcription factor that regulates cell migration and inhibites hepatocellular carcinoma metastases [61]; its splice variants are known to promote metastasis in several tumor types [62]. Similarly, GNAQ is supposed be relevant in metastases development in certain tumor types [63]. PPP2R1A is a negative regulator of signal transduction, gene expression and cell cycle [64], and its mutation influences tumor-endothelium interaction in melanoma metastases [65]. Notice that many other genes that are supposed to characterize MSI-high progression are wildtype in this tumor, e.g., FBXW7, BRAF, ARID1A, FAM123B, etc., [49], as a further evidence of the high level of ITH even within the same tumor subtype.

We finally compared the ordering estimated by TRaIT to the predictions obtained by SCITE (Supplementary Figure S20). Both approaches predict the same formation of the metastatic lesion, yet some significant differences are present. First, SCITE predicts that the mutation of ATM triggers tumor initiation, prior to the mutations included in TRaIT's clonal group, which are ordered in a 34 events-long linear chain. Yet, this specific order has score equivalent to several other models (Supplementary Figure S20) and, thus, might be unreliable.

Besides, in SCITE's model TGFBR2 is associated to region P3-1 (in accordance with PRIM, but not with Edm), GNAQ's stopgain is upstream to both P3-3 and L branches, and some relations appear in inverted temporal ordering, e.g., between SMAD4 and KMT2C. Finally, by construction, SCITE can not infer any confluent evolutionary trajectory, as its statistical model relies on the ISA.

4 Conclusion

The increasing availability of high-resolution multi-sample sequencing data allows one to study ITH at an unprecedented resolution, to understand origination and development of tumors both at the genotype and phenotype level, and to better stratify and treat cancer patients.

Multi-region and SCS data harbours signals that can be informative of different aspects of tumor evolution. In fact, several techniques have been developed that either deconvolve clonal signatures, determine the ordering of growing clones or accumulating mutations, or estimate clonal fractions and cellular prevalence. The concerted application of these techniques allows to draw complex pictures of cancer evolution. As we show for a triple negative breast cancer, one can use the same dataset to detect *both* the signatures of the prevalent clones, *and* to infer the temporal precedence (i.e., ordering) of mutations that generated them. With the right tools, one can hence understand *which* clones are annotated in the data, and *how* they were shaped by evolutionary pressures.

The majority of techniques that perform such analyses ground their roots in standard phylogenetic theory, or in some of its cancer-specific derivations. These techniques are very effective, but sometimes they also implement a noteworthy deal of technical assumptions regarding sequence substitution models, alleles fixation, noise or search scheme etc. As a consequence, it could be hard to quantify how much the final predictions are shaped by the model and its assumptions, or actually suggested by the data. For instance, complex noise-learning models to leverage the imbalances of SCS data might resolve the ordering of clonal mutations in arbitrary ways. This manifests as long trunks whose actual order can not be estimated from current data, and the inclusion of subclonal mutations in dubious positions in the trunk (Supplementary Figure S19).

Similarly, when one seeks for a maximum parsimony phylogenetic tree of tumor evolution, several equivalent-scoring solutions could be returned. When that happens, one has to implement disambiguation heuristics to select *one* output model [11, 12]. This could be one of the computed trees, or a new tree that is a combination of those (e.g., a bootstrap consensus [66]). Despite these routines are often adopted, they are somewhat arbitrary and some deal of care should be warned.

On top of these, recent evidences on the violation of the ISA suggest that this assumption might not be always appropriate; the ISA is preponderant in the derivation of most inferential techniques, and future methods should find a way to consistently account for its violations [35].

In this paper, we deviate from phylogenetic methods and present the TRaIT computational framework, whose methods give statistically robust estimates of mutational orderings in a variety of settings. Our models are simple, and can be interpreted straightforwardly: if an edge connects two mutations (*i*) it statistically resolves their temporal ordering and (*ii*) the mutations are statistically

dependent. Both conditions are estimated from data without using complex inferential models, and assessed via statistical testing, which leads to p-values. Further assessment of models' confidence can be obtained by usual bootstrap or cross-validation approaches [49]. One should be warned that, as mutational trees or other phylogenies, TRaIT's models will display in output all mutations annotated in the input data. So, x and y could be *passengers* mutations observed by hijacking in this patient; nonetheless, their temporal relation can be disentangled, but for a more thorough carachterization of divers against passenger mutations, one would arguably need more complex tools and data combined with *causal* approaches [46, 48, 49].

The simplicity of our framework has major advantages, both from an evolutionary and a computational point of view. First of all, TRaIT's models can account for any variable that can be annotated in a tumor sample. Thus, with TRaIT one can introduce high-level information on pathways, hallmarks, phenotypic-triggering lesions or epigenetic states (e.g., methylations), as long as they are persistent during tumor evolution. Inclusion of these information in traditional sequence-based phylogenetic methods that work with sequences could be harder. Second, TRaIT implements four optimal (i.e., polynomial-time) algorithms that look for different types of signals in the sequencing data and can model more complex topologies than trees, such as direct acyclic graphs with disconnected components. Therefore, TRaIT can be used to investigate whether data suggest the presence of confounding factors, or if the tumor's data harbours several progressions, or if late mutations associate to multiple evolutionary trajectories. This latter case is a first attempt at performing inference when the ISA is violated by convergent evolution, a possibility that is missing in classical phylogenetic methods that are limited to estimating single trees from data [2]. Thus, with TRaIT's algorithms one can test a broad set of hypotheses on tumor evolution as we show in a colorectal cancer case study where we find convergent evolution towards CDKN2A, which might point to a cell cycle arrest hallmark for this tumor type.

The computational burden of our techniques is limited, compared to standard Bayesian approaches (which, however, include an estimation of uncertainty within the model). We do not compute a full posterior over our estimates, but rather a Maximum A Posteriori model constrained by Suppes' conditions. These conditions impose minimum levels of significance to the ordering predicted by our models, and are enforced as *empirical Bayes priors*. In light of the increasingly available data – especially from SCS – TRaIT's scalability properties represent an important algorithmic advancement over Bayesian computations that might become impractical with larger datasets. Our methods accommodate low-effort parallel implementations [67], which we provide in the TRONCO tool for TRanslational ONCOlogy [44, 45].

TRaIT could be improved in several ways. For instance, we could pair bulk sequencing samples to either SCS or multi-region inputs; in fact, the combination of these has been recently shown to improve the estimation of the mutational ordering [38]. Furthermore, we could extend our framework to infer, besides mutational orderings, clonal signatures and architectures, in the attempt of defining a unified framework for cancer evolutionary inference. A general ex-

[2] We refer to Supplementary Information for a discussion of the statistical complications arising from such a generalization, in our framework.

tension to models where the ISA is violated could also be investigated. From a broader perspective, our methods build on our earlier contributions on tumor evolution from single-sample bulk sequencing data [46, 48, 49] (see the Supplementary Information for further discussion). These models allowed us to define the first automatic pipeline to quantify inter-tumor heterogeneity across multiple patients [49].

To conclude, we advocate the use of our methods as complementary to phylogenetic tools for clone deconvolution, in a joint effort to better quantify the extent of ITH. To this end TRaIT represents an innovative and powerful tool to raise precision and effectiveness of large-scale analyses of single tumor evolution.

Author contribution

DR, AG and GC designed the algorithmic framework. DR, LDS and GC implemented the tool. LDS carried out the simulations on synthetic data. Data gathering was performed by DR, AG, LDS and GC. All the authors analyzed the results and interpreted the models. DR, AG, MA and GC wrote the original draft of the paper, which all authors reviewed and revised in the final form.

Acknowledgments

We thank Bud Mishra for his valuable comments on the manuscript and his insights on the effects of parallel evolution to our framework. Furthermore, we thank Guido Sanguinetti and Yuanhua Huang for useful discussions on the preliminary version of this manuscript. This work was partially supported by the SysBioNet project, a Ministero dell'Istruzione, dell'Universitá e della Ricerca initiative for the Italian Roadmap of European Strategy Forum on Research Infrastructures.

References

[1] Serena Nik-Zainal, Peter Van Loo, David C Wedge, Ludmil B Alexandrov, Christopher D Greenman, King Wai Lau, Keiran Raine, David Jones, John Marshall, Manasa Ramakrishna, et al. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, 2012.

[2] Robert J Gillies, Daniel Verduzco, and Robert A Gatenby. Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. *Nature Reviews Cancer*, 12(7):487–493, 2012.

[3] Li Ding, Timothy J Ley, David E Larson, Christopher A Miller, Daniel C Koboldt, John S Welch, Julie K Ritchey, Margaret A Young, Tamara Lamprecht, Michael D McLellan, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–510, 2012.

[4] Rebecca A Burrell, Nicholas McGranahan, Jiri Bartek, and Charles Swanton. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, 501(7467):338–345, 2013.

[5] David Basanta and Alexander RA Anderson. Exploiting ecological principles to better understand cancer progression and treatment. *Interface focus*, 3(4):20130020, 2013.

[6] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *science*, 339(6127):1546–1558, 2013.

[7] Andrea Sottoriva, Inmaculada Spiteri, Sara GM Piccirillo, Anestis Touloumis, V Peter Collins, John C Marioni, Christina Curtis, Colin Watts, and Simon Tavaré. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences*, 110(10):4009–4014, 2013.

[8] Niko Beerenwinkel, Roland F Schwarz, Moritz Gerstung, and Florian Markowetz. Cancer evolution: mathematical models and computational inference. *Systematic biology*, 64(1):e1–e25, 2015.

[9] Russell Schwartz and Alejandro A Schäffer. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, 2017.

[10] National Cancer Institute, National Genome Research Institute. The cancer genome atlas (tcga). https://tcga-data.nci.nih.gov/tcga, 2017.

[11] Marco Gerlinger, Andrew J Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England journal of medicine*, 366(10):883–892, 2012.

[12] Elza C de Bruin, Nicholas McGranahan, Richard Mitter, Max Salm, David C Wedge, Lucy Yates, Mariam Jamal-Hanjani, Seema Shafi, Nirupa Murugaesu, Andrew J Rowan, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science*, 346(6206):251–256, 2014.

[13] Sylvan C Baca, Davide Prandi, Michael S Lawrence, Juan Miguel Mosquera, Alessandro Romanel, Yotam Drier, Kyung Park, Naoki Kitabayashi, Theresa Y MacDonald, Mahmoud Ghandi, et al. Punctuated evolution of prostate cancer genomes. *Cell*, 153(3):666–677, 2013.

[14] Jianjun Zhang, Junya Fujimoto, Jianhua Zhang, David C Wedge, Xingzhi Song, Jiexin Zhang, Sahil Seth, Chi-Wan Chow, Yu Cao, Curtis Gumbs, et al. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science*, 346(6206):256–259, 2014.

[15] Lucy R Yates, Moritz Gerstung, Stian Knappskog, Christine Desmedt, Gunes Gundem, Peter Van Loo, Turid Aas, Ludmil B Alexandrov, Denis Larsimont, Helen Davies, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature medicine*, 21(7):751–759, 2015.

[16] Mariam Jamal-Hanjani, Gareth A Wilson, Nicholas McGranahan, Nicolai J Birkbak, Thomas BK Watkins, Selvaraju Veeriah, Seema Shafi, Diana H Johnson, Richard Mitter, Rachel Rosenthal, et al. Tracking the evolution of non–small-cell lung cancer. *New England Journal of Medicine*, 2017.

[17] Layla Oesper, Ahmad Mahmoody, and Benjamin J Raphael. Inferring intra-tumor heterogeneity from high-throughput dna sequencing data. In *Annual International Conference on Research in Computational Molecular Biology*, pages 171–172. Springer, 2013.

[18] Francesco Strino, Fabio Parisi, Mariann Micsinai, and Yuval Kluger. Trap: a tree approach for fingerprinting subclonal tumor composition. *Nucleic acids research*, 41(17):e165–e165, 2013.

[19] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah. Pyclone: statistical inference of clonal population structure in cancer. *Nature methods*, 11(4):396–398, 2014.

[20] Wei Jiao, Shankar Vembu, Amit G Deshwar, Lincoln Stein, and Quaid Morris. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC bioinformatics*, 15(1):1, 2014.

[21] Habil Zare, Junfeng Wang, Alex Hu, Kris Weber, Josh Smith, Debbie Nickerson, ChaoZhong Song, Daniela Witten, C Anthony Blau, and William Stafford Noble. Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput Biol*, 10(7):e1003703, 2014.

[22] 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):1, 2015.

[23] 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.

[24] Zheng Hu and Christina Curtis. Inferring tumor phylogenies from multiregion sequencing. *Cell systems*, 3(1):12–14, 2016.

[25] Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94, 2011.

[26] Xun Xu, Yong Hou, Xuyang Yin, Li Bao, Aifa Tang, Luting Song, Fuqiang Li, Shirley Tsang, Kui Wu, Hanjie Wu, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, 148(5):886–895, 2012.

[27] Yong Hou, Luting Song, Ping Zhu, Bo Zhang, Ye Tao, Xun Xu, Fuqiang Li, Kui Wu, Jie Liang, Di Shao, et al. Single-cell exome sequencing and monoclonal evolution of a jak2-negative myeloproliferative neoplasm. *Cell*, 148(5):873–885, 2012.

[28] Yingrui Li, Xun Xu, Luting Song, Yong Hou, Zesong Li, Shirley Tsang, Fuqiang Li, Kate McGee Im, Kui Wu, Hanjie Wu, et al. Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. *GigaScience*, 1(1):12, 2012.

[29] Charles Gawad, Winston Koh, and Stephen R Quake. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proceedings of the National Academy of Sciences*, 111(50):17947–17952, 2014.

[30] Yong Wang, Jill Waters, Marco L Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155–160, 2014.

[31] Nicholas E Navin. Cancer genomics: one cell at a time. *Genome biology*, 15(8):452, 2014.

[32] Nicholas E Navin. The first five years of single-cell cancer genomics and beyond. *Genome research*, 25(10):1499–1507, 2015.

[33] Dan Gusfield, Yelena Frid, and Dan Brown. Integer programming formulations and computations solving phylogenetic and population genetic problems with missing or genotypic data. In *International Computing and Combinatorics Conference*, pages 51–64. Springer, 2007.

[34] Alexander Davis and Nicholas E Navin. Computing tumor trees from single cells. *Genome biology*, 17(1):113, 2016.

[35] Nicholas E Navin and Ken Chen. Genotyping tumor clones from single-cell data. *Nature Methods*, 13(7):555–556, 2016.

[36] Kyung In Kim and Richard Simon. Using single cell sequencing data to model the evolutionary history of a tumor. *BMC bioinformatics*, 15(1):27, 2014.

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

[38] Sohrab Salehi, Adi Steif, Andrew Roth, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah. ddclone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. *Genome Biology*, 18(1):44, 2017.

[39] Ke Yuan, Thomas Sakoparnig, Florian Markowetz, and Niko Beerenwinkel. Bitphylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome biology*, 16(1):1, 2015.

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

[41] Andrew Roth, Andrew McPherson, Emma Laks, Justina Biele, Damian Yap, Adrian Wan, Maia A Smith, Cydney B Nielsen, Jessica N McAlpine, Samuel Aparicio, et al. Clonal genotype and population structure inference from single-cell tumor sequencing. *Nature methods*, 13(7):573–576, 2016.

[42] Barbara L Parsons. Many different tumor types have polyclonal tumor origin: evidence and implications. *Mutation Research/Reviews in Mutation Research*, 659(3):232–247, 2008.

[43] Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2012.

[44] Luca De Sano, Giulio Caravagna, Daniele Ramazzotti, Alex Graudenzi, Giancarlo Mauri, Bud Mishra, and Marco Antoniotti. Tronco: an r package for the inference of cancer progression models from heterogeneous genomic data. *Bioinformatics*, 32(12):1911–1913, 2016.

[45] Daniele Ramazzotti, Marco Antoniotti, Giulio Caravagna, Luca De Sano, Alex Graudenzi, Giancarlo Mauri, and Bud Mishra. Design of the TRONCO BioConductor package for TRanslational ONCOlogy. *The R Journal*, 8(2):39–59, 12 2016.

[46] Daniele Ramazzotti, Giulio Caravagna, Loes Olde Loohuis, Alex Graudenzi, Ilya Korsunsky, Giancarlo Mauri, Marco Antoniotti, and Bud Mishra. Capri: efficient inference of cancer progression models from cross-sectional data. *Bioinformatics*, 31(18):3016–3026, 2015.

[47] Patrick Suppes. *A probabilistic theory of causality*. North-Holland Publishing Company Amsterdam, 1970.

[48] Loes Olde Loohuis, Giulio Caravagna, Alex Graudenzi, Daniele Ramazzotti, Giancarlo Mauri, Marco Antoniotti, and Bud Mishra. Inferring tree causal models of cancer progression with probability raising. *PloS one*, 9(10):e108358, 2014.

[49] G. Caravagna, A. Graudenzi, D. Ramazzotti, R. Sanz-Pamplona, L. De Sano, G. Mauri, V. Moreno, M. Antoniotti, and B. Mishra. Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proceedings of the National Academy of Sciences of the United States of America*, 113(28):E4025–E4034, 2016.

[50] Francesco Bonchi, Sara Hajian, Bud Mishra, and Daniele Ramazzotti. Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics*, 2017.

[51] Daniele Ramazzotti, Alex Graudenzi, Giulio Caravagna, and Marco Antoniotti. Modeling cumulative biological phenomena with suppes-bayes causal networks. *arXiv preprint arXiv:1602.07857*, 2017.

[52] Daniele Ramazzotti. A model of selective advantage for the efficient inference of cancer clonal evolution. arXiv preprint arXiv:1602.07614, 2017.

[53] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[54] Jack Edmonds. Optimum branchings. *Mathematics and the Decision Sciences, Part*, 1:335–345, 1968.

[55] Harold N Gabow. Path-based depth-first search for strong and biconnected components. *Information Processing Letters*, 74(3-4):107–114, 2000.

[56] C Chow and Cong Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

[57] You-Wang Lu, Hui-Feng Zhang, Rui Liang, Zhen-Rong Xie, Hua-You Luo, Yu-Jian Zeng, Yu Xu, La-Mei Wang, Xiang-Yang Kong, and Kun-Hua Wang. Colorectal cancer genetic heterogeneity delineated by multi-region sequencing. *PloS one*, 11(3):e0152673, 2016.

[58] A Rose Brannon, Efsevia Vakiani, Brooke E Sylvester, Sasinya N Scott, Gregory McDermott, Ronak H Shah, Krishan Kania, Agnes Viale, Dayna M Oschwald, Vladimir Vacic, et al. Comparative sequencing analysis reveals high genomic concordance between matched primary and metastatic colorectal cancer lesions. *Genome biology*, 15(8):454, 2014.

[59] Hafid Alazzouzi, Pia Alhopuro, Reijo Salovaara, Heli Sammalkorpi, Heikki Järvinen, Jukka-Pekka Mecklin, Akeseli Hemminki, Simo Schwartz, Lauri A Aaltonen, and Diego Arango. Smad4 as a prognostic marker in colorectal cancer. *Clinical Cancer Research*, 11(7):2606–2611, 2005.

[60] Xuemei Li, Baoquan Liu, Jianbing Xiao, Ying Yuan, Jing Ma, and Yafang Zhang. Roles of vegf-c and smad4 in the lymphangiogenesis, lymphatic metastasis, and prognosis in colon cancer. *Journal of Gastrointestinal Surgery*, 15(11):2001, 2011.

[61] Wen-Lung Ma, Cheng-Lung Hsu, Chun-Chieh Yeh, Ming-Heng Wu, Chiung-Kuei Huang, Long-Bin Jeng, Yao-Ching Hung, Tze-Yi Lin, Shuyuan Yeh, and Chawnshang Chang. Hepatic androgen receptor suppresses hepatocellular carcinoma metastasis through modulation of cell migration and anoikis. *Hepatology*, 56(1):176–185, 2012.

[62] Dejuan Kong, Seema Sethi, Yiwei Li, Wei Chen, Wael A Sakr, Elisabeth Heath, and Fazlul H Sarkar. Androgen receptor splice variants contribute to prostate cancer aggressiveness through induction of emt and expression of stem cell marker genes. *The Prostate*, 75(2):161–174, 2015.

[63] Chung-Young Kim, Dae Won Kim, Kevin Kim, Jonathan Curry, Carlos Torres-Cabala, and Sapna Patel. Gnaq mutation in a patient with metastatic mucosal melanoma. *BMC cancer*, 14(1):516, 2014.

[64] Eric G Bluemn, Elysia Sophie Spencer, Brigham Mecham, Ryan R Gordon, Ilsa Coleman, Daniel Lewinshtein, Elahe Mostaghel, Xiaotun Zhang, James Annis, Carla Grandori, et al. Ppp2r2c loss promotes castration-resistance and is associated with increased prostate cancer-specific mortality. *Molecular Cancer Research*, 11(6):568–578, 2013.

[65] Dawn R Christianson, Andrey S Dobroff, Bettina Proneth, Amado J Zurita, Ahmad Salameh, Eleonora Dondossola, Jun Makino, Cristian G Bologa, Tracey L Smith, Virginia J Yao, et al. Ligand-directed targeting of lymphatic vessels uncovers mechanistic insights in melanoma metastasis. *Proceedings of the National Academy of Sciences*, 112(8):2521–2526, 2015.

[66] Joseph Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, pages 783–791, 1985.

[67] Daniele Ramazzotti, Marco S Nobile, Paolo Cazzaniga, Giancarlo Mauri, and Marco Antoniotti. Parallel implementation of efficient search schemes for the inference of cancer progression models. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2016 IEEE Conference on*, pages 1–6. IEEE, 2016.

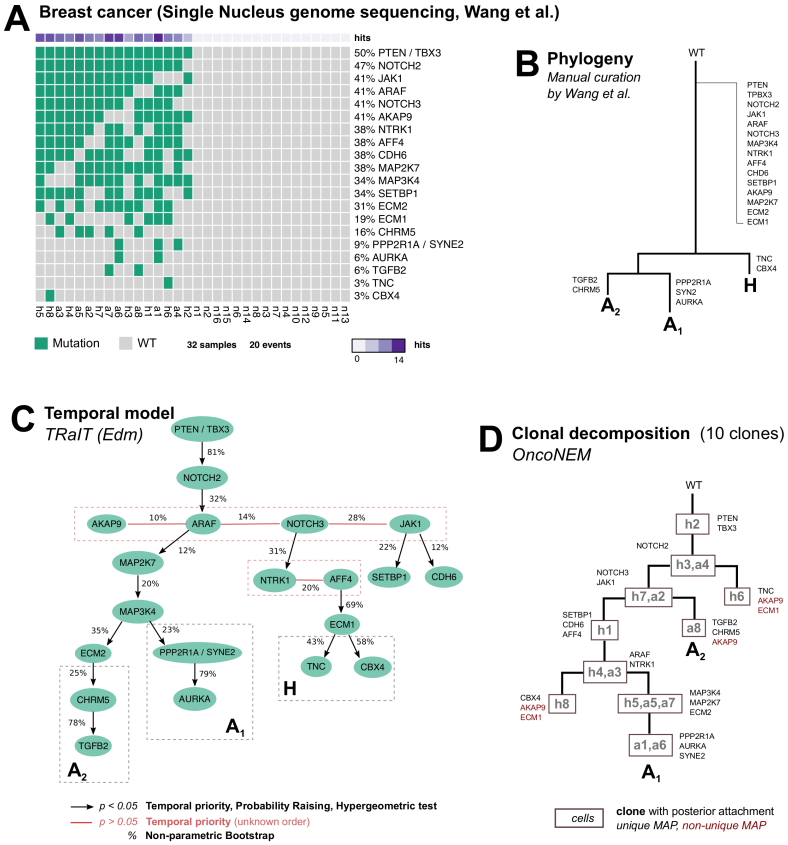

Figure 4: **A.** Input data from single-nucleus sequencing of a triple-negative breast cancer [30] (32 cells). The rate of missing values for this dataset is very low (around 1%), allelic dropout has rate 9.73×10^{-2}, and false discovery 1.24×10^{-6}. **B** Manually curated phylogenetic tree estimated in [30]. Mutations are annotated to the trunk if they are ubiquitous across cells, and detected also in a bulk control sample [30]. Subclonal mutations are those appearing only in more than one cell. **C**. Temporal mutational tree obtained with Edm algorithm; p-values are obtained by 3 tests for Suppes' conditions and overlap (hypergeometric test), and edges annotated with a posteriori non-parametric bootstrap scores (100 estimates). For these data, all TRaIT's algorithms return trees (Supplementary Figure 16), consistently with the manually curated phylogeny (A). Most edges are highly confident ($p < 0.05$), expect for groups of variables with the same frequency which have unknown ordering (red edges). The ordering of mutations in subclones $\mathtt{A}_1$, $\mathtt{A}_2$ and tumor initiation has high bootstrap estimates ($> 75\%$). **D**. We perform a concerted analysis to estimate both clones' signatures and their formation, at least for mutations with a clear statistical signal in these data. We do this by computing a clonal tree with OncoNEM, which predicts 10 clones. Mutations are assigned to clones via *maximum a posteriori* estimates. The mutational ordering of the early clonal expansion of the tumor, which involves mutations in PTEN, TBX3, NOTCH2, is consistent among both models. The same happens for most of the late subclonal events, e.g., mutations in MAP2K7, MAP3K4, PPP2R1A, SYNE2 and AURKA in subclone A1. However, the temporal ordering of intermediate events has weaker support as they have the same marginal probability, and any permutation of their ordering would be equivalent for our method.

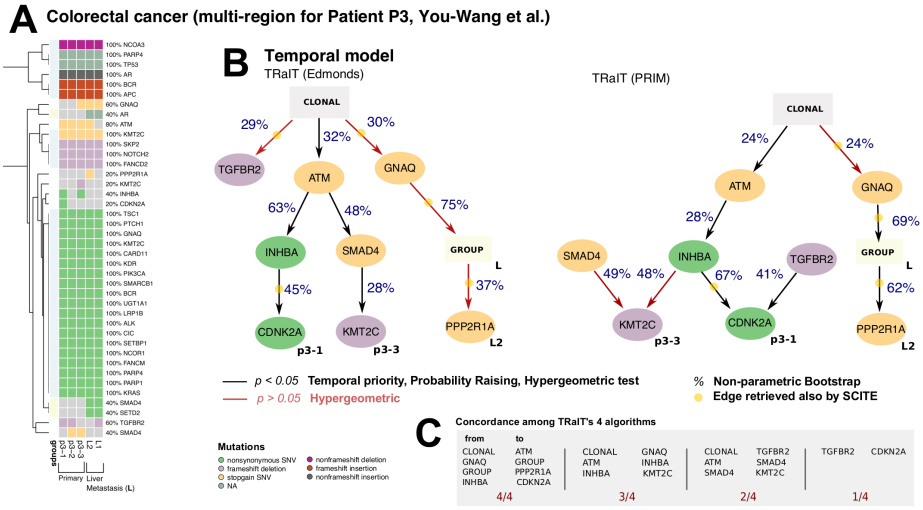

Figure 5: **A**. Multi-region sequencing data for a MSI-high colorectal cancer [57], with three regions of the primary cancer: `p3-1`, `p3-2` and `p3-3`, and two of one metastasis: `L-1` and `L-2`. To use this data with TRaIT we merge mutations that have the same signature across all regions, obtaining a *clonal* group (light blue) including 34 mutations and *sublclonal* group (light yellow) including: non-synonymous SNVs of SMAD4 and SETD2, and non-framshift insertion of AR. **B**. Models obtained by Edm and PRIM algorithms, with their confidence annotated and the overlap in the predicted ordering obtained by SCITE. PRIM predicts convergent evolution – which violates the ISA – towards a non-synonymous mutation in CDKN2A, which is also predicted by ChL (Supplementary Figure 18). All edges, in all models, are statistically significant for Suppes' conditions (temporal precedence and selection strengths). **C.** Four of the predicted ordering relations are consistently found across all TRaIT's algorithm, which gives a high-confidence explanation for the formation of the `L2` metastasis. This finding is also in agreement with predictions by SCITE (Supplementary Figure 19).