bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Profiles of expressed mutations in single cells reveal patterns of tumor evolution and therapeutic impact of intratumor heterogeneity

Farid Rashidi Mehrabadi[1,5,★], Kerrie L. Marie[2,★], Eva Pérez-Guijarro[2,★], Salem Malikić[1,★], Erfan Sadeqi Azer[5,†], Howard H. Yang[2], Can Kızılkale[6,7], Charli Gruen[2], Huaitian Liu[2], Michael C. Kelly[3], Christina Marcelus[2], Sandra Burkett[4], Aydın Buluç[6,7], Funda Ergün[5], Maxwell P. Lee[2], Glenn Merlino[2,✉], Chi-Ping Day[2,✉], and S. Cenk Sahinalp[1,✉]

[1]Cancer Data Science Laboratory, National Cancer Institute, Bethesda, MD 20892, USA
[2]Laboratory of Cancer Biology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA
[3]Single Cell Analysis Facility, Frederick National Laboratory, Bethesda, MD 20892, USA
[4]Molecular Cytogenetics Core Facility, Frederick National Laboratory, Frederick, MD 21701, USA
[5]Department of Computer Science, Indiana University, Bloomington, IN 47408, USA
[6]Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[7]Department of Electrical Engineering and Computer Sciences UC Berkeley, Berkeley, CA 94720, USA
[†]Now at Google LLC, Sunnyvale, CA 94089, USA
[★]Joint first authors
[✉]Corresponding authors: gmerlino@helix.nih.gov, chi-ping.day@nih.gov, cenk.sahinalp@nih.gov

Abstract

Advances in single cell RNA sequencing (scRNAseq) technologies uncovered an unexpected complexity in solid tumors, underlining the relevance of intratumor heterogeneity for cancer progression and therapeutic resistance. Heterogeneity in the mutational composition of cancer cells is well captured by tumor phylogenies, which demonstrate how distinct cell populations evolve, and, e.g. develop metastatic potential or resistance to specific treatments. Unfortunately, because of their low read coverage per cell, mutation calls that can be made from scRNAseq data are very sparse and noisy. Additionally, available tumor phylogeny reconstruction methods cannot computationally handle a large number of cells and mutations present in typical scRNAseq datasets. Finally, there are no principled methods to assess distinct subclones observed in inferred tumor phylogenies and the genomic alterations that seed them. Here we present `Trisicell`, a computational toolkit for scalable tumor phylogeny reconstruction and evaluation from scRNAseq as well as single cell genome or exome sequencing data. `Trisicell` allows the identification of reliable subtrees of a tumor phylogeny, offering the ability to focus on the most important subclones and the genomic alterations that are associated with subclonal proliferation. We comprehensively assessed `Trisicell` on a melanoma model by comparing the phylogeny it builds using scRNAseq data, to those using matching bulk whole exome (bWES) and transcriptome (bWTS) sequencing data from clonal sublines derived from single cells. Our results demonstrate that tumor phylogenies based on mutation calls from scRNAseq data can be robustly inferred and evaluated by `Trisicell`. We also applied `Trisicell` to reconstruct and evaluate the phylogeny it builds using scRNAseq data from melanomas of the same mouse model after treatment with immune checkpoint blockade (ICB). After integratively analyzing our cell-specific mutation calls with their expression profiles, we observed that each subclone with a distinct set of novel somatic mutations is strongly associated with a distinct developmental status. Moreover, each subclone had developed a specific ICB-resistance mechanism. These results demonstrate that `Trisicell` can robustly utilize scRNAseq data to delineate intratumoral heterogeneity and tumor evolution.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 Introduction

Cancer is a genomic disease that originates in distinct cells and grows through an evolutionary process, including acquisition of genomic alterations, environmental selection, clonal expansion, and competition. The "natural history" of cancer development results in extensive intratumor heterogeneity, which drives disease progression and therapeutic resistance [1, 2, 3, 4]. From a genomic point of view, the subclonal makeup of tumors can be better understood through modeling their evolutionary history as a branched process, in the form of a phylogenetic tree [5]. Tumor phylogenies depict relative temporal order of genomic alteration events during tumorigenesis. As such, they can help identify key mutations driving the formation of distinct subpopulations, especially those that are resistant to therapies. Furthermore, tumor phylogenies can also reveal metastatic seeding patterns and sets of mutations harbored by the metastasis seeding cells [6].

The first computational tools for reconstructing tumor phylogeny were developed for bulk genome/exome sequencing data (see [7] for a survey). Unfortunately, since bulk sequencing provides only an aggregate signal over a large number of cells, it offers a very limited resolution in the inferred trees of tumor evolution. Specifically, computational methods based on bulk sequencing data are frequently unable to distinguish between multiple distinct trees that describe the observed data equally well [8, 9, 10].

The rise of single cell sequencing (SCS) has made it possible to explore the makeup of tumors at a higher, cellular resolution. In particular, single cell transcriptomic sequencing (scRNAseq) has significantly improved our understanding of intratumor heterogeneity from a gene expression point of view (see [11] for a survey). As per its application to developmental biology, scRNAseq has enabled the identification of tumor subpopulations based on their gene expression, differentiation status and lineages, and how they respond to therapies. Unfortunately, computational methods for tracing lineages through gene expression profiles require a hypothesis of developmental origin, and as such, can not help infer the evolutionary history of the tumor in detail [12]. In contrast, recently developed barcoding techniques (see [13] for a survey) can help perform detailed lineage tracing up to the time of the original tumor biopsy, but not earlier. Finally, tumor phylogenies based on genomic alterations observed in single cells provide a very detailed evolutionary history of a tumor and are our main focus in this paper.

Many of the methods developed for tumor phylogeny inference from single cell data primarily employed genomic mutations and short indels observed in single cells. The first methods in this domain, such as SCITE [14] and OncoNEM [15], aimed to build the "most likely" tumor phylogeny through genomic SCS data under the "infinite sites" assumption (ISA). Even though some (e.g. SiFit [16] and SiCloneFit [17]) relaxed this assumption by allowing ISA violations at a given "cost" and others (e.g. SPhyR [18]) generalized it to a Dollo parsimony, the majority of available tools still employ ISA (see [19] for a survey), and aim to infer a "perfect phylogeny" by "correcting for noise" in genotype calls (due to, e.g., allele dropout or low sequence coverage), possibly with a maximum likelihood/parsimony approach. Since this is a computationally hard (in fact "NP-hard") problem [20, 21, 22, 18] any principled approach has a theoretical upper bound with respect to the scale of the SCS data it can handle. As a result, available methods either employ MCMC (Markov Chain Monte Carlo) methods and other heuristics (e.g. SCITE [14]), which can not guarantee optimality, or use Integer Linear Programming or Boolean Constraint Satisfaction problem solvers (e.g. PhISCS [23]), which have worst-case running times exponential with the input size. Even recent, custom-built branch and bound approaches such as PhISCS-BnB [24] can not cope with the scale of emerging SCS datasets with thousands of mutations detected in hundreds of cells. As a result, the only scalable tumor phylogeny inference methods in existence are the "distance-based" heuristics (e.g. neighbor-join-based ScisTree [22]), which have known theoretical limitations and can not offer performance guarantees.[1] An orthogonal set of tools utilize single cell copy number variation (CNV) calls, (e.g. offered by Chromium Single Cell CNV Solution from 10x Genomics), to infer the evolutionary history of tumors (e.g. [25]) but they work only with a limited number of (sub)clones and copy number alteration events.

The combination of single cell sequencing and phylogenetic analysis is the key to understanding intratumor heterogeneity. Unfortunately, single cell genome sequencing (scWGS) and exome sequencing (scWES) are prohibitively expensive and suffer from high levels of sequencing noise [26]; as a result, they are not commonly employed. In contrast, single cell whole transcriptome sequencing protocols, such as Smart-seq2 [27], have gained popularity due to their relatively modest cost. However, their ability to offer information about (expressed) mutations and copy number alterations has been underutilized. In fact, there have been only two attempts to (1) cluster single cells based on their mutational profiles derived from scRNAseq data and/or (2) utilize these clusters to build phylogenies. The first was DENDRO [28], which, as per ScisTree, is a greedy hierarchical clustering method based on coverage-weighted differences between cells, and thus has theoretical limitations with respect to accuracy. Another was Cardelino [29],

[1] Consider a simple example with a noise-free input comprised of three cells and $m > 4$ mutations. Suppose the first cell has only the first mutation, the second cell has only the second mutation and the third cell has all mutations except the first mutation. Among these cells, the "nearest" two are the first and second, with a Hamming distance of 2. A distance-based greedy method could conclude that the first and second cells are siblings, and their parent is a sibling of the third cell. ISA dictates that this parent node does not include any mutation. This would imply that the second mutation occurs independently both in the third cell and the second cell, violating ISA. However there is an alternative, ISA satisfying solution where the second and third cells are siblings, and their parent is a sibling of the first cell.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

which, as per SCITE, simultaneously assigns cells and mutations to nodes/clones. Since Cardelino is based on Gibbs sampling, it offers no performance guarantee and does not scale up to infer sizable/detailed phylogenies (e.g., most of the trees reported in [29] consist of 4 nodes).

To address some of the key challenges in tumor phylogeny reconstruction from single cell genomic and, in particular, transcriptomic sequencing data, we developed `Trisicell` (Triple-toolkit for single-cell tumor phylogenetics) comprised of three computational methods of independent but complementary aims and applications (see Figure 1): (1) `Trisicell-PartF` is the first method devised to compute the *partition function* of a tumor phylogeny. Given an input genotype matrix, `Trisicell-PartF` employs a localized sampling strategy to compute the probability of (and thus our confidence in) any user-specified set of cells forming a subclone seeded by one or more (again user-specified) mutations. Importantly, we prove that, as the sample size increases, the output of `Trisicell-PartF` converges to the correct value of the partition function. (2) `Trisicell-Cons` is devised to compare two or more phylogenies (more specifically cell-lineage trees) derived from the same single cell data, and build their consensus tree by "collapsing" the minimum number of their edges so that the resulting trees are isomorphic. `Trisicell-Cons` guarantees to build the optimal consensus tree across phylogenies inferred by distinct methods and/or data types, and helps in identifying robust evolutionary relationships between cells/mutations. (3) `Trisicell-Boost` is a booster for phylogeny inference methods allowing them to scale up and handle large and noisy scRNAseq datasets. For that, `Trisicell-Boost`, iteratively samples random subsets of mutations and builds the phylogeny of each subset through a user-selected inference method (e.g. SCITE or PhISCS). It then establishes the consensus of these phylogenies as the global phylogeny of the tumor sample. In combination, the three components of `Trisicell` provide means to infer large scale tumor phylogenies, compare them in a principled manner, and assess their distinct substructures/subclones, especially through the use of scRNAseq data.

We assessed `Trisicell` on both simulated and real tumor sequencing data for phylogeny reconstruction and evaluation. In particular, we applied `Trisicell` to study melanoma, the malignant disease of melanocytes.

During embryonic development, neural crest (NC) differentiates to melanoblasts; they migrate and mature to melanocytes in skin. In normal skin, melanocytes accumulate mutations induced by ultraviolet radiation (UV) or other carcinogens over time. In rare instances, some melanocytes are transformed into melanoma. It was recently shown that human melanoma can be categorized into four distinct developmental subtypes [30] that are associated with expression of specific marker genes: "undifferentiated" subtype with the highest expression of *Axl*, "NC-like" subtype with high level of *Axl* and *Erbb3*, as well as "transitory" and "melanocytic" subtypes with lower level of *Erbb3* but high level of *Mitf*. In a recent study, we have also developed a "melanocytic plasticity signature" (MPS) score, which involves the expression levels of 45 genes: a higher MPS score correlates with higher plasticity and lower differentiation [31]. Importantly, therapeutic response to immune checkpoint blockade (ICB) is associated with the MPS score in both mouse and human melanomas.

For our recent study, we have generated multiple mouse melanoma models using distinct and relevant genetic alleles and carcinogenic procedures. Among them, the "M4" mouse melanoma model represents UV-induced RAS-mutated, differentiation-transitory subtype of human melanoma. This model responds to anti-CTLA-4 in a diverse manner, replicating the clinical treatment of melanoma with immune checkpoint inhibitors. Histological analysis of the tumors from this model also demonstrated a heterogeneous distribution of melanocytic makers' expression values (*TRP2*, *TYRP1*, and *SOX10*), as well as pigmentation [31]. Therefore, it is a suitable model for studying intratumor heterogeneity.

In this study, we have applied `Trisicell` to sequencing data derived from the melanoma line B2905 derived from the M4 melanoma model. Specifically for this paper, we isolated and expanded 24 single cells from B2905 cell line to distinct "clonal sublines" that possess genomically uniform populations. Three tasks were performed: (i) On each subline, we performed high coverage bulk whole exome sequencing (bWES). (ii) We selected ~ 8 single cells from each subline and applied full-length transcriptome sequencing via the Smart-seq2 protocol [32] (scRNAseq). (iii) We finally selected 11 of these 24 sublines and implanted them in syngeneic C57BL/6 mice, each with 3 replicates, and performed bulk whole transcriptome sequencing (bWTS) on the resulting tumors (see Section 2.3 for the selection criteria).

The high coverage bWES data forms a "gold standard" for mutation calling in each of the 24 sublines. For each subline, we compared the mutations identified in bWES data to those derived from bWTS and scRNAseq data to identify those mutations that are expressed. Additionally, we evaluated "mutation calls" on both bWTS and scRNAseq data that are not present in bWES data to assess the scale of sequencing artifacts and RNA editing events [33] that lead to false positive mutation calls in transcriptome sequencing data. Since for some sublines we obtained bWTS data from three distinct replicates and scRNAseq data from eight single cells for each subline, these mutation calls are very robust. As a result, this tri-modal sequencing dataset offers a powerful way to benchmark computational methods for tumor phylogeny reconstruction and evaluation, especially through the whole and single cell transcriptome sequencing.

We first employed simulated data to benchmark `Trisicell-Boost` using two existing tumor phylogeny inference methods, namely SCITE and PhISCS, which we ran both as stand-alone programs as well as after boosting them with `Trisicell-Boost`. We observed that the run time performance of PhISCS as well as both the running time and accuracy of SCITE were substantially improved through the use of `Trisicell-Boost`, particularly when solving large

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

problem instances.

Next, we assessed `Trisicell` on the tri-modal sequencing dataset generated from 24 clonal sublines of parental (B2905) cell line derived from the "M4" mouse melanoma model [31]. Having assessed that `Trisicell-Boost` (boosting, e.g. SCITE) achieves high levels of accuracy on simulated genomic data, we first applied it to the bWES data. The resulting phylogeny clusters the most rapidly growing sublines in vivo in a distinct clade. The probabilities of specific mutations seeding this subclone, as computed by `Trisicell-PartF`, is also very high, offering further evidence for the correctness of not only the topology of the inferred phylogeny but also its mutational placements.

We then built phylogenies for both bWTS data from 11 sublines (selected out of 24, as described in Section 2.3) and the scRNAseq data from all 24 sublines. `Trisicell-Cons` demonstrated that these phylogenies were in strong agreement with what we obtained for bWES data. Importantly, these phylogenies exhibited (almost) perfect clustering across cells (~ 8 cells per subline, 175 cells in total) and replicate originating from the same subline (3 replicates per subline), further validating the implied evolutionary relationships between cells and sublines inferred through scRNAseq and bWTS data.

After extensively evaluating `Trisicell` on our tri-modal sequencing dataset, we applied it to a larger, preclinical, dataset involving B2905 melanoma cell line implanted in syngeneic (C57BL/6) mice, this time treated with anti-CTLA-4 immune checkpoint inhibitor. We sequenced a total of 764 cells from four anti-CTLA-4-treated tumors as well as two control tumors, via Smart-seq2 protocol, which, again, allowed the identification of mutations in full-length transcripts. The phylogeny we built for this large scale scRNAseq dataset features distinct subclones strongly associated with differentiation status and therapeutic response, leading to the conclusion that a subset of seed mutations drove reshaping of the developmental landscape of the tumor, resulting in substantial phenotypic heterogeneity.

2 Results

2.1 Development of `Trisicell`

`Trisicell` is comprised of three computational methods of independent use and interest: respectively for *assessing, integrating,* and *boosting* tumor phylogeny reconstruction methods so as to obtain reliable evolutionary histories of cancers in a scalable manner. Importantly, we also present matching whole transcriptome and exome sequencing data from clonal sublines, each derived from a single cell of a melanoma mouse model, as well as single cell transcriptome sequencing data from these sublines. As will be detailed below, by jointly evaluating `Trisicell` inferred and validated phylogenies on these clonal sublines with matching gene expression data, we demonstrate that certain subpopulations of the melanoma mouse model seeded by specific mutations share developmental status, implying a connection to tumor evolution. We expect that this dataset will help future method development efforts by offering a high-quality platform for benchmarking, parameter tuning, and integration with other methods.

`Trisicell` is comprised of the following three tools:

(1) Available methods for tumor phylogeny inference from single cell data aim to reconstruct the "most likely" tree using a noisy genotype matrix as their primary input. Even if all of the mutation calls in the genotype matrix are correct and the method used can guarantee optimality, the output tree is just one of the many possible interpretations of the input data as it only represents the "global" optimum. However, in many cases, we are interested in the likelihood (and the reliability) of "local structures". This is commonly encountered in structural biology, e.g. in RNA-structure prediction where the goal is typically computing the most likely/thermodynamically stable secondary structure [34]. The secondary structure of an RNA strand can be represented as a tree and the likelihood of distinct subtrees of the secondary structure can be assessed through a "partition function algorithm" [35, 36].[2] `Trisicell-PartF` features the first partition function formulation for a tumor phylogeny and employs a sampling strategy to compute it for any user-defined subtree/subclone seeded by one or more mutations. Importantly, we prove that, as the sample size increases, the output of `Trisicell-PartF` converges to the correct value of the partition function.

(2) In order to "solve" large instances of tumor phylogeny reconstruction problem, available approaches make implicit and explicit assumptions on data behavior and, as a result, may output different trees for a given input genotype matrix. In fact, even the set of mutations themselves, and thus the genotype matrix derived from a read collection, may vary significantly as a consequence of the employed mutation calling pipelines. Furthermore, the type of sequencing used for mutation calls (e.g., targeted or whole, genome or transcriptome sequencing) could result in distinct mutation calls for a given set of cells. As a result, it is many times not feasible to compare phylogenies describing the lineages of a given set of cells from a tumor sample through their "mutation tree" representation, where each node is labeled with one or more distinct mutations. Unfortunately, since available measures for comparing tumor phylogenies were typically not developed for single cell data, they focus on how mutations (and not cells) are placed

[2] In the context of RNA-structure prediction, given the collection of potential base pairs as a set of random variables, the partition function represents the marginal probability of a particular base pair (or set of "non-conflicting" subset of base pairs), i.e., the total probability of all possible configurations that feature the base pair(s).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

on distinct lineages.[3] As will be described below, our `Trisicell-Cons` features a novel measure to compare two or more *cell lineage trees*. Here we use the broad definition of a cell lineage tree as a phylogeny where leaves and some of the internal nodes are each labeled with one or more cells (in contrast to a narrower definition where only leaves are labeled with a unique identifier [22]) . Our measure is based on a specific notion of a "consensus" between two trees, accompanied with an algorithm named `Trisicell-Cons` and its implementation to build this consensus. On a given set of cell lineage trees, `Trisicell-Cons` minimizes the number of edges to be "collapsed" so that the resulting trees are isomorphic. As such, `Trisicell-Cons` generalizes the consensus taxonomic tree construction algorithm of Day [41] for dendrograms, i.e. trees where only leaves are labeled with a single unique label. In order to modify Day's algorithm for tumor phylogenies where any internal or leaf node may have an arbitrary number of labels[4], we developed a new combinatorial approach and implemented it in `Trisicell-Cons`. We note that `Trisicell-Cons` provably obtains the largest possible consensus tree between the two input phylogenies in time quadratic with the input size. As we will demonstrate, it can be employed to integrate phylogenies obtained by various methods and data types to help infer a more reliable evolutionary history of a tumor.

(3) `Trisicell-Boost` is a "booster" strategy for scaling up available tumor phylogeny inference methods, especially those based on principled but relatively slow combinatorial approaches such as SCITE and PhISCS. It relies on the observation that the "correct" phylogeny involving any subset of mutations in a tumor preserves the relative order of mutations with respect to their ancestor-descendant and different-lineage dependencies compared to the "correct" phylogeny on the entire set of mutations. In its implementation, `Trisicell-Boost` first performs a sampling and phylogeny inference step in which a specified number of mutations and cells from the input genotype matrix are sampled and, using the resulting submatrix as input, a perfect phylogeny is built through any user-selected phylogeny inference tool that works well on small-scale data. This atomic step is repeated for a user-defined number of times resulting in a collection of smaller phylogenies on distinct subsets of mutations. Since these phylogenies are obtained independently from each other, their inference tasks could be executed in parallel. Once this step is completed `Trisicell-Boost` reconciles them into a single phylogeny featuring the complete set of mutations and cells.

An overview of `Trisicell` methods is depicted in Figure 1.

2.2 Benchmarking tumor phylogeny reconstruction methods on simulated data

Using the procedure described in Supplementary Section H which has been employed to assess many of the available methods for tumor phylogeny inference (including SCITE, PhISCS, OncoNEM, and B-SCITE) we simulated 80 distinct genotype matrices under infinite sites model, with a varying number of mutations (300 and 1000), cells (300 and 1000) and false negative mutation calling rates (0.05 and 0.20). The false positive mutation calling rate was set to 0.001 and the size of the simulated tree of tumor evolution was set to 100 in all simulations (see Supplementary Section H for a more detailed discussion on the tree size and other parameter settings).[5].

On each simulated dataset, we compared `Trisicell-Boost` against both SCITE and PhISCS as stand-alone programs.[6] The accuracy of each tool for each parameter setting is provided in Figure 2. As can be seen, the results obtained by all three methods are comparable when $n = m = 300$. This is also the case when $n = 1000$ and $m = 300$ when false negative rate is 0.05, however PhISCS fails to report solutions for some instances when false negative rate is 0.2. When $m = 1000$ the performance of SCITE (with a 24-hour time limit) under both measures falls substantially below that of `Trisicell-Boost`. PhISCS also has difficulty handling certain problem instances in 24 hours when $m = 1000$, with the exception for $n = 300$ and the false negative rate of 0.05; for this setting it performs similar to `Trisicell-Boost`. It is worth noting that, with respect to different-lineage accuracy measure, `Trisicell-Boost` is near perfect across all simulations. With respect to ancestor-descendant accuracy measure, `Trisicell-Boost` also achieves very high accuracy, which increases as the number of simulated single cells is increased. An evaluation of the robustness of `Trisicell-Boost` to the presence of deletion events, which are a major cause of violations of infinite sites assumption, and additional information of running simulations can be found in Supplementary Section K.

[3] Examples include the number of ancestor-descendant and different-lineage relationships shared across the trees compared, as well as more sophisticated measures such as the Multi-Labeled Tree Edit Distance and others [23, 37, 38, 39, 40].

[4] which is important especially since many methods build phylogenies based on mutations shared by at least two cells or pre-cluster cells based on their mutational profiles

[5] Another parameter used in our simulations is the missing value rate γ, which determines the fraction of entries in the genotype matrix which are unknown. We used $\gamma = 0.2$ in our simulations. This choice is consistent with the missing value rates observed in single cell genome sequencing data but is somewhat lower than that we observed in our real scRNAseq data. Our simulations aim to compare the performance of `Trisicell-Boost` to the methods it boosts, in particular SCITE and PhISCS, which were developed for genome sequence data. For higher missing value rates, a comparison of `Trisicell-Boost` to SCITE, PhISCS, as well as DENDRO and Cardelino, tools developed specifically for RNAseq data, is provided for real scRNAseq datasets.

[6] Since the model employed by SCITE and PhISCS are identical, the trees they built were isomorphic for many of the smaller datasets; thus we do not report `Trisicell-Boost`'s results for SCITE and PhISCS separately.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.3 Benchmarking tumor phylogeny reconstruction methods on clonal sublines from the M4 melanoma model

Melanocytes are derived from the neural crest during embryonic development. In normal skin, melanocytes accumulate mutations induced by ultraviolet radiation (UV) or other carcinogens over time. In rare instances, some of them are transformed into melanoma. It was recently shown that human melanoma can be categorized into four distinct developmental subtypes. Ranked with respect to differentiation levels, they are "undifferentiated", "neural crest-like", "transitory", and "melanocytic" [30]. In a previous study [31], we have generated multiple mouse melanoma models using distinct genetic alleles and carcinogenic procedures with the aim of better modeling human melanoma. Previously, through comparative analysis, we have also identified a "melanocytic plasticity signature" (MPS) composed of 45 genes, which allows quantification of the developmental status of melanoma: a higher MPS score correlates with higher plasticity and lower differentiation [31]. We have found that therapeutic responses to immune checkpoint inhibitors are associated with MPS scores in both mouse and human melanomas. In this study, we used the B2905 cell line derived from "M4" mouse melanoma model that represents UV-induced RAS-mutated, differentiation-transitory subtype of human melanoma. This model responds to anti-CTLA-4 in a diverse manner [31], replicating the clinical treatment of melanoma with immune checkpoint inhibitors. Therefore, it is a suitable model for studying tumor heterogeneity.

In order to investigate `Trisicell-Boost`'s applicability to experimental datasets, we generated three types of sequencing data from tumors derived from the parental M4 (B2905) mouse melanoma model. First, 24 single cells from M4 were expanded into clonal sublines, and then subjected to bulk whole exome sequencing (bWES). We then implanted each subline into five syngeneic mice to monitor their growth kinetics. Out of the 24 sublines, 11 reached 100% tumor-take rate. Each of these tumors was harvested at an endpoint, and 3 tumors from each subline were subjected to bulk whole transcriptome sequencing (bWTS) as triplicates (Figure 3a). Additionally, we extracted 8 single cells from each subline and subjected them to Smart-seq2 protocol to obtain scRNAseq data (Figure 4a).

For assessing `Trisicell` on this tri-modal sequencing dataset, we first built the phylogeny of the 24 clonal sublines, through mutation calls that we made on their bWES data using `Trisicell-Boost` (Figure 4b). Next, we built the phylogeny of the 11 sublines (each with 3 replicates) through mutation calls that we made on their bWTS data (Figure 3c). Then, by the use of `Trisicell-Cons`, we constructed the consensus of the subline phylogenies that we had built through the use of bWES data and bWTS data, which exhibited only minor differences (in the location of the subtree with sublines C1, C4, and C22; see Figure 3f). Furthermore, the tree we built using bWTS data was very similar to that using bWES data with respect to commonly used measures of accuracy (Figure 3e). Note that, even though the exact set of mutation calls made on any given subline from bWTS data differed from that of bWES data, the mutation calls made on the bWTS data differed minimally across the three replicates of the same subline (see Figure 3c for those mutations also observed in bWES data and Figure S16 for the entire set of mutations). As a result, the `Trisicell-Boost` built phylogeny on bWTS data was able to cluster all replicates from each subline correctly, demonstrating its potential use on mutations derived from RNAseq data for tumor phylogeny reconstruction.[7] In contrast, DENDRO and Cardelino, the only other tools for tumor phylogeny inference through RNAseq reads, both failed to cluster replicates of certain sublines correctly, as discussed in more detail below.

The phylogeny built on bWES data features a subtree comprised of the most rapidly growing sublines (measured by the survival time[8]) in vivo (Figure 4b). These sublines, i.e., C11, C15, C16, and C18, are seeded by a distinct set of mutations, several of which are also observed in bWTS data.[9] An analysis of this subtree in both bWES and bWTS phylogenies by `Trisicell-PartF` (see Figure 3g,h) suggests that some of these seeding mutations (especially the seven observed in both bWTS and bWES data, all with high probability) may have contributed to the emergence and proliferation of the associated subclone. In particular, *Map2* is a known melanoma cell proliferation inhibitor [42]. A non-synonymous SNV in exon9:c.G739C:p.G247R of this gene is likely one of the contributors to the rapid tumor growth in these sublines.

As mentioned earlier, we have also used our bWTS dataset to evaluate DENDRO and Cardelino, the only other tools for tumor phylogeny inference via RNAseq reads, in addition to SCITE and PhISCS, this time as stand-alone tools. DENDRO produced a tree which is substantially different from the bWES tree (e.g. it clustered the most aggressive sublines with some of the least aggressive); in fact, it even failed to cluster replicates of certain sublines correctly. Cardelino on the other hand is primarily developed for assigning mutations to the nodes of a preset topology and could not infer trees of this scale - see Supplementary Sections P and Q). SCITE and PhISCS on the other hand produce phylogenies (almost) isomorphic to their boosted versions with `Trisicell-Boost`, providing further evidence for the robustness of `Trisicell-Boost` in scaling up these methods; see Supplementary Section O. (Note that for brevity we only provide the results for SCITE as PhISCS results were identical.)

[7] Even though for most sublines, the expression profiles of all three replicates were very similar, in rare cases a replicate had a substantially different gene expression profile, suggesting an epigenomic alteration.

[8] Survival time is defined as the post-inoculation time required for the tumor to reach the pre-determined endpoint, size of 1000 mm^3.

[9] bWTS data does not include all mutations or sublines.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We finally built the scRNAseq based phylogeny of ~ 8 single cells derived from each of the 24 sublines. A total of 175 (out of $24 \times 8 = 192$) cells provided sufficient read coverage to perform mutation calling. In order to reduce transcriptome sequencing artifacts and RNA editing, we only consider those mutations shared by the corresponding bWES data. Finally, to account for deletion events (leading to loss of heterozygosity LOH), if a mutation in a gene with a copy number of 1 is not observed in a cell, we set its status to "NA" (i.e. unknown) (see Supplementary Section R for details of copy number profiling). The resulting phylogeny obtained by `Trisicell-Boost` (again boosting SCITE) is provided in Figure 4c. Notice that the cells originating from the same subline are well clustered, akin to the clustering of replicates from each subline in the bWTS phylogeny. Furthermore, the major subtrees of the scRNAseq phylogeny are identical to that of the bWES phylogeny as demonstrated by their `Trisicell-Cons` consensus shown in Figure 4d. For example, the subtree featuring the most rapidly growing sublines in vivo (i.e. C11, C15, C16, C18) in bWES phylogeny is preserved in the scRNAseq phylogeny. Additionally, ancestor-descendant and different-lineage mutation pair relationships in the scRNAseq mutation tree with respect to the bWES mutation tree are highly preserved (Figure 4e). Overall, the phylogeny demonstrates that `Trisicell-Boost` can be used to infer detailed and scalable tumor phylogenies through the use of scRNAseq data, especially when combined with bulk whole genome or exome sequencing data for mutation calling. (The corresponding phylogenies we built for all "mutations" observed in bWTS and scRNAseq data can be found in Supplementary Section N.)

We also evaluated DENDRO and Cardelino, as well as SCITE and PhISCS as stand-alone tools, on the scRNAseq dataset. As in the case of some larger simulations, PhISCS could not complete the task in 24 hours; this is because it necessarily searches for the minimum possible noise reduction for obtaining a phylogeny. SCITE on the other hand produced a result since it is designed to search for a solution in a user-specified time, but its output could be sub-optimal. The tree built by SCITE is given in Figure S19: even though SCITE clusters cells from each subline well, the relative placement of the sublines does not agree with that in the bWES phylogeny as demonstrated by their `Trisicell-Cons` built consensus. DENDRO, in contrast, failed to cluster cells from the same subline Figure S21. Similarly, Cardelino, even when used for cell clustering only, fails to perform well on this dataset (see Supplementary Section Q).

2.4 Application of `Trisicell` to anti-CTLA-4-treated tumors derived from the M4 melanoma model

To investigate the origin of intratumoral heterogeneity and its impact on the resistance to anti-CTLA-4, we performed a preclinical study involving six mice implanted with M4 melanoma, four of which were treated with anti-CTLA-4 antibodies, and the remaining two with generic immunoglobulin IgGb as control (Figure 5a).[10] A total of 764 cells from all six tumors that passed the quality control and had sufficient read depth were selected for downstream analysis.

`Trisicell` analysis. To study the subclonal structure of the anti-CTLA-4 treated *M4* melanoma, we applied `Trisicell` to reconstruct the phylogeny from all six tumors. The cells were first locally grouped into 64 mutationally homogeneous clusters and each cluster's reads were re-evaluated to establish their mutation profiles (a similar approach for clustering cells prior to mutation calling have been used earlier, e.g. for CNV profiles [25, 43]). `Trisicell-Boost` inferred phylogeny on this dataset is provided in Figure 5b.

As a next step, for each leaf node we calculated the median of (1) "melanocytic plasticity signature" (MPS) [31] and (2) expression of neural crest (NC) lineage developmental markers including *Axl* (undifferentiated/NC), *Erbb3* (NC/transitory), and *Mitf* (differentiated) [30]. To compare the phenotypes, we also analyzed the proportion of (1) cells from distinct UMAP clusters (see below for UMAP analysis); (2) cells from distinct tumors; and (3) cells from control vs anti-CTLA-4 treatment groups. When tracing specific mutations in each node, the "hotspot" mutations at *Kras*, *Sf3b1*, and *Trp53* (*p53*) were found in the five internal nodes closest to the root. This is expected since these are the well-known "driver" mutations and they frequently occur in human melanoma. Interestingly, the phylogeny features five major subtrees (or clades), which are strongly associated with MPS values, as well as expression of *Axl* and *Mitf*: clade B1 (MPS^{high}, Axl^{+}), clade B2 (MPS^{high}, Mitf^{+}), clade B3 ($\text{MPS}^{\text{medium}}$, Axl^{-}, Mitf^{+}), clade B4 ($\text{MPS}^{\text{medium}}$, Axl^{+}, Mitf^{+}), and clade B5 (MPS^{low}, Mitf^{+}). Across these subtrees, the expression of *Erbb3* is similar to *Mitf*. The seeding mutations of these subtrees include those in *Pik3r2*, *Gna11*, *Adgr3*, *Pten*, *Brd1*, and *Top2a*, genes that are recurrently mutated in human melanoma.

The identification of distinct clades in the phylogeny allowed us to trace the subclonal evolution as a consequence of anti-CTLA-4 treatment. To examine the association between the tumors and clades, we calculate the proportion of cells from each tumor and treatment group for each leaf, as shown in the sixth and seventh panels under the phylogeny in Figure 5b, respectively. While four clades (B2 to B5) do not enrich cells from specific tumors, B1 has

[10] The tumors were harvested when their size reached about 500 mm^3 as endpoint. For scRNAseq analysis, the two control group tumors were harvested at day 41 and 54. From anti-CTLA-4 group, one tumor was harvested at day 54, another at day 75 and final two at day 85 (see Supplementary Section F). The variable time of endpoint demonstrates the diverse growth kinetics and therapeutic responses of this model.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

significantly higher number of cells from tumor m122 of the anti-CTLA-4 treatment group. In correspondence, we investigated the subclonal response to the treatment by calculating the proportion of cells from each clade for each tumor; see Figure 5d. Observe that during the growth of tumor m122, clade B1 became predominant while B2 and B3 diminished, which suggests that these three clades were subjected to strong selection pressure. The fraction of clade B1 also slightly increased in tumor m103 and m109, but not in m115. Interestingly, our previous study had demonstrated that melanoma resistance is associated with high plasticity, consistent with the fact that B1 clade exhibits the highest MPS scores and expression of "undifferentiated" markers[31].

Our new results suggest that subclonal selection may not happen in every tumor responding to therapy. Since it has been reported that cancer cells can adapt to the therapeutic stress via a non-genetic mechanisms [44], we analyzed deferentially expressed genes in each clade by comparing anti-CTLA-4 treated tumors against control tumors (see Supplementary Section D) and then subjected them to Gene Set Enrichment Analysis (GSEA) (see Supplementary Section E). Interestingly, differentially expressed genes of clade B1 are mostly enriched in inflammatory pathways, especially in m122, suggesting that cells of clade B1 were "intrinsically" resistant and thus selected by the immune response. This is consistent with our previous discovery that undifferentiated subtype of melanoma is resistant to immune checkpoint blockade [31], as clade B1 cells exhibit high MPS scores and Axl expression. Cells in other clades seem to activate pathways relevant to mesenchymal stem cells (myogenesis and cholesterol homeostasis in B2 and B3), mitotic arrest (mitotic spindle and G2M checkpoint in B4), and energy metabolism (oxidative phosphorylation in B5).

Expression profile based clustering of cells with UMAP. In order to compare phylogenetic analysis with gene expression profile analysis we employed Seurat package [45][11] which resulted in seven clusters with distinct gene expression patterns (see Figure 5c), corresponding to differentially expressed gene sets (see Supplementary Section G). We examined the expression of lineage markers in these clusters and noticed that clusters c5 and c6 in Figure 5c have higher median MPS scores and more Axl^+ cells than others, suggesting that they are less differentiated. Among these clusters c6 includes no $Erbb3^+$ cells indicating that it is even less differentiated than c5. In contrast, the remaining clusters have lower MPS scores and more $Mitf^+$ cells, indicating that they are more differentiated. Noticeably, the distribution of the expression patterns of these genes is heterogeneous, even among the cells from the same cluster. For example, Axl^+ cells are also present in the five clusters with lower MPS scores. This implies that the UMAP clusters are only weakly associated with developmental status.

A comparison of `Trisicell` inferred phylogeny and UMAP clusters. We calculated the proportion of cells from each UMAP cluster (see Figure 5c) in each leaf node of the phylogeny. As can be seen in the fifth panel of Figure 5b, clade B5 is enriched with UMAP clusters c3 and c7, and clade B3 is enriched with UMAP cluster c1; all these clusters are $Mitf^+$. In contrast, the two Axl^+ clades, B1 and B4, are not enriched with cells from any particular UMAP cluster. These results indicate that the mutation-based clades in the phylogeny are associated with distinct developmental status defined by MPS score and lineage markers, rather than the overall gene expression patterns as suggested by UMAP clusters.

In conclusion, our study demonstrates that `Trisicell` inferred tumor phylogeny from transcriptomic mutational profiles of single cells can robustly trace distinct subclones, making it feasible to study their specific treatment responses. Our results are consistent with recent studies reporting on distinct subclonal populations that develop specific resistance or adaptation mechanisms to therapy [49, 50].

3 Discussion

The large scale of and high levels of noise in single cell whole transcriptome sequencing (scRNAseq) datasets have limited their use in tumor phylogeny reconstruction and mutational heterogeneity analysis. Here we introduce `Trisicell`, an algorithmic toolkit to not only (i) build scalable tumor phylogenies by boosting available phylogeny reconstruction methods to handle large scRNAseq datasets with high levels of noise, but also (ii) compare multiple phylogenies built for a tumor sample by the use of distinct data types and methods, and establish their consensus phylogeny, and (iii) compute the partition of a tumor phylogeny to evaluate the robustness of its distinct subtrees and their seeding mutations.

We evaluated `Trisicell` on a tri-modal sequencing dataset comprised of bulk whole exome and transcriptome sequencing data from single cell derived clonal sublines (each with a homogeneous mutational profile) of a melanoma cell line M4, as well as scRNAseq data from multiple cells from each subline, obtained through Smart-seq2 protocol. `Trisicell` produced robust results on this dataset with a broad consensus among the phylogenies it built for the three data types. Importantly, `Trisicell`'s partition function formulation was able to determine that specific subtrees of

[11] Briefly, Seurat uses graph-based clustering approaches such as [46, 47] that embed cells in a graph structure where edges between cells indicate similar feature expression patterns. It then attempts to partition this graph into interconnected "communities" using Louvain or smart local moving (SLM) methods [48].

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the phylogeny with distinct growth patterns have very likely been seeded by a distinct set of mutations, suggesting that these mutations are possible subclonal drivers of tumor growth (Figure 4b). In general, the phylogeny strongly correlates with in vivo growth and developmental status, supporting that some of these mutations exhibit function and drive subclonal selection during tumor progression.

We note that only a fraction of the mutation calls on the exome and transcriptome data from the M4 melanoma are shared (see Figure 3d). This may be attributed to the fact that not all genes are expressed, and that post-transcriptional processing and modification can introduce sequence variants not present in the genome [51]. Nevertheless, `Trisicell` inferred phylogenies based on mutation calls using exome and transcriptome data turned out to be nearly identical. This concordance suggests that it is primarily the expressed mutations that are under selection pressure and drive the subclonal evolution of the tumors; other mutations could be neutral or negatively selected under stress. This observation is also supported by the fact that the proportion of mutations that are expressed decreases on lineages from the root to the leaf nodes (see Figure 3b and Figure 3c). The implied functionality of expressed mutation calls for further analysis, which we leave to future studies.

The tri-modal melanoma sequencing dataset we generated for this study is of independent interest, as it offers means to assess the accuracy and robustness of tumor phylogeny inference tools based on both bulk and single cell RNAseq data. Additionally, we applied `Trisicell` to analyze six tumors from a preclinical study of M4 melanoma model, among which four were from mice treated with anti-CTLA-4 and the other two from the control treatment. A large set of cells harvested from each tumor were then subjected to Smart-seq2 protocol to produce a large scale scRNAseq dataset. We then applied `Trisicell` to this dataset and cross-evaluated the phylogeny based on mutation calls from each cell, as compared to clusters derived from gene expression profiles. We observed that the subtrees of the phylogeny are better associated with developmental status (as defined by MPS) and neural crest lineage markers than UMAP clusters, which are based on overall gene expression (Figure 5c). Even though the five main subtrees (clades) in the scRNAseq phylogeny all share driver mutations frequently observed in melanoma (i.e. in *Kras*, *Sf3b1*, and *Trp53*), each subtree features a specific combination of mutations and developmental status.

Surprisingly, anti-CTLA-4 treatment did not have an impact on cell lineages. Moreover, among the four anti-CTLA-4-treated tumors, significant subclonal selection occurred in one sample (m122), and adaptation by transcriptional change occurred in the remaining three, consistent with recent studies which suggest that therapeutic resistance is mostly caused by a combination of genetic and non-genetic mechanisms [44]. These results demonstrate that `Trisicell` inferred phylogenies allow us to trace lineages and subtrees of cells receiving a particular treatment in detail, adding a key dimension to the studies of intratumor heterogeneity.

In conclusion our results demonstrate the importance of phylogenetic analysis in studying tumor evolution and intratumor heterogeneity. This can be achieved through the innovative use of scRNAseq data, which have been primarily used for identifying gene expression differences across distinct cell types. The ability of `Trisicell` to build large scale, robust tumor phylogenies, to identify and evaluate distinct subtrees and seeding mutations greatly expands possible applications of scRNAseq data including subclonal driver mutation discovery and lineage tracing.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4 Figures

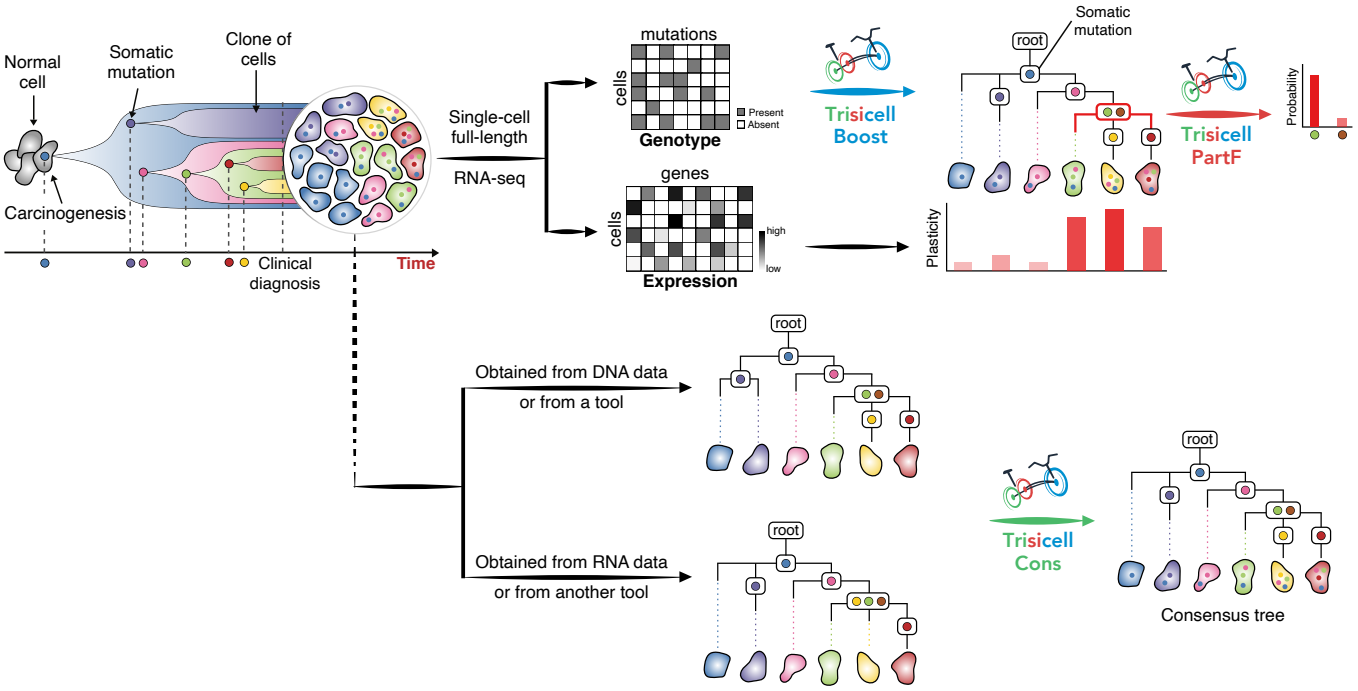

Figure 1 | Overview of `Trisicell`. Tumors are typically heterogeneous and are made of distinct populations of cells, each with a distinct set of somatic mutations. Whole single cell RNAseq (scRNAseq) data from a tumor is used for calling expressed mutations in each cell; scRNAseq data can be additionally used to derive single cell expression profiles. `Trisicell-Boost` can then be applied to cell-specific mutation sets for scalable and robust tumor phylogeny reconstruction. Additional tumor phylogeny reconstruction tools may infer alternative trees, which can all be integrated into a single consensus phylogeny through the use of `Trisicell-Cons`.[a] Expression profiles of cells harbored in distinct subtrees of the resulting phylogeny may suggest distinct subclonal phenotypes. Expressed mutations seeding these subtrees can be evaluated by `Trisicell-PartF` to identify high confidence mutations strongly associated with each subclone.

[a] Alternative phylogenies can also be obtained through other means of sequencing (e.g. DNA).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

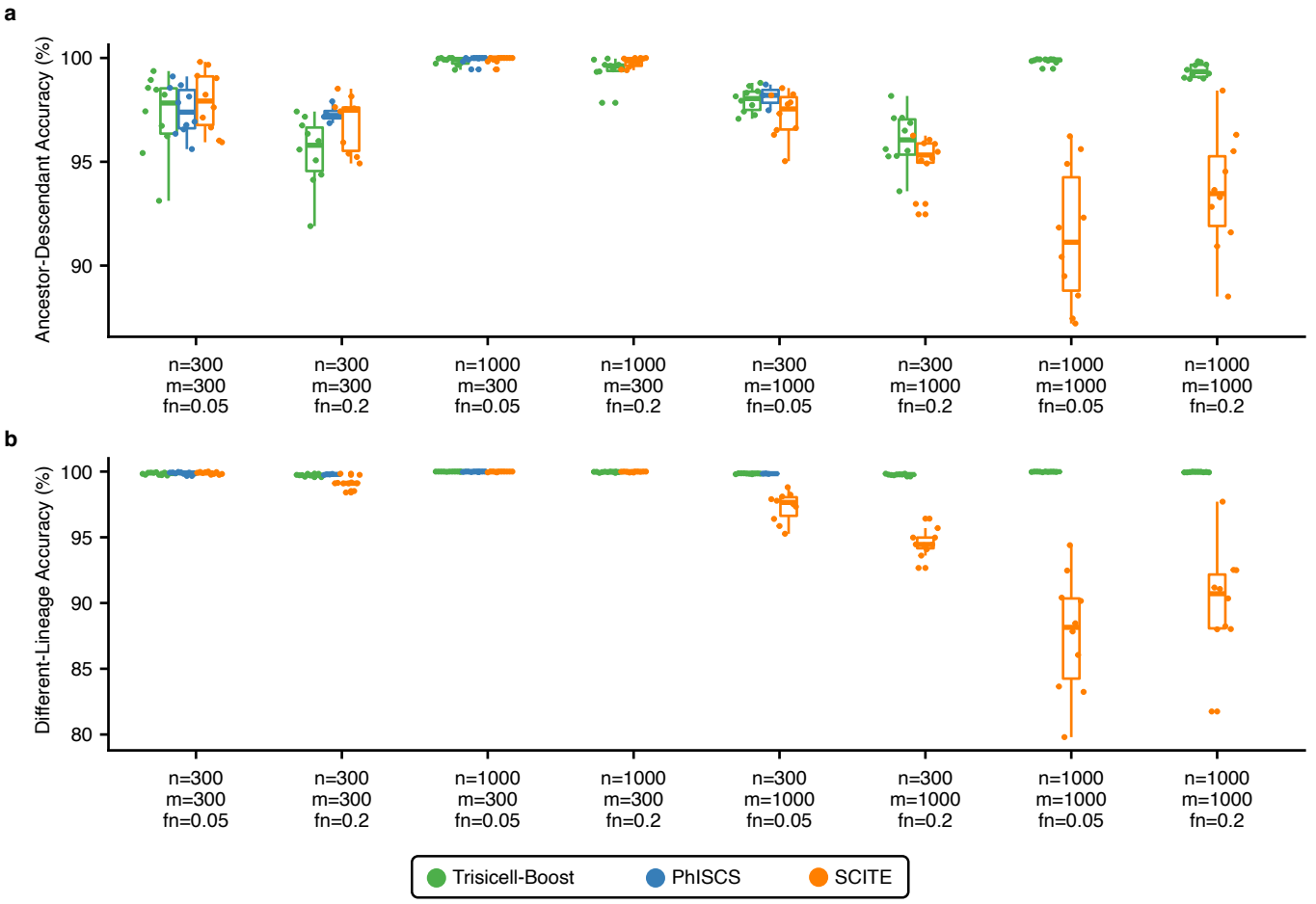

Figure 2 | Benchmarking `Trisicell-Boost` on simulated data. A Comparison of SCITE and PhISCS with their `Trisicell-Boost` boosted version[a] on simulated data, with respect to two commonly used measures of accuracy (see Supplementary Section I for definitions). Boosting SCITE and PhISCS does not result in loss of accuracy on smaller datasets that they can handle well as stand-alone tools. However, their performance is substantially improved by the boost they received from `Trisicell-Boost` on larger datasets.
Details on the simulations and accuracy measures are provided in Supplementary Sections H and I. The parameters n, m and fn respectively denote the number of simulated single cells, mutations and the false negative rate in single cell sequencing data. For each value of m (300 or 1000), 10 distinct trees of tumor evolution were simulated and then for each value of n (300 or 1000) and fn (0.05 or 0.20) genotype matrices were derived from these trees. In all experiments, the false positive rate was set to 0.001 (i.e., 0.1%) and the missing entries rate was set to 0.05 (i.e., 5%). Note that PhISCS could not terminate within the 24-hour time limit on some instances of the problem - these cases are excluded from the results. The exact number of these instances can be found in Figure S14a.

[a]Even though SCITE and PhISCS use different combinatorial approaches, they have the same objective and constraints and thus their output on smaller trees (i.e. those which are employed by `Trisicell-Boost`) are isomorphic. As a result, we do not report on boosted versions of SCITE and PhISCS separately.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

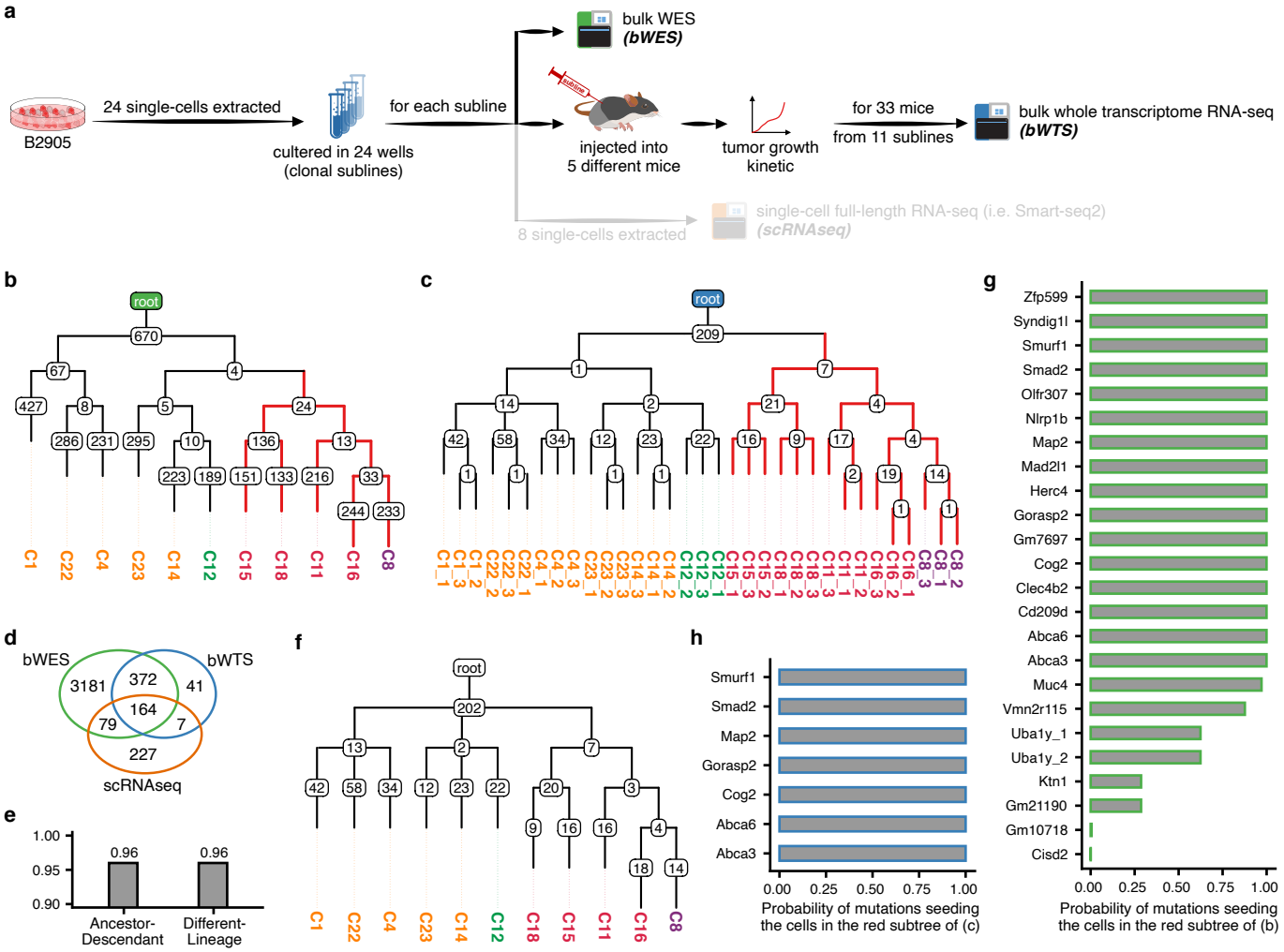

Figure 3 | Comparison of `Trisicell` built phylogenies using bulk whole transcriptome sequencing (bWTS) vs bulk whole exome sequencing (bWES) data from 11 clonal sublines.[a] **(a)** Overview of the experimental protocol specific to bWES and bWTS data generation. **(b)** bWES-based phylogeny by `Trisicell-Boost`. The number of somatic de novo mutations is shown for each node. The color used for the label of each subline indicates its "aggressiveness". In decreasing order of aggressiveness they are: red, orange, purple, and green.[b] **(c)** bWTS-based phylogeny by `Trisicell-Boost`; each subline had five independently grown replicates; the fastest growing three (see Figure S2) were sequenced. **(d)** The number of shared and private "mutation" calls across bWES and bWTS, as well as scRNAseq data (see Figure 4 for results on scRNAseq data). **(e)** Commonly used measures of accuracy for the bWTS phylogeny with respect to the bWES phylogeny (see Supplementary Section I for definitions). **(f)** `Trisicell-Cons` inferred consensus tree for bWES and bWTS phylogenies. The number of somatic de novo mutations shown in the nodes of the consensus tree are for those shared between these two phylogenies. **(g)** `Trisicell-PartF` calculated the probability of each mutation seeding the aggressive subtree of the bWES phylogeny, comprised of sublines C15, C18, C11, C16, and C8.[c] **(h)** `Trisicell-PartF` calculated the probability of each expressed mutation seeding the subtree comprised of the most rapidly growing tumors in the bWTS phylogeny.[d]

[a] See Figure S2 for the criteria used to select these 11 sublines out of 24.

[b] Aggressiveness of a subline is measured with respect to the median survival time (in days) and grouped by quartiles (red: <Q25, orange: Q25-median, purple: median-Q75, green: >Q75).

[c] A mutation with a high probability is likely to have occurred at the root of a subtree that is comprised of the sublines C15, C18, C11, C16, and C8.

[d] All expressed mutations have probability 1 but some that are not expressed have lower probabilities.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

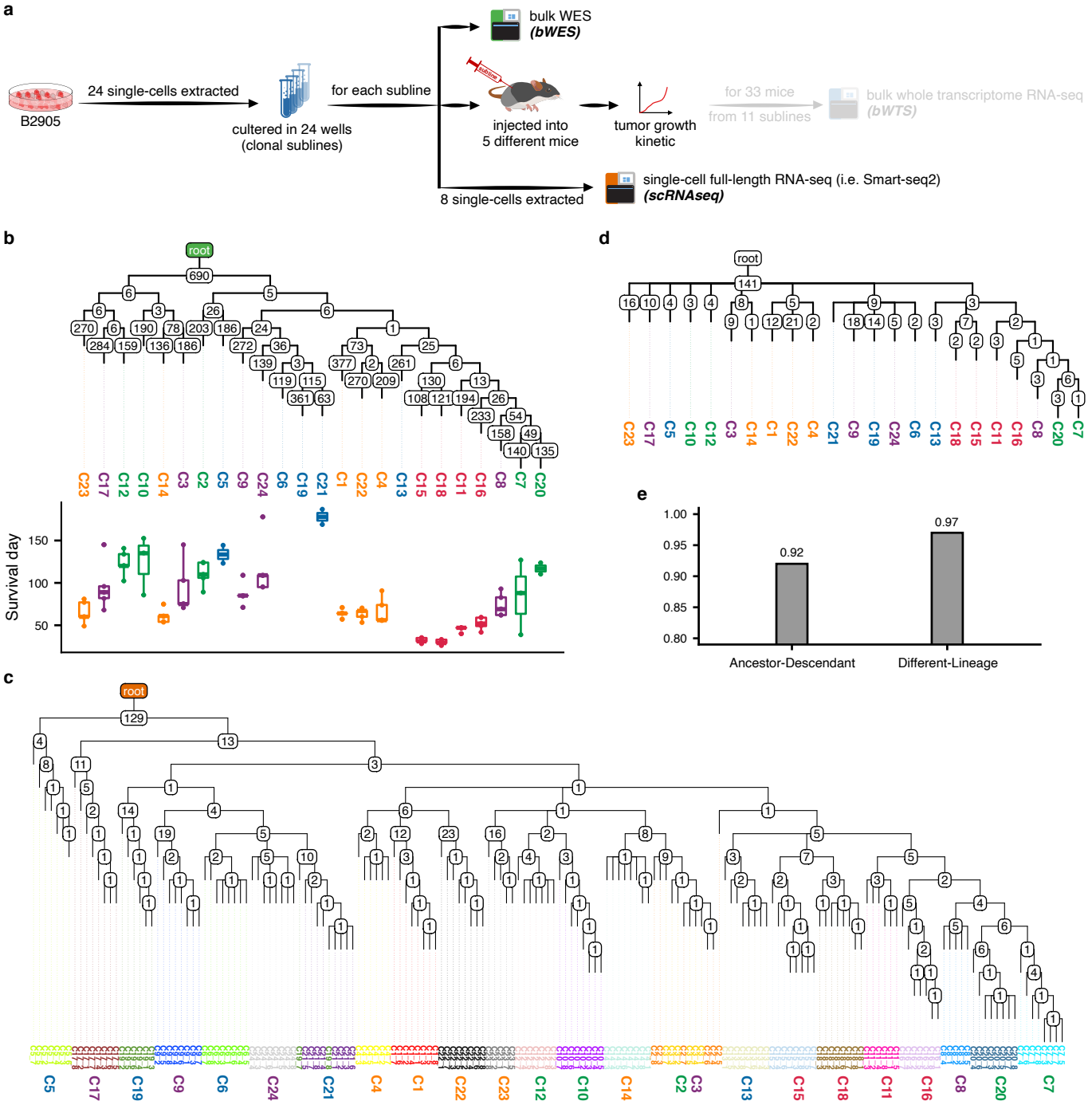

Figure 4 | Comparison of `Trisicell` built phylogenies using single cell RNA-seq (scRNAseq) vs bulk whole exome sequencing (bWES) data from 24 clonal sublines. (a) Overview of the experimental protocol specific to bWES and scRNAseq data generation. **(b)** bWES-based phylogeny inferred by `Trisicell-Boost`. The number of somatic de novo mutations is shown for each node. The color used for the label of each subline indicates its aggressiveness. In decreasing order of aggressiveness they are: red, orange, purple, green, and blue.[a] **(c)** scRNAseq-based phylogeny by `Trisicell-Boost`; from each subline $\sim$ 8 cells were sampled and sequenced. Cell colors indicate the subline they were sampled from. The cells for each subline are well clustered, with the exception of C2 and C3, which are mixed.[b] **(d)** `Trisicell-Cons` inferred consensus tree for bWES and scRNAseq phylogenies. The number of somatic de novo mutations shown in the nodes of the consensus tree is for those shared between these two phylogenies. **(e)** Commonly used measures of accuracy for the scRNAseq phylogeny with respect to the bWES phylogeny (see Supplementary Section I for definitions).

[a] Aggressiveness of a subline is measured with respect to the median survival time (in days) and grouped by quartiles (red: <Q25, orange: Q25-median, purple: median-Q75, green: >Q75, blue: <3 tumors formed).

[b] In the consensus tree of panel (d), only C3 is represented.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

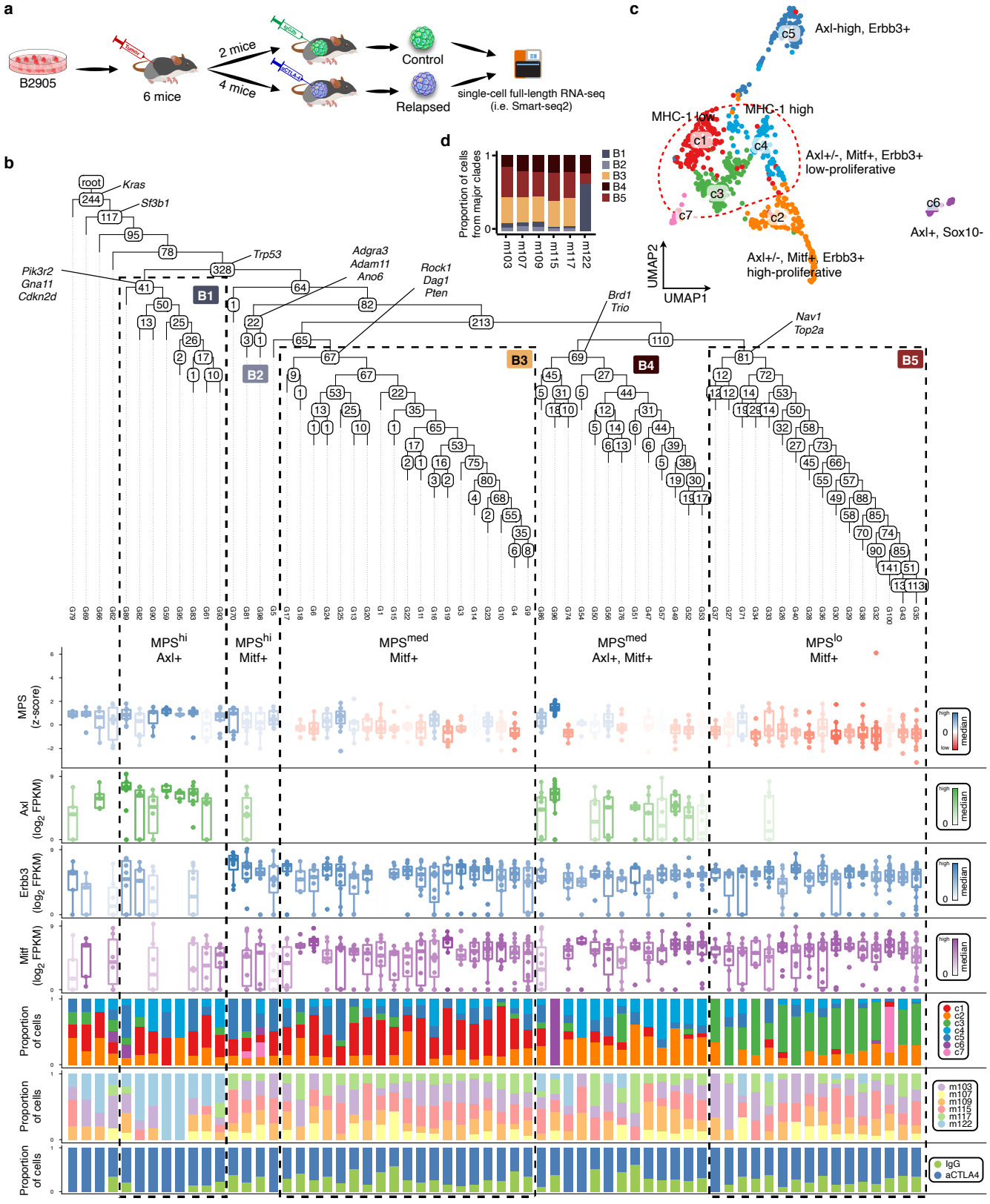

Figure 5 | **Trisicell-Boost inferred scRNAseq phylogeny for (M4 derived) tumors treated with anti-CTLA-4 immune checkpoint inhibitor. (a)** Overview of the experimental protocol specific to this scRNAseq data. **(b)** **Trisicell-Boost** inferred phylogeny where each leaf represents a group of cells with similar mutational profiles. The number of somatic de novo mutations is shown for each node. For each leaf, (i) the corresponding melanocytic plasticity signature (MPS) score and the expression values of lineage markers, (ii) *Axl*, (iii) *Erbb3*, and (iv) *Mitf* are shown in the panels below the phylogeny. Based on these, five major subtrees/clades, B1, B2, B3, B4, and B5 can be distinguished. Additional panels depict the proportion of cells in each leaf with respect to (v) expression profile based UMAP clustering, (vi) mouse id, and (vii) the type of treatment they received. **(c)** Clustering of cells with respect to overall gene expression profiles via UMAP (clusters are annotated by lineage markers). **(d)** The proportion of cells from each major subtree/clade of the phylogeny in each mouse.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5 Methods

Input data and notation. Given a sequencing experiment involves n single cells (or clonal sublines as in the case of our bWES and bWTS datasets), we denote them by $\mathbf{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n\}$. After data processing and mutation calling, suppose that m putative mutations were detected (a mutation is considered to be detected if it is reported as present in at least one single cell) and denote the set of detected mutations by $\mathbf{M} = \{M_1, M_2, \ldots, M_m\}$. The observed status of mutations in cells can be represented by a ternary "genotype matrix" I with n rows (cells), m columns (mutations), where each entry is from the set $\{0, 1, ?\}$. Here $I_{ij} = 1$ indicates that mutation M_j is reported to be present in cell $\mathcal{C}_i$ and $I_{ij} = 0$ indicates that M_j is reported to be absent in $\mathcal{C}_j$. In some cases we do not have sufficient read count information to call any of the present or absent states so we set $I_{ij} =?$ and call that a "missing entry". We assume that mutations from regions affected by deletions are filtered from the input and that commonly used infinite sites assumption (ISA) holds, that is, each mutation is acquired exactly once during tumor evolution and it is never lost in any of the descendants of the cell of its first occurrence.[12]

5.1 `Trisicell-PartF`: a partition function method to assess the likelihood of specific somatic mutations seeding a subclone

We first introduce a partition function formulation for a genotype matrix and describe a novel sampling method to calculate the marginal probability of a user-defined set of cells forming a distinct subtree of the tumor phylogeny, seeded by a user-defined mutation. This formulation allows us to assess our confidence in a particular set of cells to form a subtree/subclone, seeded by a particular mutation.[13]

Preliminaries. Let $[x] = \{1, 2, \ldots, x\}$ for any positive integer x. Recall that we fix n and m and consider n cells, m mutations to study matrices $H \in \{0, 1\}^{n \times m}$, where $H_{i,j} = 1$ indicates that i^{th} cell harbors j^{th} mutation. We use H_j to denote the j^{th} column of H. With a slight abuse of notation, we use binary vectors of length n and subsets of $[n]$ interchangeably. For example H_j denotes the profile of j^{th} mutation both as a binary vector and as a subset of cells that harbors j. Based on the definition of three-gamets rule from [23], we define $\mathcal{G} \subset \{0, 1\}^{n \times m}$ to be the set of matrices that satisfy three-gamets rule. Recall cell lineage trees introduced earlier; we further restrict this definition to *Binary Cell Lineage Tree* (BCLT). BCLTs are cell lineage trees with n observed cells as their leaves and latent subclones as their internal nodes, and each internal node has exactly two children. Notice that any cell lineage tree can be transformed to a BCLT. Also, let $\mathsf{BCLT}(H)$ denote an arbitrary BCLT consistent with H for any $H \in \mathcal{G}$. Furthermore, let $\mathcal{T}_n$ denote the set of all possible BCLTs possible for n cells. We drop the subscript n as it is fixed throughout this section. We use bold fonts for random variables to distinguish them from their non-stochastic counterparts. For example we use $\boldsymbol{X} \in \{0, 1\}^{n \times m}$ to refer to the unobserved ground truth. Finally, we use $\boldsymbol{H} \sim \mathcal{D}$ to indicate that the random variable $\boldsymbol{H}$ follows the distribution implied by some distribution $\mathcal{D}$.

Obtaining P: a matrix of independent probabilities. We define a matrix $P \in [0, 1]^{n \times m}$ based on the input matrix I and noise rates α and β. The value $P_{i,j}$ is equal to the probability that i^{th} cell harbors j^{th} mutation in the ground truth, i.e., $\boldsymbol{X}_{i,j} = 1$, given $I_{i,j}$, α and β. Formally, $P_{i,j} = P[\boldsymbol{X}_{i,j} = 1 | I_{i,j}, \alpha, \beta]$.

We can use the Bayes rule, along with the assumption that $\boldsymbol{X}_{i,j}$ is equally likely to take values 0 or 1 (before observing I) to calculate I as follows. Fix α and β, and consider the case $I_{i,j} = 1$ first:

$$\begin{aligned} P_{i,j} &= P[\boldsymbol{X}_{i,j} = 1 | I_{i,j} = 1] \\ &= \frac{P[I_{i,j} = 1 | \boldsymbol{X}_{i,j} = 1] P[\boldsymbol{X}_{i,j} = 1]}{P[I_{i,j} = 1 | \boldsymbol{X}_{i,j} = 0] P[\boldsymbol{X}_{i,j} = 0] + P[I_{i,j} = 1 | \boldsymbol{X}_{i,j} = 1] P[\boldsymbol{X}_{i,j} = 1]} \\ &= \frac{0.5 \cdot P[I_{i,j} = 1 | \boldsymbol{X}_{i,j} = 1]}{0.5 \cdot P[I_{i,j} = 1 | \boldsymbol{X}_{i,j} = 0] + 0.5 \cdot P[I_{i,j} = 1 | \boldsymbol{X}_{i,j} = 1]} \\ &= \frac{1 - \beta}{\alpha + 1 - \beta} \end{aligned} \tag{1}$$

Similarly, for $I_{i,j} = 0$ case, we have $P_{i,j} = P[\boldsymbol{X}_{i,j} = 1 | I_{i,j} = 0] = \frac{\beta}{\beta + 1 - \alpha}$.

There are more sophisticated methods for obtaining P from the output of SCS process (e.g., see [52]). Our method, presented in this section, is general and works independent of how P is calculated.

[12] Nevertheless, we have assessed the robustness of our methods in the presence of undetected deletions that result in mutational losses and possible ISA violations.

[13] The formulation naturally generalizes to any user defined set of mutations - including the empty set, in which case it represents the probability of a user defined set of cells to form a subclone without a specific seeding mutation.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In the following we present our formulation for partition function problem. Intuitively, the problem aims to measure our confidence for a template of claims about ground truth phylogenetic tree $\boldsymbol{X}$: Fix a mutation c and a subset of cells R. How confident one can be to claim that the cells in R form a subtree with the root edge of the subtree labeled with c, given P implied by our observations. Notice that this claim is equivalent to $\boldsymbol{X}_c = R$.

In the rest of this section we let $\mathcal{D}(P)$ denote the following distribution over $\{0,1\}^{n\times m}$. If $\boldsymbol{H} \sim \mathcal{D}(P)$ then the probability of $\boldsymbol{H}_{i,j} = 1$ is equal to $P_{i,j}$, independent of all other coordinates in $\boldsymbol{H}$.

Problem. Given a matrix of independent probabilities $P \in [0,1]^{n\times m}$, a mutation $c \in [m]$ and a subset of cells $R \subset [n]$, estimate $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{X}_c = R|\boldsymbol{H} \in \mathcal{G}]$. That is, what is the probability that the cth column of $\boldsymbol{H}$ is equal to R, given that $\boldsymbol{H}$ is coming from the distribution defined by P conditioned on being in $\mathcal{G}$.

Approach. First, we present a way to sample BCLTs. In short, our algorithm builds a sample tree in a bottom-up manner giving priority to nodes that correspond to the similar rows in P as well as nodes that have more mutations in common. Then a number of such samples is constructed and used to estimate $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{X}_c = R \wedge \boldsymbol{H} \in \mathcal{G}]$ and $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{H} \in \mathcal{G}]$, separately. Then our original goal $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{X}_c = R|\boldsymbol{H} \in \mathcal{G}]$ can be calculated via following equation due to the definition of conditional probability:

$$Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{X}_c = R|\boldsymbol{H} \in \mathcal{G}] = \frac{Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{X}_c = R \wedge \boldsymbol{H} \in \mathcal{G}]}{Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{H} \in \mathcal{G}]} \tag{2}$$

5.1.1 Sampling BCLTs

Algorithm 1 builds a random tree recursively. At each call to this algorithm a pair of rows is chosen based on two criteria: (I) expected number of mutations in common and (II) similarity of the two cells. Then the rows corresponding to this pair of cells are substituted by one new row representing the cells parent. Then the algorithm is recursively called for the new matrix. Let $\mathcal{E}$ denote the distribution of trees that can be output by this algorithm. Moreover, this algorithm can be modified to output the probability that each tree T is sampled, i.e., $Prob_{\boldsymbol{T}\sim\mathcal{E}}[\boldsymbol{T} = T]$.

Algorithm 1 | Sample BCLT

Input: $P \in [0,1]^{n\times m}$ and parameters,
Output: A sample BCLT.

normalization_factor $\leftarrow 0$
for $(a,b) \in [n],\ a \neq b$ **do**
  $s_{a,b} \leftarrow (P_a \odot P_b) - |P_a - P_b|$ ▷ Pairs with higher score are more probable to be chosen.
  normalization_factor $\leftarrow$ normalization_factor $+ e^{s_{a,b}}$
for $(a,b) \in [n],\ a \neq b$ **do**
  $p_{a,b} \leftarrow e^{s_{a,b}}$/normalization_factor ▷ Ensuring that $p_{.,.}$ form a valid probability distribution over pairs of rows.
Choose one $(a,b) \in [n],\ a \neq b$ according to probabilities $p_{a,b}$
parent $\leftarrow \min(P_a, P_b)$
Remove a^{th} and b^{th} rows from P and add parent
Calculate the tree for the new matrix recursively.
Add a and b as leaves and return resulting tree.

5.1.2 Estimating the probability of $\boldsymbol{X}_c = R \wedge \boldsymbol{H} \in \mathcal{G}$

Fix c, R, A, P. Given a T sampled by Algorithm 1 we define a quantity below that will be used to estimate $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{X}_c = R \wedge \boldsymbol{H} \in \mathcal{G}]$:

$$\text{numerator\_estimator}(T) = Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{X}_c = R \wedge \boldsymbol{H} \in \mathcal{G}|\mathsf{BCLT}(\boldsymbol{H}) = T]\frac{Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\mathsf{BCLT}(\boldsymbol{H}) = T]}{Prob_{\boldsymbol{T}\sim\mathcal{E}}[\boldsymbol{T} = T]}. \tag{3}$$

We show in Lemma 5.1 that this value is an unbiased estimator of Equation (2). Algorithms 1 to 3 can output $Prob_{\boldsymbol{T}\sim\mathcal{E}}[\boldsymbol{T} = T]$ (along with each T), $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\mathsf{BCLT}(\boldsymbol{H}) = T]$, and $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{X}_c = R \wedge \boldsymbol{H} \in \mathcal{G}|\mathsf{BCLT}(\boldsymbol{H}) = T]$, respectively. The proof can be found in Supplementary Section A.

Lemma 5.1. *Let v be the variance of the estimator in Equation (3). Algorithm 4 outputs a value $\hat{p}$ such that $Prob\Big[|\hat{p} - Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{X}_c = R \wedge \boldsymbol{H} \in \mathcal{G}]| > \epsilon\Big] < \delta$, when N is set to $\frac{v}{\delta\epsilon^2}$.*

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Algorithm 2 | Calculate $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\mathsf{BCLT}(\boldsymbol{H}) = T]$ used in Equation (3) given one BCLT

Input: $P \in [0,1]^{n\times m}$, a BCLT T
Output: The value of $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\mathsf{BCLT}(\boldsymbol{H}) = T]$.

1: **for** each mutation $j \in [m]$ **do**
2: **for** each subtree F in T **do**
3: Calculate $p_{j,F}$ as the probability for the event j^{th} column in $\boldsymbol{H}$ is equal to F
4: return $\prod_j \sum_F p_{j,F}$

Algorithm 3 | Calculate $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{X}_c = R \wedge \boldsymbol{H} \in \mathcal{G} | \mathsf{BCLT}(\boldsymbol{H}) = T]$ used in Equation (3) given one BCLT

Input: $P \in [0,1]^{n\times m}$, $c \in [m]$, $R \subset [n]$, a BCLT T
Output: The value of $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{X}_c = R \wedge \boldsymbol{H} \in \mathcal{G} | \mathsf{BCLT}(\boldsymbol{H}) = T]$.

1: **for** each subtree F in T **do**
2: Calculate $p_{c,F}$ as the probability for the event c^{th} column in $\boldsymbol{H}$ is equal to F
3: return $\frac{p_{c,R}}{\sum_F p_{c,F}}$

Algorithm 4 | Estimate $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{X}_c = R \wedge \boldsymbol{H} \in \mathcal{G}]$

Input: $P \in [0,1]^{n\times m}$, $c \in [m]$, $R \subset [n]$ N sample size
Output: An estimate to $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{X}_c = R \wedge \boldsymbol{H} \in \mathcal{G}]$.

1: **for** $i \in [N]$ **do**
2: $T_i \leftarrow$ a sample BCLT using Algorithm 1
3: $e_i \leftarrow estimate(T_i)$ through Equation (3) and Algorithms Algorithms 2 and 3
4: return Average$(e_1, \ldots, e_N)$

5.1.3 Estimate $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{H} \in \mathcal{G}]$

The approach for estimating $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{H} \in \mathcal{G}]$ is similar to the previous section, except that we will use the following estimator (instead of numerator_estimator in Equation (3)).

$$\mathsf{denominator\_estimator}(T) = \frac{Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\mathsf{BCLT}(\boldsymbol{H}) = T]}{Prob_{\boldsymbol{T}\sim\mathcal{E}}[\boldsymbol{T} = T]} \tag{4}$$

Lemma 5.2 proves a theoretical guarantee for denominator_estimator similar to Lemma 5.1. The proof can be found in Supplementary Section A.

Algorithm 5 | Estimate $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{H} \in \mathcal{G}]$

Input: $P \in [0,1]^{n\times m}$, $c \in [m]$, $R \subset [n]$ N sample size
Output: An estimate to $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{X}_c = R \wedge \boldsymbol{H} \in \mathcal{G}]$.

1: **for** $i \in [N]$ **do**
2: $T_i \leftarrow$ a sample BCLT using Algorithm 1
3: $e_i \leftarrow denominator\_estimator(T_i)$ through Equation (4).
4: return Average$(e_1, \ldots, e_N)$

Lemma 5.2. *Let v be the variance of the estimator in Equation (4). Algorithm 4 outputs a value $\hat{p}$ such that $Prob\Big[|\hat{p} - Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{H} \in \mathcal{G}]| > \epsilon\Big] < \delta$, when N is set to $\frac{v}{\delta\epsilon^2}$.*

5.1.4 Putting all together

Given P, c, and R, our algorithm estimates $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{X}_c = R | \boldsymbol{H} \in \mathcal{G}]$ in four phases as follows:

1. Draws a predetermined number of sample BCLTs (Algorithm 1).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2. Estimates $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{X}_c = R \wedge \boldsymbol{H} \in \mathcal{G}]$ with the guarantee given in Lemma 5.1 (Algorithm 4).
3. Estimates $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{H} \in \mathcal{G}]$ with the guarantee given in Lemma Lemma 5.2 (Algorithm 4).
4. Outputs an estimate for $Prob_{\boldsymbol{H}\sim\mathcal{D}(P)}[\boldsymbol{X}_c = R \wedge \boldsymbol{H} \in \mathcal{G}]$ by dividing the estimate from step 2 by the estimate from step 3.

An experimental evaluation of this method on simulated data can be found in Supplementary Section J.

5.2 `Trisicell-Cons`: an algorithm for building consensus tumor phylogenies

As mentioned earlier, it is desirable to establish a "consensus" between two or more phylogenies that represent the evolution of a tumor sample. Since the set $\mathbf{C}$ of cells (or clonal sublines) of the input genotype matrix representing a tumor is fixed but the mutation calls may vary, we focus on establishing a consensus between cell lineage trees rather than mutation trees.[14] Thus we introduce a simple, deterministic, exact algorithm for building a consensus tree of two or more cell lineage trees, which runs in time quadratic with the input size. The ancestor-descendant relationships between cells implied by the consensus of multiple cell lineage trees, each obtained through distinct methods and data types, are likely to be more reliable. Edges maintained in the consensus tree are also more likely to be labeled by mutations that "seed" the associated subtree.

Given sequencing data from n cells, labeled $\mathbf{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots \mathcal{C}_n\}$, let rooted trees T^a and T^b represent two distinct evolutionary scenarios for $\mathbf{C}$. We assume for both that: (i) each edge has a positive "weight", (ii) each node has 0, 1 or more labels and each leaf node has at least 1 label, (iii) each internal node without a label has at least two children, (iv) each label occurs exactly once in each tree, (v) the set of labels are identical in the two trees.[15] The weight of an edge represents the number of mutations that distinguish a (possibly inferred) parent node from its child. We call such a tree a *generalized cell lineage tree* and as the label sets of the two trees are identical, we say they are *comparable*.

We now define the *merge-with-parent* operation on any given non-root node v of a tree: the operation merges v with its parent w and deletes edge (w, v) so that (a) the new label set of w becomes the union of the original label sets of w and v, and (b) the new set of children of w is the union of the original sets of children of w and v minus v (as it ceases to exist). Observe that after *merge-with-parent* operations, the sets of labels on each node remain disjoint from one another. The *cost* of a merge-with-parent operation is defined to be the weight of the edge (w, v). Our goal is to perform a sequence of merge-with-parent operations on the two trees with minimum total cost so that the resulting trees are *isomorphic* – and can be thought of as a single tree T^c, which we call *the consensus tree* of T^a and T^b.

Definition 5.1. *Given a pair of comparable cell lineage trees and a node v in either tree, let $L(v)$ denote the complete set of labels in the subtree rooted at v (including the labels of v itself). Then the two trees are isomorphic, if there is a bijection between their nodes such that for any pair of matched nodes v and w, $L(v) = L(w)$.*

For any two nodes v, w in one tree, $L(w) \subseteq L(v)$ if w is a descendant of v, and, if there is no ancestor/descendant relationship between v and w, $L(v)$ and $L(w)$ are disjoint. Also, note that a matched leaf in T^a must be matched to a leaf in T^b, and their labels must be identical. This isomorphism implies that the two trees look identical:

Lemma 5.3. *Any two comparable cell lineage trees T^a and T^b are isomorphic if and only if there is a bijection between their nodes where (i) the label sets of matched nodes are identical, and (ii) children of matched nodes are also matched.*

The proofs of Lemma 5.3 and the following lemmata in this section can be found in Supplementary Section B.

We now describe our algorithm to compute T^c. For simplicity, we consider a node to be its own lowest ancestor.

First Step. For any pair of labels x and y where x is in an ancestor of the node of y in T^a, but not in T^b, the algorithm performs subsequent merge-with-parent operations in T^b, from the node of x, up to the lowest common ancestor of the nodes of x and y so x becomes an ancestor of y in T^b. (A symmetric action happens when T^a and T^b are reversed.) Note that, if x is a proper ancestor of y (i.e., they are not in the same node) in T^a and y is a proper ancestor of x in T^b, the sequences of merge operations triggered in the two trees will end up placing x and y in the same node in both T^a and T^b. We now show that the above operations are necessary.

Lemma 5.4. *Let x reside in an ancestor of the node containing y in T^a (for simplicity, we can say x is an ancestor of y) and not in T^b. Let z be the lowest common ancestor or x and y in T^b. Then, any construction of T^c must involve removing the edges between x and z in T^b through merge-with-parent operations.*

We can see that once the first step concludes, ancestor-descendant relationships are consistent across the two trees.

[14] In addition, the problem of building a consensus across mutation trees through "edit operations" including deletion or move of mutation labels from one edge to another is NP-hard [23].The single polynomial time computable "consensus" construction approach allows edit operations on leaf nodes only [37].

[15] The trees could have been built by two distinct methods on the same set of mutation calls on each single cell, or could have been built by the same method using different subsets of mutations.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Lemma 5.5. *After the first step, any two labels x and y (a) are either in the same node in both trees, or (b) have the same ancestor-descendant relationship in both trees, or (c) have no ancestor-descendant relationship in either tree.*

The output of the first step will thus be two comparable cell lineage trees in which no pair of labels have "different" ancestor-descendant relationship between the trees. We will call them T'^a and T'^b.

Second Step. Based on the observation above, the algorithm now replaces each node's label set with a single new label. The result of this step will be two trees T''^a and T''^b, respectively corresponding to trees T'^a and T'^b, in which (1) each leaf node has exactly 1 label and each internal node has either 1 or 0 labels, where each label represents a set of labels from the original trees; (2) the label set of the two trees are identical. As such, T''^a and T''^b are (still) comparable cell lineage trees that satisfy the following.

Lemma 5.6. *For each leaf v in a tree, there exists exactly one leaf w in the other tree where $L(v) = L(w)$. For each internal node v in each tree there exists at most one internal node w in the other tree such that $L(v) = L(w)$.*

We call a node v in each tree for which there is a corresponding node w in the other tree such that $L(v) = L(w)$ a *matched* node. All other nodes are said to be *unmatched*. The following holds by definition.

Lemma 5.7. *If all nodes of any of the trees are matched and the trees have the same number of nodes then the two trees are isomorphic.*

Lemma 5.8. *For each tree, any node v with a label l must have a matched node w in the other tree with label l.*

Third Step. The algorithm processes T''^a and T''^b to compute their consensus T''^c, which is convertible to T^c. W.l.o.g., consider any matched internal node v in T''^a for which the subtree rooted at v has no matched internal nodes (including when all children of v are leaves). Let w be the node matched to v in T''^b. The algorithm computes the consensus of the subtrees rooted at v and w (these subtrees must be comparable cell lineage trees) by performing merge-with-parent operations on all internal nodes such that the subtrees end up as *star trees*; a star tree is a rooted tree with no internal nodes.

In case v and w are the roots of respective trees T^a and T^b the algorithm terminates at this point since the resulting trees must be isomorphic.[16] If they are not roots, then the algorithm "replaces" the subtrees rooted by v and w by respective leaf nodes v and w, both with a new label l and iterates.

Lemma 5.9. *The algorithm described above computes a consensus tree T^c in $((|T^a| + |T^b|)^2)$ time and linear space.*

Lemma 5.10. *The algorithm terminates with isomorphic trees.*

We can now argue about the optimality of the algorithm.

Lemma 5.11. *Each merge-with-parent operation by the algorithm is necessary for obtaining a consensus tree.*

Since all the edges that the algorithm removes are those that must be removed, the Corollary below follows.

Corollary 5.1. *The algorithm terminates with a consensus by removing edges whose total cost is minimum possible.*

Corollary 5.2. *The algorithm can be generalized to $k > 2$ input trees by recursively constructing the consensus tree of the first $k - 1$ input trees and then building the consensus of that tree and the final input tree.*

5.3 `Trisicell-Boost`: a divide and conquer booster for scalable tumor phylogeny inference

Due to false mutation calls and missing entries, the matrix I is usually different from the true genotype matrix and does not admit perfect phylogeny. There are several methods that aim to *correct* (and possibly impute missing) entries of I so that the resulting matrix Y admits a perfect phylogeny, by minimizing the weighted number of $1 \to 0$ "flips" (which represent false positives) and $0 \to 1$ "flips" (which represent false negatives) [14, 15, 22, 23, 24].

Let $\mathcal{T}$ denote (one of) the most likely mutation tree(s) of tumor evolution, where the log-likelihood of a tree is defined based on a scoring model commonly used in the existing tree inference methods. While the existing methods such as PhISCS or SCITE can in principle infer $\mathcal{T}$, these methods do not scale well in terms of the running time on large datasets. As we show, for example, on simulated genotype matrices of size 1000×1000, these tools either do not terminate within a time limit of 24 hours (e.g. PhISCS, ScisTree) or yield trees that have a sizeable number of mutational pairs that are misplaced in terms of commonly used ancestor-descendant and different-lineage measures (e.g. SCITE). This motivated us to introduce `Trisicell-Boost`, a novel method for tumor phylogeny reconstruction that can handle

[16] Note that at the time of termination, all subtrees that have been replaced by a single leaf node should be reinstated.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

such (simulated) datasets within the time limit and achieve high accuracy. As will be shown, `Trisicell-Boost` is robust to various types of noise, including the presence of mutations affected by copy number aberrations, possibly resulting in ISA violations. On "real" melanoma WES data involving clonal sublines, the `Trisicell-Boost` inferred phylogenies are not only (almost) identical to that obtained by (much slower but provably optimal) PhISCS-BnB, but also very similar to that it obtains on matching WTS data. Finally, these phylogenies, as well as that obtained for on a melanoma scRNAseq dataset, highly associate with phenotypes such as cell differentiation status and tumor growth.

In designing `Trisicell-Boost` we extensively rely on the observation that any subset of mutations from $\mathbf{M}$ also forms a mutation tree [53]. We first sample a given number of submatrices of I, each obtained by taking a random subset of z columns (mutations) of I. In our experiments we sample $\approx 50 \cdot \frac{m^2}{z^2}$ submatrices, implying that, on average, each pair of mutations appears together in approximately 50 submatrices. This plays an important role in obtaining reliable *weights* that are defined and used below. Although in principle the constant 50 can be any other positive value depending on computational resources and time, we recommend using at least 10. Similarly, we recommend that each sampled submatrix has at least 10 columns (i.e., $z \geq 10$). We then run PhISCS on each of the submatrices and obtain a mutation tree for the corresponding set of sampled mutations. Note that on the smaller input matrices, PhISCS, under the model that it uses, can provide a guarantee of optimality of the reported solutions. This is the main reason why it was our primary choice for obtaining trees on sampled submatrices, although any other tool that performs well on smaller inputs (e.g., SCITE, OncoNEM, ScisTree or several others) can be used for this purpose.

Once the tree for each sampled submatrix is obtained, we do the following: for each pair of mutations M_i and M_j and each submatrix in which both M_i and M_j were sampled, we obtain *phylogenetic dependency* between them. Given a submatrix where both are present, M_i and M_j can have one of the following dependencies in the inferred tree:

1. **Ancestor-descendant** where (the first occurence) node of M_i is an ancestor of that of M_j. We denote by a_{ij} the total number of submatrices for which M_i and M_j have this dependency in their inferred trees.
2. **Descendant-ancestor** where node of M_i is a strict descendant of the node of M_j. As per above, we denote the number of submatrices for which this is a case as d_{ij}. Note that $d_{ij} = a_{ji}$.
3. **Same-node** where both M_i and M_j occur (for the first time) at the same node. We denote the number of reported trees in which this holds by s_{ij}.
4. **Different-lineage** where M_i and M_j occur at nodes that are on different tree lineages (i.e., their nodes are different and neither of them is an ancestor of the other). We denote by l_{ij} the number of submatrices for which M_i and M_j are in this dependency in their inferred trees.
5. **Unknown dependency** where the phylogeny inference tool used reports that at least one of M_i and M_j is absent in all cells.

Let $t_{ij} = a_{ij} + d_{ij} + l_{ij} + s_{ij}$. Based on the above, we define three *weights*: A_{ij} (*ancestor-descendant weight*), D_{ij} (*descendant-ancestor weight*), and L_{ij} (*different-lineage weight*), which are respectfully defined as:

$$A_{ij} = \frac{a_{ij} + \frac{s_{ij}}{2}}{t_{ij}}, \qquad D_{ij} = \frac{d_{ij} + \frac{s_{ij}}{2}}{t_{ij}}, \qquad L_{ij} = \frac{l_{ij}}{t_{ij}}.$$

Consider now an arbitrary mutation tree T of size $m + 1$ where the root node is mutation-free representing a population of healthy cells, at each node we have exactly one mutation label from $\mathbf{M}$ and no mutational label appears more than once. We say that mutation M_i is an ancestor of M_j if and only if node labeled by M_i is an ancestor of node labeled by M_j. Analogously we define M_i being descendant of or on different lineages compared of M_j. Using the above notation we can define score $S(T)$ of the tree T as follows:

$$S(T) = \sum_{i=1}^{m} \sum_{j=i+1}^{m} \delta_{ij}^{(A)} \cdot A_{ij} + \delta_{ij}^{(D)} \cdot D_{ij} + \delta_{ij}^{(L)} \cdot L_{ij}, \tag{5}$$

where $\delta_{ij}^{(A)}$ is an indicator variable with value set to 1 if and only if mutation M_i is ancestor of mutation M_j in T. Variables $\delta_{ij}^{(D)}$ and $\delta_{ij}^{(L)}$ are defined analogously and indicate whether M_i is descendant of M_j (variable $\delta_{ij}^{(D)}$) or whether these two mutations occur on different lineages (variable $\delta_{ij}^{(L)}$).

We leave finding a tree T for which $S(T)$ is maximized as an open problem and instead propose a fast heuristic for building a tree iteratively by adding mutations in a specified order and one at a time. Full details of this heuristic are provided in Supplementary Section C.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5.4 Wet lab data generation to validate `Trisicell`

The "M4" mouse model of UV-induced melanoma used in this study is described in [31]. In brief, pups of Hgf-transgenic C57BL/6 mice received UV irradiation at post-natal day three, and melanoma occurred in 8-12 months in some mice. One of the tumors was harvested from a mouse and subjected to in vitro culture. After consecutive passage, the melanoma line was built and named B2905 cell line.

5.4.1 Establishing and characterization of the 24 clonal sublines of B2905

The parental B2905 cells were subjected to fluorescence-aided cell sorting (FACS) to sort single cells into individual wells of a 96-well plate, which was brought to in vitro culture. Cells were found growing in twenty-four wells. They were expanded by consecutive passages to build the twenty-four clonal sublines derived from single cells of B2905. To characterize each subline, exome and single cell RNA were analyzed. For exome analysis, genomic DNA was prepared from cells harvested from each subline and subjected to exome capture protocol, followed by whole exome sequencing on Illumina HighSeq platform. For single cell analysis, 16 single cells from each subline were sorted to individual well of 96-well plate by FACS. cDNA library was prepared from the RNA of single cells in each well following Smart-seq2 protocol [32]. After quality control check, the cDNA libraries of 8 selected cells from each subline (in total $8 \times 24 = 192$) were sequenced on the NovaSeq platform. For downstream analysis and mutation calling, 175 cells that had sufficient coverage across all genes were selected.

5.4.2 Analysis of in vivo tumors derived from the clonal sublines of B2905

To test the growth kinetics in vivo, one million cells of each clonal subline were implanted into five syngeneic C57BL/6 mice subcutaneously. The size of tumors was measured twice a week. Eleven sublines showed 100% tumor-take rate within 90 days. Three tumors from each of these sublines were selected as triplicates. They were subjected to RNA preparation, followed by mRNA sequencing.

5.4.3 single cell sequencing of melanoma in the preclinical study

For the preclinical study of anti-CTLA-4, one million cells of B2905 were subcutaneously implanted into twenty C57BL/6 mice. When the tumors reaching 75 mm^3, the mice were randomized to two groups of equal number to receive either mouse anti-mouse CTLA-4 antibody (clone 9D9, Bio X Cell) or mouse IgG2b as isotype control (clone MPC-11, Bio X Cell). When the tumor reached around 500 mm^3, the mouse was euthanized for tumor harvest. The two control group tumors were harvested at day 41 and 54 and from anti-CTLA-4 group, one tumor was harvested at day 54, another at day 75 and final two at day 85 (see Supplementary Section F for tumor size detail). A single cell suspension was prepared from the tumor as described previously [54], stained with antibodies, nuclear dye, and viability dye, and then subjected FACS to sort $CD45^-$;$CD31^-$ nucleated alive single cells into individual wells of 96-well plates. A total of 1008 cells from four anti-CTLA-4 treated tumors and 384 cells from the control tumors were collected. cDNA library was prepared from the RNA of single cells in each well following Smart-seq2 protocol [32]. After a quality control check, the cDNA libraries from 1248 cells (960 from four anti-CTLA-4 treated tumors and 288 cells from the control tumors) were sequenced on the NovaSeq platform. For downstream analysis and mutation calling, 764 cells (595 from four anti-CTLA-4 treated tumors and 169 cells from the control tumors) that had sufficient coverage across all genes were selected.

5.5 Somatic variant calling pipeline

Several studies have explored SNVs/Indels in scRNAseq data to decipher the heterogeneity of cell populations [55, 56, 57, 58, 59]. We used GATK Best Practices [60] to call mutations from scRNAseq, bWTS and bWES datasets.[17] Briefly, after trimming and quality control of reads by FastQ (v0.11.9) [61] and Trim Galore (v0.6.5), raw reads were mapped to the mouse genome (GRCm38, mm10) using STAR (v2.7.3a) [62] in two-pass mode which improves the mapping quality at junction sites. Then Picard (v2.22.3) was used for removing duplicated mapped reads and doing indel realignment and base re-calibration. During base quality recalibration, dbSNP variants were used as known sites. In the next step, HaplotypeCaller (v3.8) [63] was used to call the mutations for every single cell separately in GVCF mode. After that, GenotypeGVCFs was used to call the mutations in a joint mode (as pseudo-bulk). These steps keep specificity and sensitivity of calling mutations high. Afterwards, the mutations were annotated using ANNOVAR (v2019.10.24) [64]. Following that, mutations that marked as unknown and non-exonic were filtered out. For further confidence, mouse SNPs/Indels filtration was applied to the VCF files based on normal mouse spleen SNPs in addition to a curated in-house database. Finally, somatic-mutations associated with less than 10 reads in at least 2 cells were filtered out. (see Supplementary Section M for more details)

[17] Depends on the data there are some modifications in the pipeline.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5.6 single cell gene expression quantification and quality control

RSEM (v1.3.2) [65] was used to extract expected counts, TPM and FPKM information. The Seurat package (v4.0.0) [45] was used for clustering and UMAP analysis. The low-quality cells were removed if any of the following is true: 1) percentage of mitochondrial genes is above 20%; 2) the number of reads is less than 500,000; 3) nFeature_RNA $<$ 1000 (see Supplementary Sections F and L for more details)

To remove batch effect due to different datasets, we used the functions of `FindIntegrationAnchors` and `IntegrateData` in Seurat package. Clustering analysis was performed with `FindClusters` (resolution = 0.5).

5.7 Differential gene expression, pathway, and gene set enrichment analyses

Differential gene expression was performed using DESeq2 (v1.30.0) [66]. The contrasts include aCTLA4 cells vs IgG cells in each branch in the phylogenetic tree. FDR was used as 0.1.

The rank list of differentially expressed genes (DEGs) was used to perform gene set enrichment analysis (GSEA) with the R package FGSEA (v1.16.0) [67]. We performed GSEA with the Hallmark gene sets available in the R package MSigDB (v7.2.1) [68]. Results are presented in Supplementary Section E.

Data availability

Sequence data have been deposited at NCBI GEO with the accession code GSExxxxxx.

Code availability

`Trisicell` is available at `https://trisicell.readthedocs.io`

Acknowledgements

We thank Maria Hernandez, Kimia Dadkhah and Yongmei Zhao (at Frederick National Laboratory for Cancer Research) for helping us sequence our samples. This work is supported in part by the Intramural Research Program of the National Institutes of Health, National Cancer Institute and utilized the computational resources of the NIH Biowulf high performance computing cluster (`http://hpc.nih.gov`) and Gurobi (`http://www.gurobi.com`) to solve some optimization problems. F.R.M. was supported in part by Indiana U. Grand Challenges Precision Health Initiative.

Competing interests

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

[1] Rebecca A. Burrell and Charles Swanton. Tumour heterogeneity and the evolution of polyclonal drug resistance. *Molecular Oncology*, 8(6):1095–1111, July 2014. doi: 10.1016/j.molonc.2014.06.005. URL https://doi.org/10.1016/j.molonc.2014.06.005.

[2] Ash A Alizadeh, Victoria Aranda, Alberto Bardelli, Cedric Blanpain, Christoph Bock, Christine Borowski, Carlos Caldas, Andrea Califano, Michael Doherty, Markus Elsner, Manel Esteller, Rebecca Fitzgerald, Jan O Korbel, Peter Lichter, Christopher E Mason, Nicholas Navin, Dana Pe'er, Kornelia Polyak, Charles W M Roberts, Lillian Siu, Alexandra Snyder, Hannah Stower, Charles Swanton, Roel G W Verhaak, Jean C Zenklusen, Johannes Zuber, and Jessica Zucman-Rossi. Toward understanding and exploiting tumor heterogeneity. *Nature Medicine*, 21(8): 846–853, August 2015. doi: 10.1038/nm.3915. URL https://doi.org/10.1038/nm.3915.

[3] Mel Greaves. Evolutionary Determinants of Cancer. *Cancer Discovery*, 5(8):806–820, July 2015. doi: 10.1158/2159-8290.cd-15-0439. URL https://doi.org/10.1158/2159-8290.cd-15-0439.

[4] Ibiayi Dagogo-Jack and Alice T. Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15(2):81–94, November 2017. doi: 10.1038/nrclinonc.2017.166. URL https://doi.org/10.1038/nrclinonc.2017.166.

[5] Peter C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, October 1976. doi: 10.1126/science.959840. URL https://doi.org/10.1126/science.959840.

[6] Marco L. Leung, Alexander Davis, Ruli Gao, Anna Casasent, Yong Wang, Emi Sei, Eduardo Vilar, Dipen Maru, Scott Kopetz, and Nicholas E. Navin. Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome Research*, 27(8):1287–1299, May 2017. doi: 10.1101/gr.209973.116. URL https://doi.org/10.1101/gr.209973.116.

[7] Russell Schwartz and Alejandro A. Schäffer. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, 18(4):213–229, February 2017. doi: 10.1038/nrg.2016.170. URL https://doi.org/10.1038/nrg.2016.170.

[8] Jack Kuipers, Katharina Jahn, and Niko Beerenwinkel. Advances in understanding tumour evolution through single-cell sequencing. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1867(2):127–138, April 2017. doi: 10.1016/j.bbcan.2017.02.001. URL https://doi.org/10.1016/j.bbcan.2017.02.001.

[9] Salem Malikic. *Inference of tumor subclonal composition and evolution by the use of single-cell and bulk DNA sequencing data*. PhD thesis, Applied Sciences: School of Computing Science, 2019. URL https://summit.sfu.ca/item/19469.

[10] 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, and Peter Van Loo and. The evolutionary history of 2,658 cancers. *Nature*, 578(7793):122–128, February 2020. doi: 10.1038/s41586-019-1907-7. URL https://doi.org/10.1038/s41586-019-1907-7.

[11] Bora Lim, Yiyun Lin, and Nicholas Navin. Advancing Cancer Research and Medicine with Single-Cell Genomics. *Cancer Cell*, 37(4):456–470, April 2020. doi: 10.1016/j.ccell.2020.03.008. URL https://doi.org/10.1016/j.ccell.2020.03.008.

[12] Sophie Tritschler, Maren Büttner, David S. Fischer, Marius Lange, Volker Bergen, Heiko Lickert, and Fabian J. Theis. Concepts and limitations for learning developmental trajectories from single cell genomics. *Development*, 146(12):dev170506, June 2019. doi: 10.1242/dev.170506. URL https://doi.org/10.1242/dev.170506.

[13] Daniel E. Wagner and Allon M. Klein. Lineage tracing meets single-cell omics: opportunities and challenges. *Nature Reviews Genetics*, 21(7):410–427, March 2020. doi: 10.1038/s41576-020-0223-2. URL https://doi.org/10.1038/s41576-020-0223-2.

[14] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome Biology*, 17 (1), May 2016. doi: 10.1186/s13059-016-0936-x. URL https://doi.org/10.1186/s13059-016-0936-x.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[15] Edith M. Ross and Florian Markowetz. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biology*, 17(1), April 2016. doi: 10.1186/s13059-016-0929-9. URL https://doi.org/10.1186/s13059-016-0929-9.

[16] Hamim Zafar, Anthony Tzen, Nicholas Navin, Ken Chen, and Luay Nakhleh. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biology*, 18(1), September 2017. doi: 10.1186/s13059-017-1311-2. URL https://doi.org/10.1186/s13059-017-1311-2.

[17] 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, October 2019. doi: 10.1101/gr.243121.118. URL https://doi.org/10.1101/gr.243121.118.

[18] Mohammed El-Kebir. SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics*, 34(17):i671–i679, September 2018. doi: 10.1093/bioinformatics/bty589. URL https://doi.org/10.1093/bioinformatics/bty589.

[19] Salem Malikić, Farid Rashidi Mehrabadi, Erfan Sadeqi Azer, Mohammad Haghir Ebrahimabadi, and S. Cenk Sahinalp. Studying the history of tumor evolution from single-cell sequencing data by exploring the space of binary matrices. *bioRxiv*, July 2020. doi: 10.1101/2020.07.15.204081. URL https://doi.org/10.1101/2020.07.15.204081.

[20] Hans L. Bodlaender, Michael R. Fellows, Michael T. Hallett, H.Todd Wareham, and Tandy J. Warnow. The hardness of perfect phylogeny, feasible register assignment and other problems on thin colored graphs. *Theoretical Computer Science*, 244(1-2):167–188, August 2000. doi: 10.1016/s0304-3975(98)00342-9. URL https://doi.org/10.1016/s0304-3975%2898%2900342-9.

[21] Duhong Chen, O. Eulenstein, D. Fernandez-Baca, and M. Sanderson. Minimum-Flip Supertrees: Complexity and Algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(2):165–173, April 2006. doi: 10.1109/tcbb.2006.26. URL https://doi.org/10.1109/tcbb.2006.26.

[22] Yufeng Wu. Accurate and efficient cell lineage tree inference from noisy single cell data: the maximum likelihood perfect phylogeny approach. *Bioinformatics*, August 2019. doi: 10.1093/bioinformatics/btz676. URL https://doi.org/10.1093/bioinformatics/btz676.

[23] 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*, 29(11):1860–1877, October 2019. doi: 10.1101/gr.234435.118. URL https://doi.org/10.1101/gr.234435.118.

[24] Erfan Sadeqi Azer, Farid Rashidi Mehrabadi, Salem Malikić, Xuan Cindy Li, Osnat Bartok, Kevin Litchfield, Ronen Levy, Yardena Samuels, Alejandro A Schäffer, E Michael Gertz, Chi-Ping Day, Eva Pérez-Guijarro, Kerrie Marie, Maxwell P Lee, Glenn Merlino, Funda Ergun, and S Cenk Sahinalp. PhISCS-BnB: a fast branch and bound algorithm for the perfect tumor phylogeny reconstruction problem. *Bioinformatics*, 36(Supplement_1):i169–i176, July 2020. doi: 10.1093/bioinformatics/btaa464. URL https://doi.org/10.1093/bioinformatics/btaa464.

[25] Simone Zaccaria and Benjamin J. Raphael. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. *Nature Biotechnology*, 39(2):207–214, September 2020. doi: 10.1038/s41587-020-0661-6. URL https://doi.org/10.1038/s41587-020-0661-6.

[26] Hamim Zafar, Nicholas Navin, Luay Nakhleh, and Ken Chen. Computational approaches for inferring tumor evolution from single-cell genomic data. *Current Opinion in Systems Biology*, 7:16–25, February 2018. doi: 10.1016/j.coisb.2017.11.008. URL https://doi.org/10.1016/j.coisb.2017.11.008.

[27] Simone Picelli, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, 10(11):1096–1098, September 2013. doi: 10.1038/nmeth.2639. URL https://doi.org/10.1038/nmeth.2639.

[28] Zilu Zhou, Bihui Xu, Andy Minn, and Nancy R. Zhang. DENDRO: genetic heterogeneity profiling and subclone detection by single-cell RNA sequencing. *Genome Biology*, 21(1), January 2020. doi: 10.1186/s13059-019-1922-x. URL https://doi.org/10.1186/s13059-019-1922-x.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[29] Davis J. McCarthy, , Raghd Rostom, Yuanhua Huang, Daniel J. Kunz, Petr Danecek, Marc Jan Bonder, Tzachi Hagai, Ruqian Lyu, Wenyi Wang, Daniel J. Gaffney, Benjamin D. Simons, Oliver Stegle, and Sarah A. Teichmann. Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes. *Nature Methods*, 17(4):414–421, March 2020. doi: 10.1038/s41592-020-0766-3. URL https://doi.org/10.1038/s41592-020-0766-3.

[30] Jennifer Tsoi, Lidia Robert, Kim Paraiso, Carlos Galvan, Katherine M. Sheu, Johnson Lay, Deborah J.L. Wong, Mohammad Atefi, Roksana Shirazi, Xiaoyan Wang, Daniel Braas, Catherine S. Grasso, Nicolaos Palaskas, Antoni Ribas, and Thomas G. Graeber. Multi-stage Differentiation Defines Melanoma Subtypes with Differential Vulnerability to Drug-Induced Iron-Dependent Oxidative Stress. *Cancer Cell*, 33(5):890–904.e5, May 2018. doi: 10.1016/j.ccell.2018.03.017. URL https://doi.org/10.1016/j.ccell.2018.03.017.

[31] Eva Pérez-Guijarro, Howard H. Yang, Romina E. Araya, Rajaa El Meskini, Helen T. Michael, Suman Kumar Vodnala, Kerrie L. Marie, Cari Smith, Sung Chin, Khiem C. Lam, Andres Thorkelsson, Anthony J. Iacovelli, Alan Kulaga, Anyen Fon, Aleksandra M. Michalowski, Willy Hugo, Roger S. Lo, Nicholas P. Restifo, Shyam K. Sharan, Terry Van Dyke, Romina S. Goldszmid, Zoe Weaver Ohler, Maxwell P. Lee, Chi-Ping Day, and Glenn Merlino. Multimodel preclinical platform predicts clinical response of melanoma to immunotherapy. *Nature Medicine*, 26 (5):781–791, April 2020. doi: 10.1038/s41591-020-0818-3. URL https://doi.org/10.1038/s41591-020-0818-3.

[32] Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, 9(1):171–181, January 2014. doi: 10.1038/nprot.2014.006. URL https://doi.org/10.1038/nprot.2014.006.

[33] Eli Eisenberg and Erez Y. Levanon. A-to-I RNA editing — immune protector and transcriptome diversifier. *Nature Reviews Genetics*, 19(8):473–490, April 2018. doi: 10.1038/s41576-018-0006-1. URL https://doi.org/10.1038/s41576-018-0006-1.

[34] David H. Mathews, Jeffrey Sabina, Michael Zuker, and Douglas H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288 (5):911–940, May 1999. doi: 10.1006/jmbi.1999.2700. URL https://doi.org/10.1006/jmbi.1999.2700.

[35] Robert M. Dirks and Niles A. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of Computational Chemistry*, 24(13):1664–1677, October 2003. doi: 10.1002/jcc.10296. URL https://doi.org/10.1002/jcc.10296.

[36] Hamidreza Chitsaz, Raheleh Salari, S. Cenk Sahinalp, and Rolf Backofen. A partition function algorithm for interacting nucleic acid strands. *Bioinformatics*, 25(12):i365–i373, May 2009. doi: 10.1093/bioinformatics/btp212. URL https://doi.org/10.1093/bioinformatics/btp212.

[37] Nikolai Karpov, Salem Malikic, Md. Khaledur Rahman, and S. Cenk Sahinalp. A multi-labeled tree dissimilarity measure for comparing "clonal trees" of tumor progression. *Algorithms for Molecular Biology*, 14(1), July 2019. doi: 10.1186/s13015-019-0152-9. URL https://doi.org/10.1186/s13015-019-0152-9.

[38] Sayaka Miura, Tracy Vu, Jiamin Deng, Tiffany Buturla, Olumide Oladeinde, Jiyeong Choi, and Sudhir Kumar. Power and pitfalls of computational methods for inferring clone phylogenies and mutation orders from bulk sequencing data. *Scientific Reports*, 10(1), February 2020. doi: 10.1038/s41598-020-59006-2. URL https://doi.org/10.1038/s41598-020-59006-2.

[39] Katharina Jahn, Niko Beerenwinkel, and Louxin Zhang. The Bourque Distances for Mutation Trees of Cancers. In *20th International Workshop on Algorithms in Bioinformatics (WABI 2020)*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020. doi: 10.4230/LIPICS.WABI.2020.14. URL https://drops.dagstuhl.de/opus/volltexte/2020/12803/.

[40] Zach DiNardo, Kiran Tomlinson, Anna Ritz, and Layla Oesper. Distance measures for tumor evolutionary trees. *Bioinformatics*, 36(7):2090–2097, November 2019. doi: 10.1093/bioinformatics/btz869. URL https://doi.org/10.1093/bioinformatics/btz869.

[41] William H. E. Day. Optimal algorithms for comparing trees with labeled leaves. *Journal of Classification*, 2(1): 7–28, December 1985. doi: 10.1007/bf01908061. URL https://doi.org/10.1007/bf01908061.

[42] Zhiqi Song, Chun-Di He, Changkai Sun, Yanni Xu, Xin Jin, Yi Zhang, Ting Xiao, Yakun Wang, Ping Lu, Yi Jiang, Huachen Wei, and Hong-Duo Chen. Increased expression of MAP2 inhibits melanoma cell proliferation, invasion and tumor growth in vitro and in vivo. *Experimental Dermatology*, 19(11):958–964, January 2010. doi: 10.1111/j.1600-0625.2009.01020.x. URL https://doi.org/10.1111/j.1600-0625.2009.01020.x.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[43] Emma Laks, Andrew McPherson, Hans Zahn, Daniel Lai, Adi Steif, Jazmine Brimhall, Justina Biele, Beixi Wang, Tehmina Masud, Jerome Ting, Diljot Grewal, Cydney Nielsen, Samantha Leung, Viktoria Bojilova, Maia Smith, Oleg Golovko, Steven Poon, Peter Eirew, Farhia Kabeer, Teresa Ruiz de Algara, So Ra Lee, M. Jafar Taghiyar, Curtis Huebner, Jessica Ngo, Tim Chan, Spencer Vatrt-Watts, Pascale Walters, Nafis Abrar, Sophia Chan, Matt Wiens, Lauren Martin, R. Wilder Scott, T. Michael Underhill, Elizabeth Chavez, Christian Steidl, Daniel Da Costa, Yussanne Ma, Robin J.N. Coope, Richard Corbett, Stephen Pleasance, Richard Moore, Andrew J. Mungall, Colin Mar, Fergus Cafferty, Karen Gelmon, Stephen Chia, The CRUK IMAXT Grand Challenge Team, Marco A. Marra, Carl Hansen, Sohrab P. Shah, and Samuel Aparicio. Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing. *Cell*, 179(5):1207–1221.e22, November 2019. doi: 10.1016/j.cell.2019.10.026. URL https://doi.org/10.1016/j.cell.2019.10.026.

[44] Jean-Christophe Marine, Sarah-Jane Dawson, and Mark A. Dawson. Non-genetic mechanisms of therapeutic resistance in cancer. *Nature Reviews Cancer*, 20(12):743–756, October 2020. doi: 10.1038/s41568-020-00302-4. URL https://doi.org/10.1038/s41568-020-00302-4.

[45] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zagar, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar B. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *bioRxiv*, October 2020. doi: 10.1101/2020.10.12.335331. URL https://doi.org/10.1101/2020.10.12.335331.

[46] Chen Xu and Zhengchang Su. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, 31(12):1974–1980, February 2015. doi: 10.1093/bioinformatics/btv088. URL https://doi.org/10.1093/bioinformatics/btv088.

[47] Jacob H. Levine, Erin F. Simonds, Sean C. Bendall, Kara L. Davis, El ad D. Amir, Michelle D. Tadmor, Oren Litvin, Harris G. Fienberg, Astraea Jager, Eli R. Zunder, Rachel Finck, Amanda L. Gedman, Ina Radtke, James R. Downing, Dana Pe'er, and Garry P. Nolan. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, 162(1):184–197, July 2015. doi: 10.1016/j.cell.2015.05.047. URL https://doi.org/10.1016/j.cell.2015.05.047.

[48] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, October 2008. doi: 10.1088/1742-5468/2008/10/p10008. URL https://doi.org/10.1088/1742-5468/2008/10/p10008.

[49] Sumaiyah K. Rehman, Jennifer Haynes, Evelyne Collignon, Kevin R. Brown, Yadong Wang, Allison M.L. Nixon, Jeffrey P. Bruce, Jeffrey A. Wintersinger, Arvind Singh Mer, Edwyn B.L. Lo, Cherry Leung, Evelyne Lima-Fernandes, Nicholas M. Pedley, Fraser Soares, Sophie McGibbon, Housheng Hansen He, Aaron Pollet, Trevor J. Pugh, Benjamin Haibe-Kains, Quaid Morris, Miguel Ramalho-Santos, Sidhartha Goyal, Jason Moffat, and Catherine A. O'Brien. Colorectal Cancer Cells Enter a Diapause-like DTP State to Survive Chemotherapy. *Cell*, 184(1): 226–242.e21, January 2021. doi: 10.1016/j.cell.2020.11.018. URL https://doi.org/10.1016/j.cell.2020.11.018.

[50] Antonio Ahn, Aniruddha Chatterjee, and Michael R. Eccles. The Slow Cycling Phenotype: A Growing Problem for Treatment Resistance in Melanoma. *Molecular Cancer Therapeutics*, 16(6):1002–1009, June 2017. doi: 10.1158/1535-7163.mct-16-0535. URL https://doi.org/10.1158/1535-7163.mct-16-0535.

[51] Qingfei Jiang, Leslie A. Crews, Frida Holm, and Catriona H. M. Jamieson. RNA editing-dependent epitranscriptome diversity in cancer stem cells. *Nature Reviews Cancer*, 17(6):381–392, April 2017. doi: 10.1038/nrc.2017.23. URL https://doi.org/10.1038/nrc.2017.23.

[52] Hamim Zafar, Yong Wang, Luay Nakhleh, Nicholas Navin, and Ken Chen. Monovar: single-nucleotide variant detection in single cells. *Nature Methods*, 13(6):505–507, April 2016. doi: 10.1038/nmeth.3835. URL https://doi.org/10.1038/nmeth.3835.

[53] Mohammadamin Edrisi, Hamim Zafar, and Luay Nakhleh. A Combinatorial Approach for Single-cell Variant Detection via Phylogenetic Inference. In Katharina T. Huber and Dan Gusfield, editors, *19th International Workshop on Algorithms in Bioinformatics (WABI 2019)*, volume 143 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 22:1–22:13, Dagstuhl, Germany, 2019. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-95977-123-8. doi: 10.4230/LIPIcs.WABI.2019.22. URL http://drops.dagstuhl.de/opus/volltexte/2019/11052.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[54] Chi-Ping Day, John Carter, Carrie Bonomi, Dominic Esposito, Bruce Crise, Betty Ortiz-Conde, Melinda Hollingshead, and Glenn Merlino. Lentivirus-mediated bifunctional cell labeling for in vivo melanoma study. *Pigment Cell & Melanoma Research*, 22(3):283–295, June 2009. doi: 10.1111/j.1755-148x.2009.00545.x. URL https://doi.org/10.1111/j.1755-148x.2009.00545.x.

[55] Olivier Poirion, Xun Zhu, Travers Ching, and Lana X. Garmire. Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage. *Nature Communications*, 9(1), November 2018. doi: 10.1038/s41467-018-07170-5. URL https://doi.org/10.1038/s41467-018-07170-5.

[56] Allegra A. Petti, Stephen R. Williams, Christopher A. Miller, Ian T. Fiddes, Sridhar N. Srivatsan, David Y. Chen, Catrina C. Fronick, Robert S. Fulton, Deanna M. Church, and Timothy J. Ley. A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. *Nature Communications*, 10(1), August 2019. doi: 10.1038/s41467-019-11591-1. URL https://doi.org/10.1038/s41467-019-11591-1.

[57] Anna S. Nam, Kyu-Tae Kim, Ronan Chaligne, Franco Izzo, Chelston Ang, Justin Taylor, Robert M. Myers, Ghaith Abu-Zeinah, Ryan Brand, Nathaniel D. Omans, Alicia Alonso, Caroline Sheridan, Marisa Mariani, Xiaoguang Dai, Eoghan Harrington, Alessandro Pastore, Juan R. Cubillos-Ruiz, Wayne Tam, Ronald Hoffman, Raul Rabadan, Joseph M. Scandura, Omar Abdel-Wahab, Peter Smibert, and Dan A. Landau. Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature*, 571(7765):355–360, July 2019. doi: 10.1038/s41586-019-1367-0. URL https://doi.org/10.1038/s41586-019-1367-0.

[58] Jun Ding, Chieh Lin, and Ziv Bar-Joseph. Cell lineage inference from SNP and scRNA-Seq data. *Nucleic Acids Research*, 47(10):e56–e56, March 2019. doi: 10.1093/nar/gkz146. URL https://doi.org/10.1093/nar/gkz146.

[59] Leif S. Ludwig, Caleb A. Lareau, Jacob C. Ulirsch, Elena Christian, Christoph Muus, Lauren H. Li, Karin Pelka, Will Ge, Yaara Oren, Alison Brack, Travis Law, Christopher Rodman, Jonathan H. Chen, Genevieve M. Boland, Nir Hacohen, Orit Rozenblatt-Rosen, Martin J. Aryee, Jason D. Buenrostro, Aviv Regev, and Vijay G. Sankaran. Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell*, 176(6):1325–1339.e22, March 2019. doi: 10.1016/j.cell.2019.01.022. URL https://doi.org/10.1016/j.cell.2019.01.022.

[60] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498, April 2011. doi: 10.1038/ng.806. URL https://doi.org/10.1038/ng.806.

[61] Simon Andrews, Felix Krueger, Anne Segonds-Pichon, Laura Biggins, Christel Krueger, and Steven Wingett. FastQC: a quality control tool for high throughput sequence data, 2010. URL http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

[62] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, October 2012. doi: 10.1093/bioinformatics/bts635. URL https://doi.org/10.1093/bioinformatics/bts635.

[63] Ryan Poplin, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, Laura D. Gauthier, Ami Levy-Moonshine, David Roazen, Khalid Shakir, Joel Thibault, Sheila Chandran, Chris Whelan, Monkol Lek, Stacey Gabriel, Mark J Daly, Ben Neale, Daniel G. MacArthur, and Eric Banks. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, November 2017. doi: 10.1101/201178. URL https://doi.org/10.1101/201178.

[64] K. Wang, M. Li, and H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164–e164, July 2010. doi: 10.1093/nar/gkq603. URL https://doi.org/10.1093/nar/gkq603.

[65] Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1), August 2011. doi: 10.1186/1471-2105-12-323. URL https://doi.org/10.1186/1471-2105-12-323.

[66] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), December 2014. doi: 10.1186/s13059-014-0550-8. URL https://doi.org/10.1186/s13059-014-0550-8.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.26.437185; this version posted March 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[67] Gennady Korotkevich, Vladimir Sukhov, Nikolay Budin, Boris Shpak, Maxim N. Artyomov, and Alexey Sergushichev. Fast gene set enrichment analysis. *bioRxiv*, June 2016. doi: 10.1101/060012. URL https://doi.org/10.1101/060012.

[68] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, September 2005. doi: 10.1073/pnas.0506580102. URL https://doi.org/10.1073/pnas.0506580102.