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

OXFORD

Subject Section

Phertilizer: growing a clonal tree from ultra-low coverage single-cell DNA sequencing of tumors

Leah L. Weber [1,*], Chuanyi Zhang [2,*], Idoia Ochoa [2,3,†] and Mohammed El-Kebir [1,4†]

[1]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA,
[2]Department of Electrical Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA and
[3]Department of Electrical and Electronics Engineering, Tecnun School of Engineering, University of Navarra, Donostia, Spain.

[*]These authors contributed equally.
[†]To whom correspondence should be addressed.

Abstract

Motivation: Emerging ultra-low coverage single-cell DNA sequencing (scDNA-seq) technologies have enabled high resolution evolutionary studies of copy number aberrations (CNAs) within tumors. While these sequencing technologies are well suited for identifying CNAs due to the uniformity of sequencing coverage, the sparsity of coverage poses challenges for the study of single-nucleotide variants (SNVs). In order to maximize the utility of increasingly available ultra-low coverage scDNA-seq data and obtain a comprehensive understanding of tumor evolution, it is important to simultaneously analyze the evolution of SNVs alongside CNAs from the same set of tumor cells.
Results: We present Phertilizer, a method to infer a clonal tree, capturing both SNV and CNA evolution, from ultra-low coverage single-cell DNA sequencing of a tumor. Based on a probabilistic model, our method recursively partitions the data by identifying key evolutionary events in the history of the tumor. We demonstrate the performance of Phertilizer on simulated data as well as on two real datasets, finding that Phertilizer effectively utilizes the copy-number signal inherent in the data to more accurately uncover clonal structure and genotypes compared to previous methods.
Availability: `https://github.com/elkebir-group/phertilizer`
Contact: {idoia,melkebir}@illinois.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Cancer results from an evolutionary process that yields a heterogeneous tumor composed of multiple subpopulations of cells, or *clones*, with distinct sets of somatic mutations (Nowell, 1976). These mutations include single-nucleotide variants (SNVs) that alter a single base and copy-number aberrations (CNAs) that amplify or delete large genomic regions. Over the last decade, new developments in single-cell DNA sequencing (scDNA-seq) methods have helped uncover a wealth of insights regarding intra-tumor heterogeneity and cancer evolution (Morita *et al.*, 2020; Baslan *et al.*, 2020; Kim *et al.*, 2018; Miles *et al.*, 2020). Importantly, in order to fully characterize the evolution of a tumor, both CNAs and SNVs should be studied and characterized by a single tumor phylogeny that depicts their coevolution. This is a challenging task due to limitations in current scDNA-seq technologies and corresponding tumor phylogeny algorithms.

One important tradeoff of all scDNA-seq technologies is the depth versus uniformity of sequencing coverage. Some technologies like DLP+ (Laks *et al.*, 2019) and ACT (Minussi *et al.*, 2021), which we refer to as ultra-low coverage scDNA-seq, provide extremely low coverage at each position (< 0.02x) but offer close to uniform sampling of the whole genome. As a result, these technologies have paved the way for enriching our understanding of the role CNAs play in cancer progression and tumor evolution (Minussi *et al.*, 2021; Laks *et al.*, 2019; Zahn *et al.*, 2017). The drawback of these ultra-low coverage sequencing techniques is the sparsity of coverage, which make it difficult to identify and characterize the evolution of SNVs. On the other hand, scDNA-seq technologies such as Mission Bio Tapestry (Pellegrino *et al.*, 2018) and MDA have higher coverage ($> 1\times$), due to preamplification of the DNA, thus making them well suited for identifying point mutations and accurate reconstruction of clonal genotypes (Roth *et al.*, 2016) and tumor phylogenies (Jahn *et al.*, 2016; Malikic *et al.*, 2019; El-Kebir, 2018; Zafar *et al.*, 2019) based only

© Weber, Zhang, Ochoa and El-Kebir, 2022.

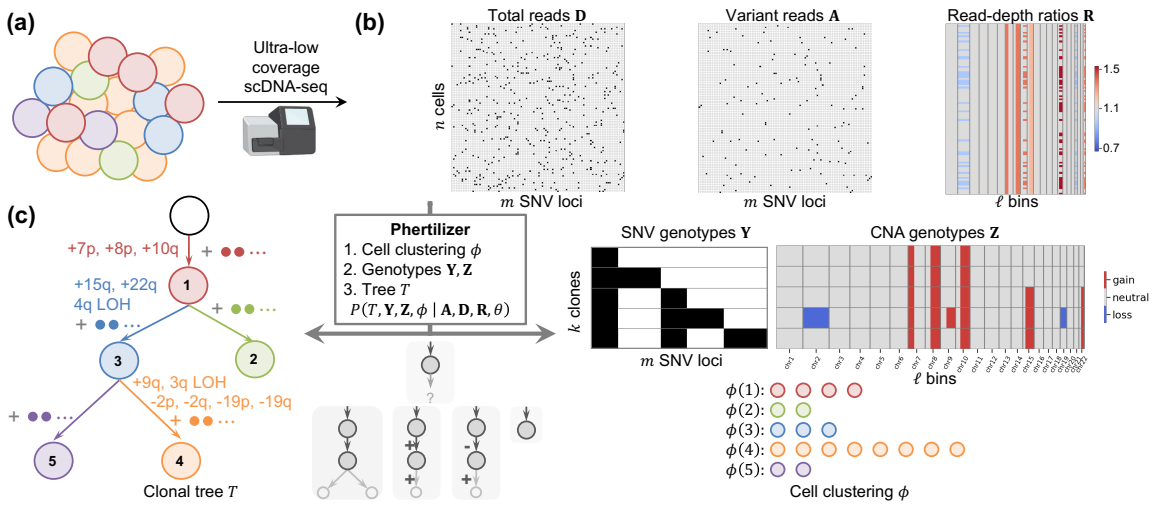

Fig. 1: **Phertilizer infers a clonal tree with SNV and CNA genotypes given ultra-low coverage single-cell sequencing data**. (a) A tumor is composed of groups of cells, or clones with distinct genotypes. (b) Ultra-low coverage scDNA-seq produces total read counts $\mathbf{D} \in \mathbb{N}^{n \times m}$ and variant read counts $\mathbf{A} \in \mathbb{N}^{n \times m}$ for n cells and m SNV loci, and read-depth ratios $\mathbf{R} \in \mathbb{R}^{n \times \ell}$ for the same cells across ℓ bins. (c) Phertilizer infers cell clustering ϕ, SNV genotypes $\mathbf{Y}$, CNA genotypes $\mathbf{Z}$, and a clonal tree T with maximum posterior probability $P(T, \mathbf{Y}, \mathbf{Z}, \phi \mid \mathbf{A}, \mathbf{D}, \mathbf{R})$.

on SNVs. However, these higher coverage scDNA-seq methods are not ideal for identifying CNAs as a result of amplification bias, which either provides a biased sampling of the genome or intentionally targets only loci that have a known impact in similar cancers.

Current methods for tumor phylogeny and/or clone inference naturally choose to focus on the features (SNV or CNA events) for which the data is ideally suited (Jahn *et al.*, 2016; Malikic *et al.*, 2019; El-Kebir, 2018; Zafar *et al.*, 2019; Roth *et al.*, 2016; Zaccaria and Raphael, 2021; Markowska *et al.*, 2021; Liu *et al.*, 2022; Wang *et al.*, 2021; Petkovic *et al.*, 2021). With the development of SBMClone, Myers *et al.* (2020) took the first step of using ultra-low coverage sequencing data to infer SNV clones. SBMClone performs co-clustering of cells and SNVs on ultra-low coverage single-cell DNA sequencing data using stochastic block modeling to uncover an underlying block structure of cells and SNVs. While SBMClone performed well on simulated data, especially with higher coverage ($> 0.2\times$), it was unable to identify clear structure in a 10x genomics breast cancer dataset without ad-hoc use of additional copy number clone information. Similarly, Rozhoňová *et al.* (2021) developed a cell clustering method that focuses only on a set of informative loci. While both of these clustering methods capitalize on the ever increasing throughput of ultra-low coverage scDNA-seq methods, neither method constrains the output by a tree and CNA features are used only in an ad-hoc manner or for orthogonal validation. As a result, both methods imply that CNA and SNV data features should be segregated and analyzed in separate bioinformatics pipelines.

The other emerging trend from the analysis of ultra-low coverage scDNA-seq is pseudobulk analysis (Laks *et al.*, 2019). In this approach, Laks *et al.* (2019) identified copy number clones using existing methods. Then, the cells forming a clone were pooled into a pseudobulk sample and used to identify SNVs. Finally, phylogeny inference under a Dollo (Dollo, 1893) evolutionary model was performed with the copy number clones as the leaves of the tree. Although this method yields an SNV tumor phylogeny, it does not capture the evolutionary history of the CNA events. Furthermore, by forcing the leaves to the copy number clones, this method does not allow for further refinement of these clones based on SNV evolution.

Here we introduce Phertilizer, the first method to grow a clonal tree from ultra-low single-cell DNA sequencing of tumors that captures both SNV and CNA evolution. The key insight of Phertilizer is that to overcome SNV coverage sparsity, we rely on both the high throughput of ultra-low coverage scDNA-seq and the strong copy number signal inherent in the data to guide the building of the clonal tree. By analogy to the planting and growing of trees, our method Phertilizer takes as input the variant and total read counts at a given set of loci and the total number of reads in binned regions of the genome and seeks to grow a clonal tree by recursively inferring a set of elementary clonal trees that are used as building blocks to construct the output clonal tree with maximum posterior probability. As opposed to current methods, our simulations demonstrate that Phertilizer accurately infers phylogenies and cell clusters when the number of cells matches current practice. On real data, we find that Phertilizer effectively utilizes the copy-number signal inherent in the data to uncover clonal structure, yielding high-fidelity SNV genotypes.

2 Problem statement

We seek to infer a phylogenetic tree describing the evolutionary history of a tumor given ultra-low coverage sequencing data from n cells from which we identified m SNV loci. Specifically, we are given variant reads $\mathbf{A} = [a_{iq}]$ and total reads $\mathbf{D} = [d_{iq}]$ where a_{iq} and d_{iq} are the variant and total read counts for SNV locus $q \in [m]$ in cell $i \in [n]$, respectively (Fig. 1b). While the number of cells is large with the latest generation of ultra-low coverage single-cell DNA sequencing technology ($n > 1000$ cells), the *coverage*, or the average number of reads that span a single locus is uniform but low ($0.01\times$) — that is, for each cell i we observe $a_{iq} = d_{iq} = 0$ reads for 99 out of every 100 SNV loci q. This sparsity renders phylogeny inference using SNVs extremely challenging. We propose to overcome this challenge using the following three key ideas.

First, similarly to current methods, we leverage the clonal structure present in tumors, i.e., the observed cells cluster into a small number of clones. Thus, we seek to group the n cells into k clones such that $k \ll n$ via a cell and SNV genotype clustering defined as follows.

Definition 1. Function $\phi : [k] \to 2^{[n]}$ is a *cell clustering* provided its image encodes a partition of cells $[n]$ into k (disjoint and non-empty) parts.

Definition 2. Matrix $\mathbf{Y} \in \{0,1\}^{k \times m}$ encodes *SNV genotypes* where $y_{jq} = 1$ ($y_{jq} = 0$) indicates that SNV q is present (absent) in clone j.

Second, because ultra-low coverage scDNA-seq is sequenced uniformly, we may leverage the copy number signal inherent in the data to infer CNA genotypes and improve the cell clustering performance. To this end we partition the reference genome into ℓ equal length regions, hereafter referred to as bins. We expect that all cells in a clone have the same CNA genotype defined as follows.

Definition 3. Matrix $\mathbf{Z} \in \{+,-,0\}^{k \times \ell}$ encodes *CNA genotypes* where $z_{jp} = +$ ($z_{jp} = -$) indicates that bin p in clone j underwent an amplification (deletion), otherwise bin p is neutral for $z_{jp} = 0$.

Third, similarly to methods such as SCITE (Jahn *et al.*, 2016), SciCloneFit (Zafar *et al.*, 2019) and SPhyR (El-Kebir, 2018), which operate on medium-to-high coverage scDNA-seq data, we consider that the observed cells are generated as the result of a tree-like evolutionary process that constrains the order of SNV clusters. In particular, we use the infinite sites model (Kimura, 1969) defined as follows.

Definition 4. A tree T with nodes $V(T) = \{v_1, \ldots, v_k\}$ rooted at node v_1 is a *clonal tree* with SNV genotypes $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_k]^\top$ and CNA genotype $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_k]^\top$ provided (i) each node v_j is labeled by SNV genotype $\mathbf{y}_j$ and CNA genotype $\mathbf{z}_j$, and (ii) each SNV q is gained exactly once and subsequently never lost, i.e., there is either exactly one directed edge $(v_{j'}, v_j)$ such that $y_{j'q} = 0$ and $y_{jq} = 1$ or $y_{1q} = 1$, and no edge $(v_{j'}, v_j)$ exists such that $y_{jq} = 1$ and $y_{j'q} = 0$.

In addition to observing variant and reference read counts $\mathbf{A}, \mathbf{D}$, we also observe read-depth ratios (RDRs) $\mathbf{R} = [r_{ip}]$, where r_{ip} is proportional to the copy number of bin $p \in [\ell]$ of cell $i \in [n]$ (Appendix 1.1.1). To relate the data to our latent variables of interest $(T, \mathbf{Y}, \mathbf{Z}, \phi)$, we introduce a generative model in Section 3.1 that takes sequencing errors into account as well as the effect of copy number on SNV variant read counts $\mathbf{A}$ and the RDR values $\mathbf{R}$. This enables us to define the posterior probability $P(T, \mathbf{Y}, \mathbf{Z}, \phi \mid \mathbf{A}, \mathbf{D}, \mathbf{R})$ that we use in the following problem.

Problem 1 (Clonal Tree Inference (CTI))**.** Given variant reads $\mathbf{A} \in \mathbb{N}^{n \times m}$, total reads $\mathbf{D} \in \mathbb{N}^{n \times m}$ and read-depth ratios $\mathbf{R} \in \mathbb{R}_{\geq 0}^{n \times \ell}$, find a clonal tree T with genotypes $(\mathbf{Y}, \mathbf{Z})$ and cell clustering ϕ with maximum posterior probability $P(T, \mathbf{Y}, \mathbf{Z}, \phi \mid \mathbf{A}, \mathbf{D}, \mathbf{R})$.

3 Materials and methods

3.1 Generative model

Our generative model captures the evolution and the subsequent process of ultra-low coverage sequencing of n cells with m SNVs and ℓ bins yielding our data $\mathbf{A}, \mathbf{D} \in \mathbb{N}^{n \times m}$ and $\mathbf{R} \in \mathbb{R}_{\geq 0}^{n \times \ell}$. Given a number $k \in \mathbb{N}$ of clones, we generate a rooted tree T with k nodes with a uniform prior $P(T \mid k)$. We generate SNV genotypes $\mathbf{Y}$ for tree T under the Dollo parsimony model with $P(\mathbf{Y} \mid T)$ uniformly distributed. We model the cell clustering $\phi : [k] \to 2^{[n]}$ with $P(\phi)$ distributed uniformly over all possible partitions of n cells into k clones/parts. The prior probability $P(\mathbf{D})$ of total read counts $\mathbf{D}$ is also uniform. The generative model has three remaining components: the (i) latent variant allele frequency model, (ii) the variant read count model and (iii) the read-depth ratio model, which we describe next (Fig. 2).

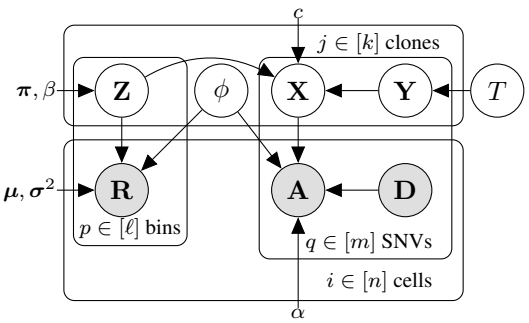

Fig. 2: **Plate diagram.** The observed data consists of variant and total read counts $\mathbf{A}, \mathbf{D}$ and RDR values $\mathbf{R}$, generated from an unobserved cell clustering ϕ and clonal tree T with SNV and CNA genotypes $\mathbf{Y}, \mathbf{Z}$ whose interplay is captured by latent variant allele frequencies (VAFs) $\mathbf{X}$. Given hyperparameters $\theta = (\boldsymbol{\pi}, \beta, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, c, \alpha)$ and observed data $(\mathbf{A}, \mathbf{D}, \mathbf{R})$, we seek $(T, \mathbf{Y}, \mathbf{Z}, \phi)$ with maximum posterior probability $P(T, \mathbf{Y}, \mathbf{Z}, \phi \mid \mathbf{A}, \mathbf{D}, \mathbf{R})$, marginalizing over latent VAFs $\mathbf{X}$.

3.1.1 Latent variant allele frequency (VAF) model

Let $\bar{d}_{jq}$ be the number of chromosomal copies of the SNV locus q in clone j, and let $\bar{a}_{jq} \leq \bar{d}_{jq}$ be the number of chromosomal copies that harbor the SNV. We call the ratio $\bar{a}_{jq}/\bar{d}_{jq}$ the *latent variant allele frequency* (VAF), which we model as the random variable x_{jq}. Latent VAFs $\mathbf{X} \in \mathbb{R}^{k \times m}$ depend on both SNV and CNA genotypes $\mathbf{Y}, \mathbf{Z}$. Given a maximum copy number $c \in \mathbb{N}$ ($c > 2$), we define the multiset $S(y, z)$ of latent VAFs associated with SNV state $y \in \{0, 1\}$ and CNA state $z \in \{+, -, 0\}$. Then, if SNV locus q occurs within bin p we have that $P(x_{iq} \mid y_{jq}, z_{jp})$ equals the proportion of occurrences of x_{iq} in $S(y_{jq}, z_{jp})$. For example, if we have maximum copy number $c = 3$, the SNV is present (i.e., $y = 1$) and the corresponding bin p is amplified (i.e., $s = +$), we have that $S(1, +) = \{1/3, 2/3, 3/3\}$ and that $P(x_{iq} = 2/3 \mid y_{jq} = 1, z_{jp} = +) = 1/3$. Appendix B.3.1 contains the precise definition of $P(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}, \theta)$.

3.1.2 Variant read count model

Given latent VAFs $\mathbf{X}$, total read counts $\mathbf{D}$ and cell clustering ϕ, we model alternate read counts $\mathbf{A}$ as a binomial distribution accounting for sequencing errors. While previous models for medium-to-high coverage scDNA-seq data specifically account for errors introduced during PCR amplification (Weber *et al.*, 2021; De Bourcy *et al.*, 2014), the lack of whole-genome amplification in the latest generation of ultra-low scDNA-seq technology restricts errors to base calling errors that occur during sequencing (Laks *et al.*, 2019; Minussi *et al.*, 2021). Let α be the probability of misreading a single nucleotide during sequencing, then we model the variant read count a_{iq} of SNV locus q in cell i as

$$a_{iq} \mid \phi, \mathbf{x}_{\cdot q}, d_{iq} \sim \mathrm{Binom}\left(d_{iq},\, x_{jq}(1-\alpha) + \frac{(1-x_{jq})\alpha}{3}\right),$$

where j is the unique clone corresponding to cell i, i.e., $i \in \phi(j)$. Appendix B.3.2 provides the derivation of $P(\mathbf{A} \mid \mathbf{D}, \mathbf{X}, \phi, \theta)$.

3.1.3 Read-depth ratio model

We model CNA genotypes $\mathbf{Z}$ as a Markov process parameterized by an initial state probability vector $\boldsymbol{\pi} = [\pi_+, \pi_-, \pi_0]^\top$ and a transition probability $\beta \in [0, 1]$. Specifically, for each clone j and initial bin p per chromosome we have $P(z_{jp} = s \mid \boldsymbol{\pi}) = \pi_s$. The remaining bins p depend on the copy-number state of the previous bin $p-1$ per the Markov property, i.e. we have $P(z_{jp} = s \mid z_{j,p-1} = t, \beta) = 1 - \beta$ if $s = t$ and $\beta/2$ otherwise. We model each RDR value r_{ip} as a Gaussian emission from

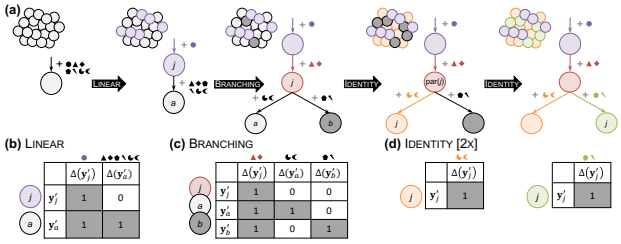

Fig. 3: **Phertilizer solves the CTI problem by enumerating clonal trees using three elementary operations in a recursive fashion.** (a) Each operation yields a new clonal tree T' by extending a leaf v_j of the previous clonal tree T and reassigning its SNVs $\Delta(\mathbf{y}_j)$ and cells $\phi(j)$. The resulting SNV genotypes $\mathbf{Y}'$ and cell clustering ϕ' are constrained as depicted: (b) `Linear`, (c) `Branching`, and (d) `Identity`.

hidden state $z_{jp} \in \{+, -, 0\}$, i.e. $r_{ip} \mid \phi, \mathbf{z}_{\cdot p}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 \sim \mathcal{N}(\mu_s, \sigma_s^2)$, where $s \in \{+, -, 0\}$ is the CNA state of bin p of the unique clone j corresponding to cell i (i.e. $i \in \phi(j)$ and $z_{jp} = s$). Appendix B.3.3 details the derivation of $P(\mathbf{Z} \mid \theta)$ and $P(\mathbf{R} \mid \phi, \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$.

3.1.4 Posterior probability

Given hyperparameters $\theta = (\boldsymbol{\pi}, \beta, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, c, \alpha)$ and observed data $(\mathbf{A}, \mathbf{D}, \mathbf{R})$, we seek to identify latent variables $(T, \mathbf{Y}, \mathbf{Z}, \phi)$ with maximum posterior probability $P(T, \mathbf{Y}, \mathbf{Z}, \phi \mid \mathbf{A}, \mathbf{D}, \mathbf{R}, \theta)$. Using Bayes' theorem and further simplification, we show in Appendix B.3.4 that this is equivalent to $\arg\max_{T,\mathbf{Y},\mathbf{Z},\phi}\{\log P(\mathbf{Z} \mid \theta) + \log P(\mathbf{R} \mid \mathbf{Z}, \phi, \theta) + \log \sum_{\mathbf{X}} P(\mathbf{A} \mid \mathbf{D}, \mathbf{X}, \phi, \theta) P(\mathbf{X} \mid \mathbf{Y}, \mathbf{Z}, \theta)\}$.

3.2 Phertilizer

At the root of Phertilizer (pun intended) are three elementary tree operations that break down the CTI problem into smaller subproblems that we seek to solve to optimality. Specifically, each operation takes as input a clonal tree T with genotypes $\mathbf{Y}, \mathbf{Z}$ and cell clustering ϕ and yields a new clonal tree T' with updated genotypes $\mathbf{Y}', \mathbf{Z}'$ and cell clustering ϕ' by transforming exactly one leaf v_j of T (Fig. 3). We begin by introducing some notation. Given node v_j not equal to the root v_1, the function $\mathrm{par}(j)$ returns its parent node $v_{\mathrm{par}(j)}$ in clonal tree T. Recall that $\mathbf{y}_j \in \{0, 1\}^m$ is the SNV genotype of clone j corresponding to node v_j in T. With a slight overloading of notation, $\mathbf{y}_j$ also indicates the set of SNVs present in clone j. We denote the set of SNVs introduced on the incoming edge $(v_{\mathrm{par}(j)}, v_j)$ of node v_j by $\Delta(\mathbf{y}_j) = \mathbf{y}_j \setminus \mathbf{y}_{\mathrm{par}(j)}$.

A `Linear` operation transforms node v_j of clonal tree T with cells $\phi(j)$ and mutations $\Delta(\mathbf{y}_j)$ into a clonal tree T' with a new node v_a and directed edge (v_j, v_a), partitioning original cells $\phi(j)$ and mutations $\Delta(\mathbf{y}_j)$ among nodes v_j and v_a such that the resulting SNV genotypes $\mathbf{y}'_j, \mathbf{y}'_a$ adhere to the infinite sites model, i.e. $\mathbf{y}'_j \subseteq \mathbf{y}'_a$ and $\mathbf{y}'_j \cup \mathbf{y}'_a = \Delta(\mathbf{y}_j)$ (Fig. 3b). Similarly, a `Branching` operation results in a clonal tree T' with two new nodes v_a, v_b that are children of v_j such that the original cells $\phi(j)$ are partitioned between child nodes v_a, v_b, the mutations $\Delta(\mathbf{y}_j)$ are partitioned between all three nodes v_j, v_a, v_b, and the genotypes $\mathbf{y}'_a, \mathbf{y}'_b, \mathbf{y}'_c$ also adhere to the infinite sites model, i.e. $\mathbf{y}'_j \subseteq \mathbf{y}'_a, \mathbf{y}'_j \subseteq \mathbf{y}'_b, \Delta(\mathbf{y}'_a) \cap \Delta(\mathbf{y}'_b) = \emptyset$ and $\mathbf{y}'_j \cup \mathbf{y}'_a \cup \mathbf{y}'_b = \Delta(\mathbf{y}_j)$. We note that this operation may assign no cells to v_j, thus modeling an extinct clone with SNVs that are common to its children but absent from its parent (Fig. 3c). Lastly, an `Identity` operation on v_j does not alter the tree and prevents the application of other elementary operations on that node (Fig. 3d).

While each elementary tree algorithm varies out of necessity for the specific constrained CTI problem being solved, the three key ideas overlap. First, we partition cells ϕ_j and SNVs $\mathbf{y}_j$ in an iterative fashion. That is, we fix one partition (*e.g.* SNVs) and solve for the other partition (*e.g.* cells), and alternate until convergence or until we reach a specified number of iterations. Drawing parallels between image segmentation, where utilizing pixel location features results in better clustering (Shi and Malik, 2000), we incorporate both RDR values $\mathbf{R}$ and variant read counts $\mathbf{A}$ into a single feature as opposed to only using sparse SNV features when finding a cell partition. We implement this via the well known normalized cut algorithm (Shi and Malik, 2000). Since our method is iterative, improved cell partitions enables us to subsequently obtain better quality SNV partitions. Given a fixed cell partition, we use our probabilistic model to assign each SNV being partitioned to the edge with maximum likelihood. Second, we use the Viterbi algorithm to perform CNA genotyping on non-empty nodes (Viterbi, 1967). Third, to prevent overfitting, we introduce additional regularization via parameters λ, τ, restricting the only valid operation to `Identity` for leaf nodes with fewer than λ attached cells and/or fewer than τ introduced SNVs. Details are provided in Appendix A.3.

Phertilizer is implemented in Python 3, is open source (BSD-3-Clause license), and is available at `https://github.com/elkebir-group/phertilizer`. Finally, we note that as normal cells might not be readily available to compute input RDR values, Phertilizer offers a mode where it instead uses binned read counts (Appendix A.3). In this mode, Phertilizer infers a clonal tree T, SNV genotypes $\mathbf{Y}$ and cell clustering ϕ but does not infer CNA genotypes $\mathbf{Z}$.

4 Results

4.1 Simulation results

Our simulation study evaluates three things. (i) How well is each method able to identify clusters of cells with similar SNV and CNA genotypes? (ii) How accurate are the inferred clonal trees? (iii) How accurate are the SNV/CNA genotypes with respect to a known ground truth? Additionally, we assess robustness to violations of the infinite sites assumption. To achieve this, we generated simulation instances with varying number $k \in \{5, 7\}$ of nodes, number $n \in \{1500, 2500, 5000\}$ of sequenced cells and the number $m \in \{1500, 2500, 5000\}$ of SNVs with a mean sequencing coverage of $0.01\times$. In addition, we varied the SNV evolutionary model, using either a Dollo (SNV loss, Dollo (1893)) or an infinite sites model (without SNV loss, Kimura (1969)).

We benchmarked Phertilizer against SCG (Roth *et al.*, 2016) and SBMClone (Myers *et al.*, 2020) either run in isolation or in conjuction with SPhyR (El-Kebir, 2018) to additionally perform tree inference (denoted as SBMClone+SPhyR). We note that the cell clustering and SNV genotypes inferred by SBMClone are not altered by SPhyR, rather it seeks to reconcile these into a phylogeny under the Dollo evolutionary model. We focus our discussion on simulations with $k = 7$ clones and $m = 2500$ SNVs, and show results for SCG and SBMClone run in hierarchical mode (SBMClone-H). Results for other regimes and for SBMClone run in default mode as well as further details on simulation setup are provided in the supplement (Appendix B.1).

Cell clustering. To assess cell clustering, we considered two performance metrics: (i) cell ARI (adjusted Rand index, cf. Rand (1971)) and (ii) the number of inferred cell clusters. As typical scDNA-seq data currently contain around $800 - 2000$ sequenced cells, good performance on the low end ($n = 1500$ cells) is critical. Phertilizer achieved the highest overall median cell ARI of 0.70 (Fig. 4a), with only a small deterioration in performance between $n = 5000$ (median: 0.72) and $n = 1500$

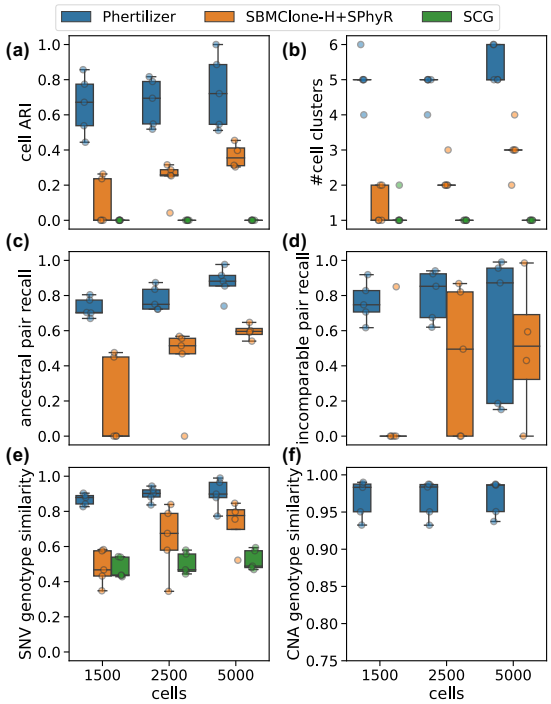

Fig. 4: **Phertilizer outperforms existing methods on simulated data.** We show results for $k = 7$ nodes and $m = 2500$ SNVs, comparing Phertilizer to SBMClone run in hierarchical model in conjuction with SPhyR (SBMClone-H+SPhyR) and SCG. (a-b) Cell clustering performance. (c-d) Tree inference performance, considering placement of SNV gains on the tree. No results for SCG are shown, as it lacks the capability to infer a tree. (e) SNV genotyping performance. (f) CNA genotyping performance. No results for SBMClone-H+SPhyR and SCG are shown, as they lack the capability of inferring CNA genotypes.

cells (median: 0.67). In comparison, SBMClone achieved significantly lower performance, with overall median cell ARI of 0.26. Moreover, SBMClone's performance on the smaller input size ($n = 1500$) was significantly worse (median: 0.0) than the largest input size ($n = 5000$, median: 0.36). As anticipated, SCG performed extremely poorly with an overall median cell ARI of only 0.0 since it was designed for clustering scDNA-seq data with much higher coverage. The difference in cell ARI performance between methods is due to Phertilizer's ability to recover a larger number of cell clusters (median: 5) when compared to SBMClone (median: 2) and SCG (median: 1) (Fig. 4b). Finally, we observed similar performance trends for $k = 5$ clones and $m \in \{1500, 5000\}$ SNVs, maintaining good performance even when the ground truth included SNV losses (Fig. S1-S12).

Clonal tree inference. We next evaluated the ability of Phertilizer to accurately infer a known ground truth clonal tree T^*. We used the ancestral pair recall (APR) and incomparable pair recall (IPR) metrics (El-Kebir *et al.*, 2015), which assess whether SNV gains are correctly arranged in the inferred tree T. That is, APR (IPR) is the fraction of ordered (unordered) pairs of SNV gains on the same (distinct) branch of T^* that are inferred on a single (distinct) branch in T. When T^* is linear, we set IPR to 1.

For APR, Phertilizer outperformed (median: 0.77) SBMClone-H+SPhyR (median: 0.50) (Fig. 4c). Similar to cell ARI, we also observed smaller variability in median APR as the number of input cells decreases than for SBMClone-H+SPhyR. The same performance trend was observed

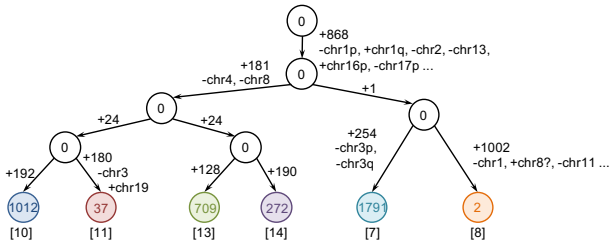

Fig. 5: **Phertilizer infers a branching clonal tree in correspondence with CHISEL's copy number tree (Zaccaria and Raphael, 2021).** Phertilizer infers a clonal tree with $k = 11$ nodes and 6 cell clones. The incoming edges are labeled with both the number of gained SNVs as well as CNA state change. Additionally, numbers below the leaf nodes indicate the number of assigned cells.

for IPR where Phertilizer (median: 0.83) significantly outperformed SBMClone+SPhyR (median: 0.22) (Fig. 4d). However, even though the median remained high on $n = 5000$ cells, we noted an increase in standard deviation from $n = 1500$ cells (sd: 0.115) to $n = 5000$ cells (sd: 0.425) for Phertilizer. These trends were consistent across varying simulation regimes, such as $k = 5$ and on both a small and larger number of SNVs (Fig. S1-S3, S7-S9). Finally, we observed robustness to the presence of SNV loss in tree inference and outperformed SBMClone-H+SPhyR for this regime, especially for APR (Fig. S4-S6, S10-S12).

Genotype inference. Lastly, we evaluated the ability of Phertilizer to accurately infer SNV and CNA genotypes when compared to a known ground truth. We define a genotype similarity score as $1 - 1/\sum_{i=1}^{n} D(\hat{\mathbf{g}}_i, \mathbf{g}_i^*)$, where $D(\hat{\mathbf{g}}_i, \mathbf{g}_i^*)$ is the Hamming distance between the inferred genotype $\hat{\mathbf{g}}_i$ and the ground truth genotype $\mathbf{g}_i^*$ for cell i. Thus, a score of 1 implies that every sequenced cell was correctly genotyped.

On SNV genotype similarity, Phertilizer (median: 0.89) significantly outperformed both SBMClone-H+SPhyR (median: 0.58) and SCG (median: 0.48) (Fig. 4e), also in the presence of SNV loss (Fig. S4-S6, S10-S12). As seen in Fig. 4f where the y-axis is between $0.75-1.00$, Phertilizer had excellent performance on CNA genotype similarity with an overall median of 0.98. We also found that Phertilizer performed extremely well on genotype metrics across all regimes (Fig. S1-S12). Phertilizer's high performance on both SNV and CNA genotyping demonstrates the benefit of incorporating both an evolutionary model and copy-number information into the genotype inference problem in the presence of extremely sparse sequencing coverage (Fig. 4e-f).

4.2 Real data results

4.2.1 Breast cancer data sequenced by 10X Genomics

We applied Phertilizer to ultra-low coverage scDNA-seq data from a breast cancer patient sequenced via 10X Genomics technology. The data consisted of $n = 4055$ tumor cells and an additional 4226 normal cells. In previous work, Zaccaria and Raphael (2021) performed haplotype specific copy-number calling via CHISEL. We used their identified cell clones for comparison. We also used these clusters to compute a guideline SNV genotype profile, designating an SNV as present for a clone if at least two cells assigned to that clone had a variant read. To assess the accuracy of our inferred CNA genotypes, we discretized the copy number profile's of CHISEL clones into states $s \in \{+, -, 0\}$, assigning a neutral state of 0 when the copy number was 4 due to a truncal whole genome duplication and the gain and loss states accordingly. Additionally, we benchmarked our method against SBMClone-H using the same baselines for comparison.

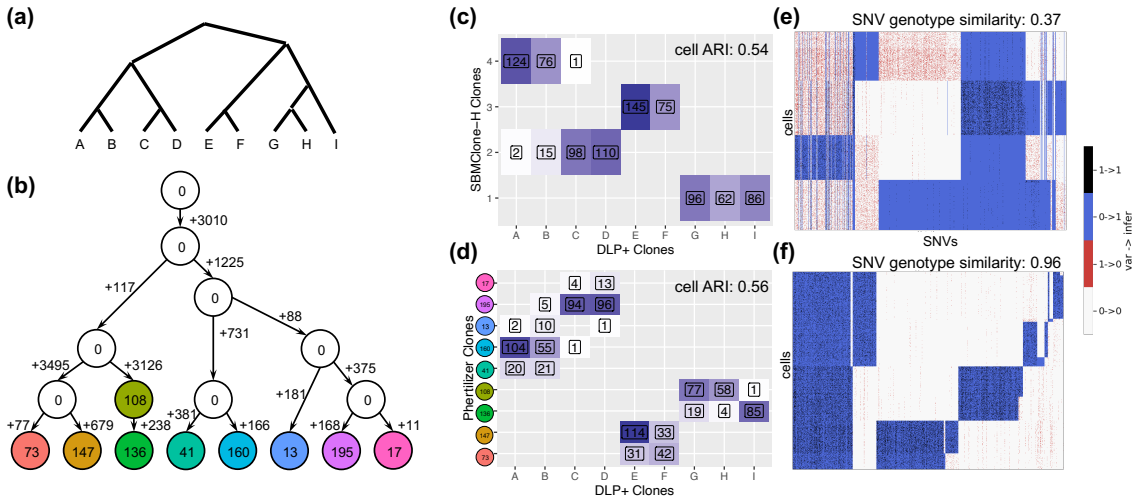

Fig. 6: **Phertilizer outperforms SBMClone on high-grade serous ovarian cancer DLP+ data from (Laks *et al.*, 2019).** (a) Laks *et al.* (2019) inferred 9 clones through copy-number profiling. (b) Phertilizer inferred a clonal tree with 9 clones as well. Edges labels indicate the number of introduced SNVs, whereas nodes labels indicate the number of corresponding cells. (c) SBMClone inferred only 4 clones, resulting in a cell ARI of 0.54. (d) Phertilizer achieved a comparable cell ARI of 0.56. (e) The SNV genotype similarity was 0.37 for SBMClone, (f) whereas Phertilizer achieved a SNV genotype similarity of 0.96.

To run Phertilizer, we conducted a grid search over hyperparameters and selected the tree with overall maximum posterior probability. We inferred a clonal tree with $k = 11$ nodes and 6 clones, obtaining a cell ARI of 0.56 and a SNV genotype similarity of 0.74 (Fig. S13) when compared to the CHISEL clones (Zaccaria and Raphael, 2021). In contrast, SBMClone achieved a cell ARI of 0.0, clustering almost all the cells into a single clone (Fig. S13) and an SNV genotype similarity of 0.54. The copy number tree inferred by Zaccaria and Raphael (2021) on this data is branching, such that the tree is dominated by two large clones, one in each branch. Phertilizer was able to accurately identify these two major clades (Fig. 5) and accurately characterizing their genotypes (Fig. S13). However, we observed that Phertilizer assigned 1002 SNVs to node a with only two cells. Upon inspection of the data, we noted low support in terms of variant reads for these SNVs, which explains why Phertilizer assigns them to a spurious cluster. More accurate variant calling on ultra-low coverage scDNA-seq data, along with more stringent filtering requirements would have prevented such spurious clustering. In addition, we achieved a CNA genotype score of 0.85 when comparing against the CHISEL's discretized copy number profiles. Thus, we demonstrate Phertilizer's potential for maximizing the information signal from a single sequencing dataset.

4.2.2 High grade serous ovarian cancer cell lines

We ran Phertilizer on $n = 890$ sequenced cells from three high grade serous ovarian cancer clonally related cancer cells lines sequenced with DLP+ (Laks *et al.*, 2019). We obtained the variant and total read counts $\mathbf{A}, \mathbf{D}$ along with binned read counts $\mathbf{B}$ for $\ell = 6{,}207$ bins (bin width of 500 KB) from the paper. Laks *et al.* (2019) identified 9 copy number clones through dimensionality reduction and clustering as well as $m = 14{,}068$ SNVs, which were reconciled in a single phylogeny with the copy number clones as leaves (reproduced in Fig. 6a).

Since no normal cells were available to construct RDR values, we ran Phertilizer in read-depth mode, obtaining a clonal tree T, SNV genotypes $\mathbf{Y}$ and cell clustering ϕ without running any copy-number calling methods *a priori* on this data. We benchmarked our performance against the DLP+ cell clusters that were obtained via more rigorous copy number calling methods using cell ARI as performance metric. To obtain a benchmark for computing the SNV genotype similarity score, we grouped the cells into the reported DLP+ clones, designating an SNV as present for a clone if at least two cells assigned to that clone had a supporting variant read. We performed a grid search over hyperparameters with 3 input seeds, using 5 restarts for each tree operation with a maximum of 10 iterations. We selected the output tree with the maximum posterior probability. Additionally, we ran SBMClone on the same input data.

Phertilizer returned a clonal tree with $k = 16$ nodes, composed of 9 clones, with branching structure similar to the DLP+ inferred tree (Fig. 6b). Both methods had good concordance with the DLP+ clones with a cell ARI of 0.54 for SBMClone-H and 0.56 for Phertilizer (Fig. 6c,d). Both Phertilizer and SBMClone-H were unable to distinguish DLP+ clones A and B or C and D as these clones appear to differ mostly in terms of CNAs based on Laks *et al.* (2019)'s inferred phylogeny. However, unlike SBMClone-H, Phertilizer was able to mostly identify clone I. Additionally, due to the large number of SNVs ($m = 14{,}068$), SBMClone-H's block clustering approach was able achieve acceptable cell clusterings without considering copy-number information. But, due to the small number of cells ($n = 890$), the SNV genotype similarity score of the clustered cells for SBMClone-H is poor (0.37, see Fig. 6e). In contrast, Phertilizer obtained a high SNV genotype score of 0.96 when compared with our benchmark SNV genotypes (Fig. 6f).

5 Discussion

The ability to infer clones using copy-number profiles from ultra-low coverage scDNA-seq data is well established (Zaccaria and Raphael, 2021). Here, we showed that the copy-number signal inherent in these data, along with evolutionary constraints, can be used in a novel way to overcome the sparse SNV signal and uncover clonal structure, resulting in high-fidelity SNV and CNA genotypes. We introduced Phertilizer, which grows clonal trees in a recursive fashion using two new elementary tree operations. We demonstrated the effectiveness of our approach relative to current methods on both simulated and real data. Importantly, we find that for the current number of cells ($800 - 2000$) used in practice, Phertilizer

performs markedly better than current methods, yielding more accurate clonal trees, cell clusters and SNV genotypes.

There are several limitations and directions for future research. First, the infinite sites model used in this work is often violated due to copy-number deletions. While we demonstrate robustness to such violations, a future direction is to use the Dollo evolutionary model (Dollo, 1893; El-Kebir, 2018). Second, our model lacks an explicit evolutionary model for CNAs. We plan to incorporate models such as those based on segmental amplifcations/deletions (El-Kebir *et al.*, 2017). Third, there is an opportunity to polish the inferred trees and improve performance, by reassigning cells and SNVs that have low support. Fourth, beyond SNVs and CNAs, we plan to support structural variants and integrate other omics modalities such as methylation and transcription.

Funding. M.E-K. was supported by the National Science Foundation under award numbers CCF 1850502 and CCF 2046488. I.O. was supported by a Gipuzkoa Fellows grant from the Basque Government and a Ramon y Cajal Grant (Spain).

References

Baslan, T., Kendall, J., Volyanskyy, K., McNamara, K., Cox, H., D'Italia, S., Ambrosio, F., Riggs, M., Rodgers, L., Leotta, A., *et al.* (2020). Novel insights into breast cancer copy number genetic heterogeneity revealed by single-cell genome sequencing. *Elife*, **9**, e51480.

De Bourcy, C. F., De Vlaminck, I., Kanbar, J. N., Wang, J., Gawad, C., and Quake, S. R. (2014). A quantitative comparison of single-cell whole genome amplification methods. *PloS one*, **9**(8), e105585.

Dollo, L. (1893). Les lois de l'évolution. *Bulletin de la Société belge de géologie, de paléontologie et d'hydrologie*, **7**, 164–166.

El-Kebir, M. (2018). Sphyr: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics*, **34**(17), i671–i679.

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

El-Kebir, M., Raphael, B. J., Shamir, R., Sharan, R., Zaccaria, S., Zehavi, M., and Zeira, R. (2017). Complexity and algorithms for copy-number evolution problems. *Algorithms for Molecular Biology*, **12**(1), 1–11.

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

Kim, C., Gao, R., Sei, E., Brandt, R., Hartman, J., Hatschek, T., Crosetto, N., Foukakis, T., and Navin, N. E. (2018). Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. *Cell*, **173**(4), 879–893.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, **61**(4), 893.

Laks, E., McPherson, A., Zahn, H., Lai, D., Steif, A., Brimhall, J., Biele, J., Wang, B., Masud, T., Ting, J., *et al.* (2019). Clonal decomposition and dna replication states defined by scaled single-cell genome sequencing. *Cell*, **179**(5), 1207–1221.

Liu, Y., Edrisi, M., Ogilvie, H., and Nakhleh, L. (2022). Nestedbd: Bayesian inference of phylogenetic trees from single-cell dna copy number profile data under a birth-death model. *bioRxiv*.

Malikic, S., Mehrabadi, F. R., Ciccolella, S., Rahman, M. K., Ricketts, C., Haghshenas, E., Seidman, D., Hach, F., Hajirasouliha, I., and Sahinalp, S. C. (2019). 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.

Markowska, M., Cąkała, T., Miasojedow, B., Juraeva, D., Mazur, J., Ross, E., Staub, E., and Szczurek, E. (2021). Conet: Copy number event tree model of evolutionary tumor history for single-cell data. *bioRxiv*.

Miles, L. A., Bowman, R. L., Merlinsky, T. R., Csete, I. S., Ooi, A. T., Durruthy-Durruthy, R., Bowman, M., Famulare, C., Patel, M. A., Mendez, P., *et al.* (2020). Single-cell mutation analysis of clonal evolution in myeloid malignancies. *Nature*, **587**(7834), 477–482.

Minussi, D. C., Nicholson, M. D., Ye, H., Davis, A., Wang, K., Baker, T., Tarabichi, M., Sei, E., Du, H., Rabbani, M., *et al.* (2021). Breast tumours maintain a reservoir of subclonal diversity during expansion. *Nature*, **592**(7853), 302–308.

Morita, K., Wang, F., Jahn, K., Hu, T., Tanaka, T., Sasaki, Y., Kuipers, J., Loghavi, S., Wang, S. A., Yan, Y., *et al.* (2020). Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nature Communications*, **11**(1), 1–17.

Myers, M. A., Zaccaria, S., and Raphael, B. J. (2020). Identifying tumor clones in sparse single-cell mutation data. *Bioinformatics*, **36**(Supplement_1), i186–i193.

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

Pellegrino, M., Sciambi, A., Treusch, S., Durruthy-Durruthy, R., Gokhale, K., Jacob, J., Chen, T. X., Geis, J. A., Oldham, W., Matthews, J., *et al.* (2018). High-throughput single-cell DNA sequencing of acute myeloid leukemia tumors with droplet microfluidics. *Genome Research*, **28**(9), 1345–1352.

Petkovic, M., Watkins, T. B., Colliver, E. C., Laskina, S., Swanton, C., Haase, K., and Schwarz, R. F. (2021). Whole-genome doubling-aware copy number phylogenies for cancer evolution with MEDICC2. *bioRxiv*.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**(336), 846–850.

Roth, A., McPherson, A., Laks, E., Biele, J., Yap, D., Wan, A., Smith, M. A., Nielsen, C. B., McAlpine, J. N., Aparicio, S., *et al.* (2016). Clonal genotype and population structure inference from single-cell tumor sequencing. *Nature Methods*, **13**(7), 573–576.

Rozhoňová, H., Danciu, D., Stark, S., Rätsch, G., Kahles, A., and Lehmann, K.-V. (2021). Secedo: Snv-based subclone detection using ultra-low coverage single-cell dna sequencing. *bioRxiv*.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, **22**(8), 888–905.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, **13**(2), 260–269.

Wang, F., Wang, Q., Mohanty, V., Liang, S., Dou, J., Han, J., Minussi, D. C., Gao, R., Ding, L., Navin, N., *et al.* (2021). Medalt: single-cell copy number lineage tracing enabling gene discovery. *Genome Biology*, **22**(1), 1–22.

Weber, L. L., Sashittal, P., and El-Kebir, M. (2021). doubletD: detecting doublets in single-cell DNA sequencing data. *Bioinformatics*, **37**(Supplement_1), i214–i221.

Zaccaria, S. and Raphael, B. J. (2021). Characterizing allele-and haplotype-specific copy numbers in single cells with chisel. *Nature Biotechnology*, **39**(2), 207–214.

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

Zahn, H., Steif, A., Laks, E., Eirew, P., VanInsberghe, M., Shah, S. P., Aparicio, S., and Hansen, C. L. (2017). Scalable whole-genome single-cell library preparation without preamplification. *Nature methods*, **14**(2), 167–173.