

## REVIEW

# Algorithmic approaches to clonal reconstruction in heterogeneous cell populations

Wazim Mohammed Ismail\*, Etienne Nzabarushimana, Haixu Tang

School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN 47405-7000, USA

\* Correspondence: wazimoha@iu.edu

Received June 2, 2019; Revised August 9, 2019; Accepted August 25, 2019

**Background:** The reconstruction of clonal haplotypes and their evolutionary history in evolving populations is a common problem in both microbial evolutionary biology and cancer biology. The clonal theory of evolution provides a theoretical framework for modeling the evolution of clones.

**Results:** In this paper, we review the theoretical framework and assumptions over which the clonal reconstruction problem is formulated. We formally define the problem and then discuss the complexity and solution space of the problem. Various methods have been proposed to find the phylogeny that best explains the observed data. We categorize these methods based on the type of input data that they use (space-resolved or time-resolved), and also based on their computational formulation as either combinatorial or probabilistic. It is crucial to understand the different types of input data because each provides essential but distinct information for drastically reducing the solution space of the clonal reconstruction problem. Complementary information provided by single cell sequencing or from whole genome sequencing of randomly isolated clones can also improve the accuracy of clonal reconstruction. We briefly review the existing algorithms and their relationships. Finally we summarize the tools that are developed for either directly solving the clonal reconstruction problem or a related computational problem.

**Conclusions:** In this review, we discuss the various formulations of the problem of inferring the clonal evolutionary history from allele frequency data, review existing algorithms and categorize them according to their problem formulation and solution approaches. We note that most of the available clonal inference algorithms were developed for elucidating tumor evolution whereas clonal reconstruction for unicellular genomes are less addressed. We conclude the review by discussing more open problems such as the lack of benchmark datasets and comparison of performance between available tools.

**Keywords:** clonal theory; infinite sites assumption; clonal reconstruction problem; bacteria evolution; tumor evolution; combinatorial algorithm; probabilistic algorithm

**Author summary:** As cells divide, they often gain new mutations creating newborn cells that are genetically distinct from their parent cells. Each new genetically distinct cell is called a clone. The problem of inferring the number of clones in a given population of cells, the unique set of mutations that identify each clone and the ancestral history of these identified clones is known as the clonal reconstruction problem. In this review, we discuss the theoretical framework of this problem, briefly review and classify the existing algorithms based on their approach and discuss open problems in this area of research.

## INTRODUCTION

Many unicellular organisms like archaea and bacteria reproduce by asexual cell division. Evolution in such

organisms is driven by the accumulation of mutations in their genomes occurring during DNA replication [1,2]. The variations induced by the mutations fall into several categories, including single nucleotide variations (SNVs),

short insertions and deletions (indels), copy number variations (CNVs) and large structural variations (SVs) — some of which lead to novel cellular functions adaptive to specific environmental conditions [3,4]. Interestingly, rapidly dividing somatic cells such as cancer cells in multi-cellular organisms are often hypothesized to follow a similar evolutionary process as the unicellular organisms. In particular, this hypothesis (known as the clonal theory) assumes that cells in the evolving population or tissue (*i.e.*, the tumor tissue) are the descendants of one or a few founder clones, where a clone is referred to as a subpopulation of cells sharing identical genome (and thus sharing the identical set of variations among the entire population).

According to the clonal theory, during the course of evolution, cells accumulate novel variations forming new clones. An evolving population is considered to be clonal if the ancestral relationships between clones are always vertical, *i.e.*, gene transfer from parent to offspring, while the horizontal transfer (or transfer of genetic materials between different cells or organisms outside of reproduction process) and the recombination across cells are negligible [1]. As a result, the evolutionary history of the clones can be represented by a directed tree, referred to as the clonal tree, in which each vertex representing a clone has one and only one incoming edge representing the ancestor of the clone, except for the root node that represents the founder clone.

The characterization of the evolutionary history and dynamics in a clonal population is critical for understanding the mechanism of adaptation and evolution, and for detecting genetic elements under selection [5]. For instance, in cancer biology, characterizing the heterogeneity of cancer cells and reconstructing the ancestral relationships between clonal cancer genomes is a key step to identify driver mutations [6], *i.e.*, the mutations occurring in early tumorigenesis and driving tumor progression, and to devise effective therapeutic approaches. In long term evolution experiments (LTEEs) of unicellular microbes and microbial communities, characterization of clonal structures is helpful for elucidating subpopulations under selection in specific environmental conditions (*e.g.*, antibiotic treatments) [5,7–11].

A straightforward approach to characterize clones and their ancestral relationships is to sequence a large number of individual genomes sampled at random from the population. However, even with reduced cost, this approach is quite expensive, especially for large genomes like the human genome. In practice, this is often achieved by the pool-seq approach (also known as bulk sequencing) [12], in which the variations and the variant allele frequencies (VAF) are inferred from the whole genome sequencing (WGS) of samples containing randomly

pooled cells that represent a mixture of clones.

From pool-seq data, it is non-trivial to reconstruct the clonal evolutionary structure from the inferred allele frequencies of variations, and many computational approaches have been developed to tackle this problem. It is worth noting that in some cases, we have data from only one pooled sample (mixture of clones) while in other cases, multiple samples from the same evolving population are available. These samples may distinguish themselves either by time and/or space, *i.e.*, they are sampled at different time points during the evolution process and/or at different physical locations of the population. Here, we refer to them as the time-resolved and space-resolved samples, respectively.

In this review, we aim to describe various existing computational approaches addressing the clonal reconstruction problem — the problem of inferring the haplotypes of all the clones in a given population and their ancestral relationships from the variations and the variant allele frequencies within the population. These approaches will be compared based on their problem formulation, modelling strategies and the type of input data. The type of input data is a defining and discerning feature in understanding the algorithmic ideas of clonal reconstruction computation. Some algorithmic ideas are solely based around SNV data while others are based on CNV and SV data or a mixture of both. Whether these data are derived from bulk sequencing of a single sample or multi-sample or single-cell sequencing, clonal reconstruction remains an active area of research. We will address the advantages and limitations of currently existing algorithmic methods and discuss some open computational problems related to clonal reconstruction that need to be addressed in the future. Also note that for a more detailed review of the evolution of tumor phylogenetics, refer to Schwartz and Schaffer's excellent review [13].

## CLONAL RECONSTRUCTION PROBLEM

Consider an evolving clonal population with  $n$  observed variations and  $m$  samples from the population collected and used to measure the variant allele frequencies (VAFs). The resulting  $m \times n$  frequency matrix  $F$  is provided as the input to the clonal reconstruction problem. The clonal evolution model follows the infinite sites assumption (ISA) as proposed by Mitoo Kimura in 1969 [14], which states that (1) a variation occurs at a single locus at most once during the period of evolutionary process and cannot be lost by subsequent reversal mutation; (2) there is no recombination, and (3) all cells in the evolving population are assumed to be descendants of a single founding clone. Under these assumptions, the ancestral relationships between the clones in the evolving population can be

represented as a directed tree  $T$ , (*i.e.*, the clonal tree), in which the root node represents the founder clone, each other node represents a clone introduced by one or more novel variations, and each edge represents the direct ancestral relationships between the clones. Each edge in the clonal tree is labeled by the variation(s) that distinguishes the child from its parent. Each node has exactly one parent, while each parent may have multiple children. When more than one mutation occurs during the evolution from the parent to the child, they can be clustered together and considered as a single variation group. As a result, the *haplotype* of a clone, represented by the set of variations that uniquely identify the clone, corresponds to the unique path from the root (the founder clone) to the node representing the clone. Each leaf node thus represents a clone observed at the end of the evolutionary process. The clonal tree can also be represented as a  $n \times n$  binary matrix  $B$  where each row represents a clone (a node) in  $T$ , and each column represents a variation. The rows contain 1's where the corresponding variation is present in the clone and 0's where it is absent.

Now assume that we represent the unknown frequencies of the  $n$  clones in the  $m$  samples using a  $m \times n$  matrix  $C$ . Then the observed VAF matrix  $F$  must satisfy the equation

$$F = C \times B. \quad (1)$$

The goal of the clonal reconstruction problem is to characterize (1) the haplotype of each clone, (2) the frequency of each clone in each sample, and (3) the evolutionary relationships between the clones (*i.e.*, the clonal tree) that best explain the observed data  $F$ . Formally, it is often formulated as a matrix factorization problem: given the observed VAF matrix  $F$ , find the clonal frequency matrix  $C$  and the clonal matrix  $B$  such that Equation (1) is satisfied (Figure 1). Typically there are many valid combinations of  $C$  and  $B$  that satisfy Equation (1). A naive approach to solve this problem is to enumerate all possible clonal matrices  $B$  and the corresponding clonal frequency matrices  $C$  that satisfy Equation (1). Among all the valid solutions, only one solution is chosen based on biological constraints and heuristics.

## BIOLOGICAL PROBLEM FORMULATIONS

### Space-resolved versus time-resolved sampling

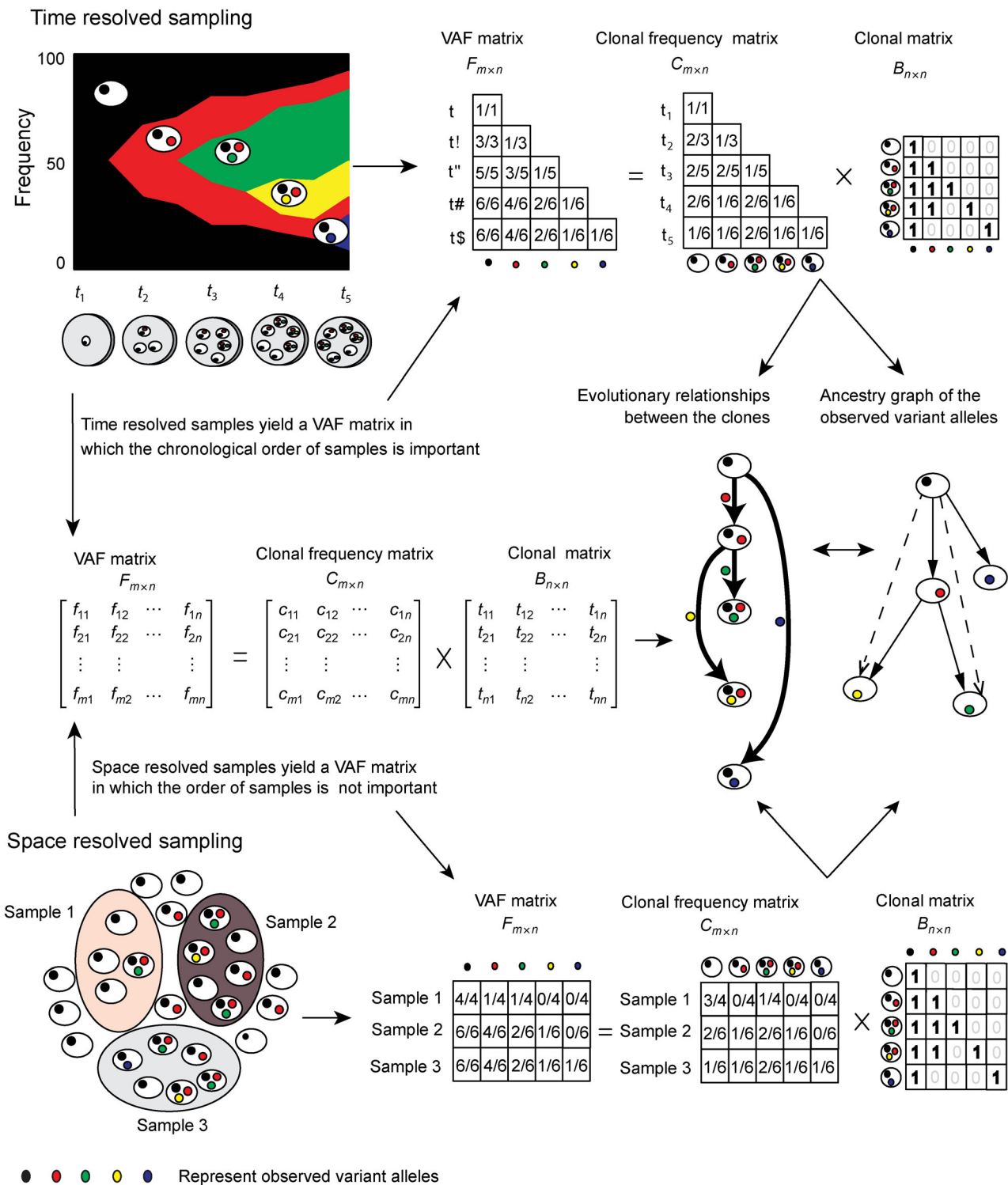
Methods to solve the clonal reconstruction problem depend on the number and types of samples obtained from the evolving population. When there is only one

sample available, very limited information can be exploited that will help reduce the solution space. There are only a few methods developed for single sample VAF data. The data from more than one sample provides additional information and constraints that will allow us to define the best solution among the many possible solutions (clonal trees). The multi-sample data that are usually available for clonal studies can broadly be classified into two categories: the space-resolved data and the time-resolved data. In many cancer genomic studies, variations are measured in multiple spatially distinct regions from the same tumor tissue [15], which falls under space-resolved frequency data. Methods that use this kind of data rely on the samples containing heterogeneous but overlapping sets of clones across samples. This allows the clonal reconstruction methods to select only those solutions that fit the model across all samples. On the other hand, whole-genome sequencing for long term evolution experiments (LTEEs) often provide time-resolved VAF data collected from the evolving population at different time points during evolution process, which provide not only the snapshots of variant calls such as VAFs in the population but also the chronological order of the mutations. This information can be exploited for the proposal of a likelihood function that depends on the order of mutations and the VAFs over time, which in turn allows us to find the maximum likelihood solution that fit the model and the data. It should be noted that we may be able to obtain a combination of space-resolved and time-resolved data in the future. Notably, in addition to these two kinds of multi-sample pool-seq data, whole genome sequencing performed on randomly isolated clones or single cell sequencing data are sometimes available at a low coverage, providing complementary information for clonal reconstruction.

### Single-nucleotide variations (SNVs) versus copy number variations (CNVs)

Irreparable alterations in DNA sequence are the driving force of evolutionary processes, which include point mutations (resulting in SNVs) and genome rearrangements (resulting in structural variations such as CNVs).

The classical evolution theory substantiates that only mutations that occur in reproductive cells (known as the germ line mutations) are essential for evolution of most sexually reproduced higher organisms, whereas mutations in non-reproductive cells (*i.e.*, the somatic mutations) are often non-essential. However, in clonal populations, novel DNA mutations in somatic cells have significant impact on the clonal expansion. Nevertheless, there is a parallelism between the mechanism of evolution in these two distinct evolving populations: in higher-organisms,



**Figure 1. Overview of the clonal reconstruction problem.** The input data to the clonal reconstruction problem could be either time-resolved (top panel) or space-resolved (bottom panel). The middle panel illustrates the formal formulation of the clonal reconstruction problem. Suppose  $m$  samples are collected and  $n$  variants are detected. The clonal reconstruction problem is formulated as the factorization problem, *i.e.*, to find both the matrix  $C$  representing the fraction of each clone contained by each sample, and the clonal matrix  $B$  representing the set of variants contained by each clone, that are consistent with the observed VAF data. Note that the matrix  $B$  can be directly derived from a clonal tree  $T$  and *vice-versa*. The ancestry graph (rightmost in the middle panel) shows all the possible ancestral relationships between the variations as dictated by the ancestry condition. The solid arrows show the clonal tree which is a spanning arborescence of the ancestry graph.

the evolution acts on the individuals while in clonal populations it acts on the cells, leading to cancer evolution theory of cancer for understanding the tumorigenesis and tumor progression from an evolutionary perspective.

According to the cancer evolution theory, cancer is viewed as an abnormal health condition resulting mainly from cumulative DNA mutations subjected to the selection pressure on the genomes of somatic cells in an organism. Every tumor is thus characterized by a glaring cell-to-cell genetic variability termed intra-tumor heterogeneity. For the most part, the evolution and clonal expansion of tumors are highly driven by point mutations. Nevertheless, CNVs and other structural variations (SVs) play key roles in tumor evolution, and their impact has been increasingly acknowledged [16–18]. Therefore, characterizing tumor clones and their evolutionary relationships provides insights into the mechanism of tumorigenesis, tumor progression, and the response to cancer therapy.

Understandably, because the cancer evolution is highly dependent on point mutations occurring in the tumor tissues, a majority of the current algorithmic research on clonal reconstruction focus on the variant allele frequency data to infer the clones and their evolutionary relationships. However, CNVs and other SVs may affect the allele frequency of the SNVs. Thus, methods that are based only on SNVs are often restricted to regions that are free of large-scale variations. To address this limitation, computational methods that take both SNVs and CNVs as input become available for clonal reconstruction in cancer genomics (as shown in Table 1). Even though SVs in evolving bacterial populations are often neglected, a general computational framework for clonal reconstruction that adequately address the biological significance of CNVs in such populations is still desirable. As CNVs evolve due to segmental duplications, reconstructing the clonal composition of CNVs from pool-seq data requires different problem formulations as the clonal reconstruction problem solely on SNVs, *e.g.*, to minimize the number of duplication events in the clonal phylogeny, as proposed recently by Eaton *et al.* [54] and Lei *et al.* [55]. RCK [56] on the other hand addresses chromosomal aberrations and attempts to reconstruct the clonal haplotypes from bulk sequencing data. In addition, the clonal reconstruction problem involving SNVs, CNVs and other SVs together remain as an open challenge.

## ALGORITHMIC APPROACHES

Many methods were proposed to solve the clonal reconstruction problem in different biological contexts. These methods broadly fall under two main categories depending on their problem formulation and algorithmic approaches: combinatorial or probabilistic. There are also

methods that combine the two types of formulations. We will discuss these formulations and algorithms in this section (see Table 1 for a summary).

### Combinatorial formulations

The combinatorial formulation of the clonal reconstruction problem (also referred to as the variant allele frequency factorization problem (VAFFP) [15]) attempts to characterize the combinatorial space of valid solutions to Equation (1) and the computational complexity of finding a valid solution for a given  $\mathcal{F}$ . Let  $\mathcal{T}$  be the space of all rooted clonal tree and  $\mathcal{C}$  be the space of all  $m \times n$  clonal frequency matrices. The combinatorial approaches aim at finding  $\mathbf{T}' \in \mathcal{T}$  and  $\mathbf{C}' \in \mathcal{C}$  by traversing the solution space in efficient ways and discarding solutions that are biologically not relevant. These approaches use various heuristics like maximum parsimony — minimizing the total number of evolutionary events required to explain the observed data, minimum number of clones and shallowness of the clonal tree to reduce the searching space and make it computationally feasible to find a suitable solution. Some methods define an optimization criteria such that to minimize the error of assigning each variation to a clone while finding the global phylogeny that satisfies pre-defined topological constraints, and is consistent with the ISA framework.

To this end, El-Kebir *et al.* [15] proved two conditions — the ancestry condition and the sum condition, that the solutions need to satisfy in order for them to be consistent with ISA. The ancestry condition states that no two variations can be assigned to the same clone unless the frequency of one variation is always greater than the other across all samples. This is a necessary condition that gives rise to a directed graph (named the Ancestry Graph)  $\mathcal{G}$  that represents all the possible ancestral relationships between the variations in  $\mathcal{F}$  (Figure 1). So the problem of finding a clonal tree in  $\mathcal{T}$  is reduced to the problem of finding a spanning arborescence (directed spanning tree) of  $\mathcal{G}$  because all the other trees in  $\mathcal{T}$  are invalid according to the ancestry condition. While the ancestry condition provides a necessary condition, sufficiency is provided by the sum condition, which states that the VAF of a parent node is at least equal to the sum of the VAFs of its children in each sample. This further reduces the searching space to find the spanning arborescence of  $\mathcal{G}$  that also satisfies the sum condition. Nonetheless, it is shown that the VAFFP is a NP-complete problem, and the authors proposed an integer linear programming (ILP) approach for finding the largest spanning arborescence in an ancestry graph that satisfies the sum condition. In this ILP formulation, the graph problem is reduced to an ILP — a mathematical optimization problem where a linear objective function is maximized (or minimized)

**Table 1** Various methods addressing the clonal reconstruction problem

Method	Approach	Multi-sample	SNVs	Read depth/CNVs	ISA	Phylogeny	Ref.
AncesTree	Combinatorial	y	y	n	y	y	[15]
CITUP	Combinatorial	y	y	n	y	y	[19]
Clomial	Probabilistic	n	y	n	n	n	[20]
CloneHD	Probabilistic	y	y	y	n	n	[21]
CNT-MD	Combinatorial	y	n	y	n	y	[22]
MIPUP	Combinatorial	y	y	n	y	y	[23]
LICHeE	Combinatorial	y	y	n	y	y	[24]
PhyloSub	Probabilistic	y	y	n	y	y	[25]
PhyloWGS	Probabilistic	y	y	y	y	y	[26]
PyClone	Probabilistic	y	y	n	y	n	[27]
Rec-BTP	Combinatorial	n	y	n	y	y	[28]
SciClone	Probabilistic	y	y	y	n	n	[29]
ThetA	Probabilistic	n	n	y	n	n	[30]
TrAp	Combinatorial	n	y	n	y	y	[31]
QuantumClone	Probabilistic	y	y	n	n	n	[32]
CTPSingle	Combinatorial	n	y	n	y	y	[33]
ClonalTREE	Probabilistic	y	y	n	y	y	[34]
SPRUCE	Combinatorial	y	y	y	n	y	[35]
TargetClone	Probabilistic	y	y	n	y	y	[36]
BitPhylogeny	Probabilistic	y	y	n	y	y	[37]
Canopy	Probabilistic	y	y	y	y	y	[38]
TITAN	Probabilistic	n	n	y	n	n	[39]
CALDER	Combinatorial	y	y	n	y	y	[40]
Bayclone	Probabilistic	y	y	n	n	n	[41]
BayClone2	Probabilistic	y	y	y	n	n	[42]
CloneFinder	Regression	y	y	n	n	y	[43]
Cloe	Probabilistic	y	y	n	y	y	[44]
TreeClone	Probabilistic	y	y	n	n	y	[45]
PairClone	Probabilistic	y	y	n	n	n	[46]
SubcloneSeeker	Combinatorial	y	y	y	n	n	[47]
SiFit	Probabilistic	SCS	y	n	n	y	[48]
SCITE	Probabilistic	SCS	y	n	y	n	[49]
OncoNEM	Probabilistic	SCS	y	n	y	y	[50]
SPhyR	Combinatorial	SCS	y	n	n	y	[51]
B-SCITE	Probabilistic	SCS/BSD	y	n	y	y	[52]
ddClone	Probabilistic	SCS/BSD	y	y	n	n	[53]

SNVs: single nucleotide variations, CNVs: copy number variations, ISA: infinite sites assumption, SCS: single cell sequencing data, BSD:bulk sequencing data, y:yes and n:no.

over a solution space specified by linear constraints. In this case, the objective function is the maximum number of edges in the spanning arborescence. It is to be noted that the ILP approach only proves the existence of a valid solution and does not discriminate between multiple solutions. It also assumes that the VAF matrix is error free. To account for errors in VAF estimation and to define and find the best solution, they proposed a probabilistic framework which leads to an approximate ancestry graph,

and then used a mixed ILP (MILP) formulation to find the largest arborescence in the approximate ancestry graph that satisfies the sum condition. This method is implemented in AncesTree [15]. It is worth noting that the same group developed another combinatorial method called SPRUCE [35] that infers the phylogeny from both SNV and CNV data, which uses the infinite alleles assumption instead of the ISA. This complements those methods which often ignore the genomic regions with copy

number aberrations. Another method which uses a similar framework as AncestryTree is LICHeE [24], which also constructs an ancestry graph and then looks for a spanning tree that satisfies the sum condition, but uses a sophisticated backtracking algorithm to search all the spanning trees. Since the running time of this algorithm depends on the number of spanning trees in a given ancestry graph and some graphs may have many spanning trees, they provide a bound to the number of trees that are explored. Recently, Myers *et al.* [40] published CALDER which introduces a variant of VAFFP called Longitudinal VAFFP (or LVAFFP) where the input VAF data from the samples are time-resolved instead of space-resolved. The longitudinal order of mutations provide additional constraints to the solution. So, not all solutions to the VAFFP are solutions to the LVAFFP. The longitudinal order constraint is modeled using another tree (the longitudinally observed clonal tree) with nodes colored based on the times the clone is observed. Spanning arborescences of  $\mathcal{G}$  that satisfy the sum condition and is consistent with the longitudinal observed clonal tree obtained from the corresponding  $\mathcal{C}$  matrix, are considered as valid solutions.

One of the earliest studies that formalizes the clonal reconstruction problem in a combinatorial framework is designed for single sample VAF input. This method—TrAP [31], attempts to find the clonal evolutionary tree that sequentially minimizes the number of clones, the number of populated clones (*i.e.*, clones whose frequencies in the sample is  $> 0$ ), and the depth of the clonal tree. The algorithm enumerates all possible trees in a greedy approach that sequentially optimizes each of the above criteria. An alternative method, CITUP [19] uses a quadratic integer programming (QIP) formulation to minimize the squared error in assigning variations to clones. To avoid overfitting, CITUP also minimizes the Bayesian information criterion (BIC), assuming that the noise is normally distributed. BIC is a method used in model selection to reduce overfitting by introducing a penalty term for the number of parameters in the model. CITUP iterates through all tree topologies up to a user defined threshold, and hence it is a combinatorial algorithm. A related method—CTPsingle [33] developed by the same group, is designed for low-coverage sequencing data from a single sample. This method takes read depths instead of VAFs as input to infer the number of clones using a probabilistic approach for clustering in advance of clonal tree reconstruction. It then uses a MILP formulation for the clonal tree reconstruction.

Yet another related work designed for single sample data is Rec-BTP [28]. This tool operates under the assumption that each clonal expansion splits the current population into exactly two sub-populations giving rise to

a rooted binary tree representing the ancestral relationships between sub-populations. Hence, the problem is formulated as to find a binary tree partition (BTP) that satisfies the sum condition, which is proved to be NP-complete. The program Rec-BTP is a recursive algorithm for solving  $\varepsilon$ -BTP problem, a variant of the BTP problem that accounts for errors in VAFs.

### Probabilistic formulations

Probabilistic approaches in clonal reconstruction attempt to model the problem as a probabilistic inference problem, where the solutions are assumed to be distributed over a standard probability distribution. Then an inference algorithm is used to produce the solution that has the highest probability of generating the observed data—in this case, the input VAF data. For example, PhyloSub [25] defines a generative probabilistic model that attempts to explain the observed VAFs in terms of an unobserved clonal tree. It uses Bayesian inference based on Markov Chain Monte Carlo (MCMC) sampling algorithm to infer a distribution over phylogenies, where the Dirichlet distribution over all potential phylogenies is used as the prior. The inference itself implicitly models the ancestry condition and the sum condition by inferring the unknown clonal tree that has the highest probability of generating the observed VAF data. PhyloWGS [26] is a method that uses the same probabilistic model as PhyloSub but applied to the CNV data in addition to SNVs. Another method that uses a similar Bayesian framework is Canopy [38], which jointly models SNVs and CNVs allowing inference of temporal order of the CNV events in relation to the SNV events.

ClonalTREE [34] like CALDER discussed in combinatorial approaches, is a method developed for clonal reconstruction from time-resolved multi-sample VAF data. It uses a framework where a likelihood function is defined based on the assumption that at any given time, the likelihood of a candidate clone to acquire a new variation and hence spawn a new clone is proportional to the frequency of the clone in the population at that time. This assumption solely relies on the advantage that time-resolved sampling provides—the order of variation occurrence. The method also provides an option to incorporate clonal sequencing data as additional constraints for reconstructing the unknown clonal haplotypes, referred to as the hybrid input.

### Other approaches

Other methods that address clonal reconstruction but do not exactly fit into the specific problem definition that we discuss here are also listed in Table 1. Among them, Clomial [20], CloneHD [21], PyClone [27], SciClone

[29], ThetA [30] and QuantumClone [32] infer the haplotypes of clones but do not infer their phylogeny. CNT-MD [22] was specifically designed for CNV data. MIPUP [23] takes a binary input matrix that represents the presence or absence of a variation in a sample based on a given threshold of VAF and attempts to find the minimum perfect phylogeny. SiFit [48], SCITE [49], OncoNEM [50] and SPhyR [51] are designed for clonal reconstruction in single cell sequencing data. TargetClone [36] is a method specifically designed for targeted sequencing data obtained from microdissected tumor samples. It relies on the assumption that each sample contained approximately one major subclone. BitPhylogeny [37] reconstructs phylogeny from methylation patterns and from single-cell exomes. B-SCITE [52] and ddClone [53] are the first computational approaches that integrate pool-seq data and SCS data to infer the clonal composition and the corresponding clonal tree. B-SCITE uses a joint likelihood model of both SCS data and pool-seq data and uses a Markov chain Monte Carlo approach to search for a single maximum likelihood mutation tree that is a joint fit for both SCS data and pool-seq data. On the other hand, in ddClone, SCS data are used to inform clusters of mutation in pool-seq data: SCS data are used to derive a non-parametric Bayesian prior used by the likelihood model based on pool-seq data to infer clonal composition, but it does not find the clonal tree. However, both B-SCITE and ddClone use SNV data. Lei *et al.* [55] proposed and formulated a mixed membership model integrating both SCS data and pool-seq data for clonal decomposition based only on CNV data using non-negative matrix factorization. In their approach, the clonal deconvolution from pool-seq data is gauged and optimized by its similarity to SCS data via a coordinate descent algorithm. Integrative approaches combining both SCS data and pool-seq data are needed. Finally, we note that a few algorithms were recently developed for dimension reduction and clustering in single-cell RNA-seq (scRNA-seq) data analysis [57], which have striking resemblances with the clonal reconstruction algorithms. Although they address a very different biological problem—in case of scRNA-seq data analysis, the goal is to characterize the similarities of gene expression profiles among cells, they benefit from the same algorithmic framework based on matrix factorization [58–60].

## CONCLUSION

Clonal theory of natural evolving unicellular organisms draws parallelism with tumor evolution. However, most of the available clonal sequencing data and algorithms were developed for elucidating tumor evolution whereas clonal reconstruction for unicellular genomes are less

addressed. In fact, the clonal reconstruction for unicellular genomes (*e.g.*, using the WGS data from bacterial LTEE [34]) serves as an excellent system for benchmarking the clonal reconstruction algorithms because (1) the genomes are much smaller and thus sequencing (even at a very high coverage) cost a relatively low price; (2) the reference genome is often well characterized; (3) the variations of interests are usually fewer and simpler (*e.g.*, typically no more than a few hundreds variations, mostly SNVs) than those occurring in cancer evolution); and (4) it is relatively straightforward to obtain time-resolved and/or space-resolved samples. Therefore, one future research is to generate high quality benchmark datasets from evolving microbial population, which can be used to evaluate different clonal reconstruction algorithms, as discussed below.

Despite the rapid advance of clonal reconstruction algorithms, there remain some open problems and challenges. One open problem is evaluating the different methods and comparing their performance. The major challenge in such an endeavor is the fact that each method uses different sets of model assumptions and objective functions to optimize which leads to difficulties in creating unbiased simulated data for evaluation. The community would greatly benefit from curation of comprehensive biological datasets consisting of pool-seq data from multiple samples matched with single-cell sequencing data that accurately define the haplotypes of clones present in each sample. Such datasets would prove to be valuable benchmarking resources that will not only help compare the performances of various methods discussed above but also help prove the model assumptions made by each method. Other open problems include solving the clonal reconstruction problem that take as input both time-resolved and space resolved samples simultaneously. Similarly, algorithmic approaches encompassing different types of data, for example, matched allele frequency and gene expression data, for clonal reconstruction are not yet available and remain a challenge for future researchers.

The most trivial approach to obtain the haplotypes of all clones in an evolving population is to acquire single-cell sequencing at a sufficient depth on many cells randomly selected from the population. In this case, the reconstruction of evolutionary relationships among the clones becomes the traditional phylogeny problem. But currently this approach is not feasible because single-cell sequencing (SCS) is still either expensive or has high error rate (*e.g.*, by using the Nanopore technique). Hence, most of the tools currently available focus variant allele frequencies derived from bulk sequencing (pool-seq) data. VAF generated from bulk sequencing are well suited to detect and reconstruct high abundant clonal populations but may fail to reconstruct clones with CNVs. Another advantage



of VAF from bulk sequencing data is that they allow temporal order of mutations. However, these topological and ancestral constraints make the characterization and identification of branching events very challenging. On the other hand, clonal reconstruction from SCS data can characterize cell-to-cell genomic heterogeneity and generate an accurate phylogeny. In addition, SCS data capture branching events, despite of the lack of the temporal order of mutations. Integrating both SCS and bulk sequencing data presents a great potential in effectively and accurately addressing the clonal reconstruction problem; however, there is only a handful of available tools that integrate both types of data [52,53,55]

## ABBREVIATIONS

DNA	deoxyribonucleic acid
WGS	whole genome sequencing
SNV	single nucleotide variation
CNV	copy number variation
SV	structural variation
LTEE	long term evolution experiment
VAF	variant allele frequency
VAFFP	variant allele frequency factorization problem
ISA	infinite sites assumption
ILP	integer linear programming
MILP	mixed integer linear programming
QIP	quadratic integer programming
BTP	binary tree partition
MCMC	Markov Chain Monte Carlo
BIC	Bayesian information criterion

## ACKNOWLEDGEMENTS

This research was partially supported by a Multidisciplinary University Research Initiative Award W911NF-09-1-0444 from the US Army Research Office, the National Institute of Health grant 1R01AI108888 and Indiana University (IU) Precision Health Initiative (PHI). We thank Drs. Megan Behringer and Michael Lynch for very inspiring discussions.

## COMPLIANCE WITH ETHICS GUIDELINES

The authors Wazim Mohammed Ismail, Etienne Nzabarushimana and Haixu Tang declare that they have no conflict of interests.

This article is a review article and does not contain any studies with human or animal subjects performed by any of the authors.

## REFERENCES

- Shapiro, B. J. (2016) How clonal are bacteria over time? *Curr. Opin. Microbiol.*, 31, 116–123
- Tibayrenc, M., Kjellberg, F. and Ayala, F. J. (1990) A clonal theory of parasitic protozoa: the population structures of *Entamoeba*, *Giardia*, *Leishmania*, *Naegleria*, *Plasmodium*, *Trichomonas*, and *Trypanosoma* and their medical and taxonomical consequences. *Proc. Natl. Acad. Sci. USA*, 87, 2414–2418
- Blount, Z. D., Barrick, J. E., Davidson, C. J. and Lenski, R. E. (2012) Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature*, 489, 513–518
- Wielgoss, S., Barrick, J. E., Tenaillon, O., Cruveiller, S., Chane-Woon-Ming, B., Médigue, C., Lenski, R. E. and Schneider, D. (2011) Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. *G3: Genes, Genom. Genet.*, 1, 183–186
- Behringer, M. G., Choi, B. I., Miller, S. F., Doak, T. G., Karty, J. A., Guo, W. and Lynch, M. (2018) *Escherichia coli* cultures maintain stable subpopulation structure during long-term evolution. *Proc. Natl. Acad. Sci. USA*, 115, E4642–E4650
- Pon, J. R. and Marra, M. A. (2015) Driver and passenger mutations in cancer. *Annu. Rev. Pathol.*, 10, 25–50
- Lenski, R. E., Rose, M. R., Simpson, S. C. and Tadler, S. C. (1991) Long-term experimental evolution in *Escherichia coli*. I. adaptation and divergence during 2,000 generations. *Am. Nat.*, 138, 1315–1341
- Lenski, R. E., Wisser, M. J., Ribeck, N., Blount, Z. D., Nahum, J. R., Morris, J. J., Zaman, L., Turner, C. B., Wade, B. D., Maddamsetti, R., *et al.* (2015) Sustained fitness gains and variability in fitness trajectories in the long-term evolution experiment with *Escherichia coli*. *P. Roy. Soc. B-Biol. Sci.* 282, 20152292
- Pluacain, J., Hindré, T., Le Gac, M., Tenaillon, O., Cruveiller, S., Médigue, C., Leiby, N., Harcombe, W. R., Marx, C. J., Lenski, R. E., *et al.* (2014) Epistasis and allele specificity in the emergence of a stable polymorphism in *Escherichia coli*. *Science*, 343, 1366–1369
- Rozen, D. E. and Lenski, R. E. (2000) Long-term experimental evolution in *Escherichia coli*. VIII. dynamics of a balanced polymorphism. *Am. Nat.*, 155, 24–35
- Wisser, M. J., Ribeck, N. and Lenski, R. E. (2013) Long-term dynamics of adaptation in asexual populations. *Science*, 342, 1364–1367
- Taus, T., Futschik, A. and Schlötterer, C. (2017) Quantifying selection with pool-seq time series data. *Mol. Biol. Evol.*, 34, 3023–3034
- Schwartz, R., Schöffner, A.A. (2017) The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.*, 18, 213–229
- Kimura, M. (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61, 893–903
- 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, i62–i70
- Ng, C. K., Cooke, S. L., Howe, K., Newman, S., Xian, J., Temple, J., Batty, E. M., Pole, J. C., Langdon, S. P., Edwards, P. A., *et al.* (2012) The role of tandem duplicator phenotype in tumour evolution in high-grade serous ovarian cancer. *J. Pathol.*, 226,

- 703–712
17. Yang, L., Luquette, L. J., Gehlenborg, N., Xi, R., Haseley, P. S., Hsieh, C. H., Zhang, C., Ren, X., Protopopov, A., Chin, L., *et al.* (2013) Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, 153, 919–929
  18. Quigley, D. A., Dang, H. X., Zhao, S. G., Lloyd, P., Aggarwal, R., Alumkal, J. J., Foye, A., Kothari, V., Perry, M. D., Bailey, A. M., *et al.* (2018) Genomic hallmarks and structural variation in metastatic prostate cancer. *Cell*, 174, 758–769.e9
  19. Malikic, S., McPherson, A. W., Donmez, N. and Sahinalp, C. S. (2015) Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31, 1349–1356
  20. Zare, H., Wang, J., Hu, A., Weber, K., Smith, J., Nickerson, D., Song, C., Witten, D., Blau, C. A. and Noble, W. S. (2014) Inferring clonal composition from multiple sections of a breast cancer. *PLOS Comput. Biol.*, 10, e1003703
  21. Fischer, A., Vázquez-García, I., Illingworth J. R. C., and Mustonen, V. (2014) High-definition reconstruction of clonal composition in cancer. *Cell Reports*, 7, 1740–1752
  22. Zaccaria, S., El-Kebir, M., Klau, G. W. and Raphael, B. J. (2017) The copy-number tree mixture deconvolution problem and applications to multi-sample bulk sequencing tumor data. In: *International Conference on Research in Computational Molecular Biology*, pp. 318–335. Springer
  23. Husić, E., Li, X., Hujdurović, A., Mehine, M., Rizzi, R., Mäkinen, V., Milanić, M. and Tomescu, A. I. (2019) MIPUP: minimum perfect unmixed phylogenies for multi-sampled tumors via branchings and ILP. *Bioinformatics*, 35, 769–777
  24. Popic, V., Salari, R., Hajirasouliha, I., Kashef-Haghighi, D., West, R. B. and Batzoglou, S. (2015) Fast and scalable inference of multi-sample cancer lineages. *Genome Biol.*, 16, 91
  25. Jiao, W., Vembu, S., Deshwar, A. G., Stein, L. and Morris, Q. (2014) Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*, 15, 35
  26. Deshwar, A. G., Vembu, S., Yung, C. K., Jang, G. H., Stein, L. and Morris, Q. (2015) PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.*, 16, 35
  27. Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A. and Shah, S. P. (2014) PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods*, 11, 396–398
  28. Hajirasouliha, I., Mahmoody, A. and Raphael, B. J. (2014) A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, 30, i78–i86
  29. Miller, C. A., White, B. S., Dees, N. D., Griffith, M., Welch, J. S., Griffith, O. L., Vij, R., Tomasson, M. H., Graubert, T. A., Walter, M. J., *et al.* (2014) SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLOS Comput. Biol.*, 10, e1003665
  30. Oesper, L., Mahmoody, A. and Raphael, B. J. (2013) THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol.*, 14, R80
  31. Strino, F., Parisi, F., Micsinai, M. and Kluger, Y. (2013) TrAp: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Res.*, 41, e165
  32. Deveau, P., Colmet Daage, L., Oldridge, D., Bernard, V., Bellini, A., Chicard, M., Clement, N., Lapouble, E., Combaret, V., Boland, A., *et al.* (2018) QuantumClone: clonal assessment of functional mutations in cancer based on a genotype-aware method for clonal reconstruction. *Bioinformatics*, 34, 1808–1816
  33. Donmez, N., Malikic, S., Wyatt, A. W., Gleave, M. E., Collins, C. C. and Sahinalp, S. C. (2017) Clonality inference from single tumor samples using low-coverage sequence data. *J. Comput. Biol.*, 24, 515–523
  34. Mohammed Ismail, W. and Tang, H. (2019) Clonal reconstruction from time course genomic sequencing data. In: *International Conference on Intelligent Biology and Medicine*
  35. El-Kebir, M., Satas, G., Oesper, L. and Raphael, B. J. (2016) Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. *Cell Syst.*, 3, 43–53
  36. Nieboer, M. M., Dorssers, L. C. J., Straver, R., Looijenga, L. H. J. and de Ridder, J. (2018) TargetClone: A multi-sample approach for reconstructing subclonal evolution of tumors. *PLoS One*, 13, e0208002
  37. Yuan, K., Sakoparnig, T., Markowetz, F. and Beerenwinkel, N. (2015) BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biol.*, 16, 36
  38. Jiang, Y., Qiu, Y., Minn, A. J. and Zhang, N. R. (2016) Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc. Natl. Acad. Sci. USA*, 113, E5528–E5537
  39. Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L. M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., *et al.* (2014) TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.*, 24, 1881–1893
  40. Myers, M. A., Satas, G. and Raphael, B. J. (2019) Calder: Inferring phylogenetic trees from longitudinal tumor samples. *Cell Syst.*, 8, 514–522.e5
  41. Sengupta, S., Wang, J., Lee, J., Müller, P., Gulukota, K., Banerjee, A. and Ji, Y. (2014) Bayclone: Bayesian nonparametric inference of tumor subclones using NGS data. In: *Pacific Symposium on Biocomputing Co-Chairs*, pp. 467–478. World Scientific
  42. Lee, J., Müller, P., Sengupta, S., Gulukota, K. and Ji, Y. (2016) Bayesian inference for intratumour heterogeneity in mutations and copy number variation. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 65, 547–563
  43. Miura, S., Gomez, K., Murillo, O., Huuki, L. A., Vu, T., Buturla, T. and Kumar, S. (2018) Predicting clone genotypes from tumor bulk sequencing of multiple samples. *Bioinformatics*, 34, 4017–4026
  44. Marass, F., Moulriere, F., Yuan, K., Rosenfeld, N. and Markowetz, F. (2016) A phylogenetic latent feature model for clonal deconvolution. *Ann. Appl. Stat.*, 10, 2377–2404
  45. Zhou, T., Sengupta, S., Müller, P. and Ji, Y. (2019) Treeclone: Reconstruction of tumor subclone phylogeny based on mutation pairs using next generation sequencing data. *Ann. Appl. Stat.*, 13,

- 874–899
46. Zhou, T., Müller, P., Sengupta, S. and Ji, Y. (2019) Pairclone: a bayesian subclone caller based on mutation pairs. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 68, 705–725
  47. Qiao, Y., Quinlan, A. R., Jazaeri, A. A., Verhaak, R. G., Wheeler, D. A. and Marth, G. T. (2014) SubcloneSeeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization. *Genome Biol.*, 15, 443
  48. Zafar, H., Tzen, A., Navin, N., Chen, K. and Nakhleh, L. (2016) Sifit: a method for inferring tumor trees from single-cell sequencing data under finite-site models. *Genome Biol.*, 18, 178
  49. Davis, A. and Navin, N. E. (2016) Computing tumor trees from single cells. *Genome Biol.*, 17, 113
  50. Ross, E. M. and Markowitz, F. (2016) OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biol.*, 17, 69
  51. El-Kebir, M. (2018) SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics*, 34, i671–i679
  52. Malikić, S., Jahn, K., Kuipers, J., Sahinalp, C. and Beerenwinkel, N. (2017) Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nat. Commu.*, 10, 2750
  53. Salehi, S., Steif, A., Roth, A., Aparicio, S., Bouchard-Côté, A. and Shah, S. P. (2017) ddClone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. *Genome Biol.*, 18, 44
  54. Eaton, J., Wang, J. and Schwartz, R. (2018) Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics*, 34, i357–i365
  55. Lei, H., Lyu, B., Gertz, E. M., Schaeffer, A. A., Shi, X., Wu, K., Li, G., Xu, L., Hou, Y., Dean, M., *et al.* (2019) Tumor copy number deconvolution integrating bulk and single-cell sequencing data. In: *International Conference on Research in Computational Molecular Biology*, pp. 174–189. Springer
  56. Aganezov, S. and Raphael, B. J. (2019) Reconstruction of clone- and haplotype-specific cancer genome karyotypes from bulk tumor samples. *bioRxiv*
  57. Chen, G., Ning, B., Shi, T. (2019) Single-cell RNA-seq technologies and related computational data analysis. *Front. Genet.*, 10, 317–317
  58. Ferreira, P. F., Carvalho, A. M. and Vinga, S. (2018) Scalable probabilistic matrix factorization for single-cell RNA-seq analysis. *bioRxiv*
  59. Durif, G., Modolo, L., Mold, J. E., Lambert-Lacroix, S. and Picard, F. (2019) Probabilistic count matrix factorization for single cell expression data analysis. *Bioinformatics*, 35, 4011–4019
  60. Sun, S., Chen, Y., Liu, Y. and Shang, X. (2019) A fast and efficient count-based matrix factorization method for detecting cell types from single-cell RNAseq data. *BMC Syst. Biol.*, 13, 28