Reconstructing clonal lineage trees incorporating single nucleotide variants (SNVs), copy number alterations (CNAs), and structural variations (SVs)

Xuecong Fu[1], Haoyun Lei[2], Yifeng Tao[2], and Russell Schwartz[1,2*]

[1]Department of Biological Sciences, Carnegie Mellon University, PA, 15213
[2]Computational Biology Department, Carnegie Mellon University, PA, 15213

January 16, 2022

Abstract

Cancer develops through a process of clonal evolution in which an initially healthy cell gives rise to progeny gradually differentiating through the accumulation of genetic and epigenetic mutations. These mutations can take various forms, including single nucleotide variants (SNVs), copy number alterations (CNAs), or structural variations (SVs), with each variant type providing complementary insights into tumor evolution as well as offering distinct challenges to phylogenetic inference. In the present work, we develop a tumor phylogeny method, TUSV-ext, that incorporates SNVs, CNAs, and SVs into a single inference framework. We demonstrate on simulated data that the method produces accurate tree inferences in the presence of all three variant types. We further demonstrate the method through application to real prostate tumor data, showing how our approach to coordinated phylogeny inference and clonal construction with all three variant types can reveal a more complicated clonal structure than is suggested by prior work consistent with extensive polyclonal seeding or migration.
Availabiliy: https://github.com/CMUSchwartzLab/TUSV-ext

1 Introduction

Cancer is widely understood as an evolutionary process (Nowell, 1976), in gradual accumulation of somatic genetic alterations occurs in parallel with selection for mutations promoting tumor growth or other aspects of disease progression. The evolutionary history is reflected in various kinds of genetic or epigenetic variations tumor cell lineages might accumulate (Davis et al., 2017), and the stochastic nature of the evolutionary process results in high genetic and epigenetic heterogeneity both between distinct tumors (inter-tumor heterogeneity) and between cell lineages in single tumors (intra-tumor heterogeneity). A vibrant field of "tumor phylogenetics" (Desper et al., 1999) has arisen with the goal of using such variations to reconstruct the history of a tumor and in turn, it is hoped, better understand the evolutionary landscapes of tumor cell lineages and how individual tumors navigate them.

The general idea of tumor phylogenetics has spawned numerous variations devoted to distinct kinds or combinations of genetic data or variation type (Schwartz and Schäffer, 2017). To date, most such methods have been developed to work with variants on bulk sequencing, in which one has mixed samples of genomic data from many tumor cells and must computationally deconvolve them to infer distinct cell lineages that can be resolved into a phylogeny (Schwartz and Shackney, 2010). Methods for deconvolutional phylogenetics of bulk genomic data have been developed predominantly for inference of genetic variation from single nucleotide (SNVs) data (Yuan et al., 2015; El-Kebir et al., 2015; Roth et al., 2014). It has become increasingly apparent that copy number alterations (CNAs) are of at least comparable importance to SNVs in tumor evolution, however, prompting increasing numbers of methods for CNAs (Oesper et al., 2013; Zaccaria et al., 2018; Zaccaria and Raphael, 2020) or combined SNVs and CNA data (Deshwar et al., 2015; El-Kebir et al., 2016; Li and Xie, 2015). More recently, it has also been shown to be possible to infer phylogenies incorporating structural variations (SVs) (Eaton et al., 2018) such as large deletions or translocations, which are more complicated to resolve phylogenetically but often key events in driving tumor evolution.

Attention in the field has increasingly turned to single-cell data (Navin et al., 2011; Gao et al., 2016), which has substantial advantages in resolving intratumor heterogeneity and bypassing the limited accuracy and resolution of computational deconvolution. However single cell sequencing also has substantial drawbacks, including technical artifacts such as allelic dropouts (Wang et al., 2012), with which single cell tumor phylogeny methods must contend (Zafar et al., 2019; Satas et al., 2020). Single cell data, particularly single-cell DNA-seq, remains costly to gather and thus all large tumor data resources are still based on bulk sequence. Some methods have been proposed to combine bulk and single-cell sequence or even other techniques to leverage their advantages together (Malikic et al., 2019; Lei et al., 2020, 2021; Fu et al., 2021), but data resources to date have not been designed to take advantage of such methods. Furthermore, single-cell data so far has limited utility for studying SVs as well as limited resolution for CNAs. As a result, deconvolutional methods remain important particularly for studies of variations other than SNVs.

To date, there has been no method that integrates all major variant types (SNVs, SVs, and CNAs) into a single inference. This is particularly problematic as these variants typically occur together and offer complementary advantages for understanding tumor evolution. SNVs are the simplest to handle phylogenetically and may be relatively numerous. CNAs are likely more important to understanding functional adaptation in cancers and can serve as a confounding factor for interpreting SNVs correctly, but they are more challenging to interpret phylogenetically, particularly in deconvolutional settings. SVs can also have profound effects on genome function and can confound interpretation of other variant types and may be particularly useful as phylogenetic markers because they are comparatively rare but this rarity also means they are poorly suited for phylogenetic inference on their own. So far, TUSV (Eaton et al., 2018) has been the only tumor phylogeny method capturing SVs, which it did in conjunction with CNAs.

In the present work, we develop a next generation TUSV-ext, the first tumor phylogeny software to accommodate SNVs, CNAs, and SVs within a single phylogenetic framework and resolve all three together as markers of tumor evolution. This is accomplished by extending the TUSV integer linear programming (ILP) framework with a model of SNVs implementing a model of Dollo parsimony, in which SNVs can be acquired once but potentially lost or duplicated multiple times through copy number change. We also extend the modeling of SVs so that existing breakpoints are also lost or duplicated through copy number changes. We demonstrate on simulated data that the method shows good performance in practice. We further apply it to a real prostate cancer dataset (Gundem et al., 2015), showing that it yields results consistent with prior findings and revealing of novel insight into how these variant types interact in evolution of a single tumor.

2 Method: Joint deconvolution and phylogenetics of SVs, CNAs and SNVs

2.1 Problem Statement

Our method takes in processed variant calls with paired breakpoint ends of different structural variants (SVs) and their estimated mean copy numbers, allelic mean copy number alterations (CNAs) of genome segments and estimated mean copy numbers of single nucleotide variants (SNVs) in vcf format. We encode these variations in three matrices, F, Q, and G, which contain mean variant copy numbers, positional relationships between breakpoints/SNVs and copy number segments, and pairing information for genomic breakpoints identifying which genome segments are involved in SVs for a set of genomic samples (details can be found in Table 1). Our goal is to deconvolve the mixed samples by minimizing an objective function:

$$\min_{U,C} |F - UC| + \lambda_1 R + \lambda_2 S, \tag{1}$$

similar to that of the original TUSV Eaton et al. (2018) to solve for a matrix of variant copy number profiles for each clone C and a matrix of clonal frequencies U with constraints from phylogenetic trees and relationships among different variants. The model uses two regularization terms to optimize for accurate deconvolution subject to penalties corresponding to consistency between variant types and an evolutionary distance measure used to favor minimum-evolution phylogenies.

Table 1: Essential parameters and variables.

Notation	Meaning
$F \in \mathbf{R}_{\geq 0}^{m\times(l+g+2r)}$	$f_{p,v}$ Average copy number of variant v in sample p
$Q \in \{0,1\}^{(l+g)\times r}$	$q_{b,s} = 1$ iff breakpoint or SNV b is in segment s, else 0
$G \in \{0,1\}^{l\times l}$	$g_{b_1,b_2} = 1$ iff breakpoint b_1 and b_2 are paired breakpoints deriving from the same SV
l	The number of breakpoints
g	The number of SNVs
r	The number of haploid copy number segments
n	The number of leaves in the phylogenetic tree
N	The number of total clones in the phylogenetic tree, $N = 2n - 1$
m	The number of samples
c_{max}	The maximum allowed copy number for each segment
$C \in \mathbf{Z}_{\geq 0}^{n\times(l+g+2r)}$	$c_{k,v}$ is the copy number of variants in clone k. When $v \in \{1,\ldots,l\}$, it represents the copy number of breakpoint v. When $v \in \{l+1,\ldots,l+g\}$, it represents the copy number of SNV v. When $v \in \{l+g+1,\ldots,l+g+r\}$ or $v \in \{l+g+r+1,\ldots,l+g+2r\}$, it represents the copy number of segment v from one allele or another
$U \in \mathbf{R}_{\geq 0}^{m\times k}$	$u_{p,k}$ is the frequency of clone k in sample p
$E \in \{0,1\}^{N\times N}$	$e_{i,j} = 1$ iff there is a directed edge from clone i to clone j in the phylogenetic tree, else 0
$A \in \{0,1\}^{N\times N}$	$a_{i,j} = 1$ iff clone i is an ancestor of clone j in the phylogenetic tree, else 0
$W \in \{0,1\}^{N\times N\times(l+g)}$	$w_{i,j,b}$ is a 0-1 binary indicator, 1 iff breakpoint or SNV b occur at edge from node i to j else 0
$D \in \{0,1\}^{(l+g)\times 2}$	$d_b = 1$ indicates that the breakpoint b is in the first allele and $d_b = 0$ indicates the second allele
$\Gamma \in Z_{\geq 0}^{N\times(l+g)\times 2}$	$\gamma_{i,b,0}$ is the segment copy number of node i at the position where breakpoint b locate in the first allele and $\gamma_{i,b,1}$ the second allele
$\Psi \in Z_{\geq 0}^{m\times(l+g)}$	$\psi_{p,b}$ is the mixed copy number of segment in sample p for both allele that contains the breakpoint or SNV b

2.2 Algorithm: Coordinate descent

We used coordinate descent to solve for matrix U, C based on given input matrix F, Q and G, as is done in Zaccaria et al. (2018); Eaton et al. (2018). We first initialize U matrix randomly under the constraint that $\sum_{k=1}^{N} u_{p,k} = 1, \forall p \in \{1 \dots m\}$ and solve for C with fixed U. Based on the solution C, we then fix C and solve for U. The process is done iteratively until convergence or maximum number of iterations are reached. Random starts can be performed to avoid local optimal solution. Each step of the coordinate descent algorithm is posed in terms of an ILP optimization problem as described in more detail in the subsequent sections.

2.3 Algorithm: Estimating U

As described in Eaton et al. (2018), we construct an ILP by introducing intermediate variable $f_{\delta,p,v}$, which describe the element-wise absolute distance between F and UC. U can be solved by minimizing only the first term of the objective function $|F - UC|$, represented by the sum of $f_{\delta,p,v}$, under the constraint that the sum of the frequencies of different clones in every sample is equal to one.

$$f_{\delta,p,v} \geq f_{p,v} - \sum_{k=1}^{N} u_{p,k} c_{k,v}, \forall p \in \{1, \dots m\}, v \in \{1, \dots, l+g+2r\} \tag{2}$$

$$f_{\delta,p,v} \geq -f_{p,v} + \sum_{k=1}^{N} u_{p,k} c_{k,v}, \forall p \in \{1, \dots m\}, v \in \{1, \dots, l+g+2r\} \tag{3}$$

$$|F - UC| = \sum_{p=1}^{m} \sum_{v=1}^{l+g+2r} f_{\delta,p,v} \tag{4}$$

$$\sum_{k=1}^{N} u_{p,k} = 1, \forall p \in \{1 \dots m\} \tag{5}$$

2.4 Algorithm: Estimating C

We then fix U and solve for C given the input matrix F, G and Q. Because C encodes the variant copy number profile of each subclone, it is related to the underlying evolutionary tree structure and can be used to constrain possible trees as described in the following subsections.

2.4.1 Phylogenetic constraints

As described in Eaton et al. (2018), in order to relate the copy number profiles to the phylogenetic tree T, we define an edge matrix $E^{N \times N}$ to describe the parent-child relationship in the tree where $e_{i,j} = 1$ indicates clone i is the direct parent of clone j. We assume the phylogenetic tree T is a binary tree without loss of generosity.

For the root node, there is no incoming edge (Eq.6). For all the other nodes except for root node, there is only one incoming edge for each node (Eq.7). For all leaf nodes, there is no outgoing edge (Eq.8), and for all the other nodes except for leaf nodes, there are two outgoing edges (Eq.9). We define the node $N = 2n - 1$ as root node, node 1 to n as leaf nodes, and node $n + 1$ to $N - 1$ as internal nodes other than root. Then we can formulate constraints as follows:

$$e_{i,N} = 0, \forall i \in \{1, \dots, N\} \tag{6}$$

$$\sum_{i=1}^{N} e_{i,j} = 1, \forall j \in \{1, \dots, N-1\} \tag{7}$$

$$e_{i,j} = 0, \forall i \in \{1, \dots, n\}, j \in \{1, \dots, N\} \tag{8}$$

$$\sum_{j=1}^{N} e_{i,j} = 2, \forall i \in \{n+1, \dots, N\} \tag{9}$$

We added two constraints to ensure that there is no cycle in the tree. We define an ancestor matrix to describe the ancestor-descendent relationship in the tree where $a_{i,j} = 1$ indicates clone i is an ancestor of clone j. Similarly, the ancestor matrix must specify that the root node has no ancestor and is an ancestor for all other nodes, and that leaf nodes have no descendants (Eq. 10-11). The ancestor matrix should also be consistent with the edge matrix. For instance, any parent should be its child's ancestor (Eq. 12), and any ancestor of a parent should be inherited by its child (Eq. 13-14). Those constraints can be added to the ILP as follows

$$a_{N,j} = 1, \forall j \in \{1, \ldots, N-1\} \tag{10}$$
$$a_{i,N} = 0, \forall i \in \{1, \ldots, N\} \tag{11}$$
$$a_{i,j} \geq e_{i,j}, \forall i, j \in \{1, \ldots, N\} \tag{12}$$
$$a_{h,j} \geq e_{i,j} + a_{j,i} - 1, \forall i, j \in \{1, \ldots, N\}, g \in \{1, \ldots, i-1, i+1, \ldots, N\} \tag{13}$$
$$a_{h,j} \geq 1 - e_{i,j} + a_{j,i}, \forall i, j \in \{1, \ldots, N\}, g \in \{1, \ldots, i-1, i+1, \ldots, N\} \tag{14}$$

Beyond the common constraints used in Eaton et al. (2018), the structure of a tree should also contain no cycles, which can be described by saying that for any two node i and j, either i is the ancestor of j or j is the ancestor of i. Also a node should not be its own ancestor or descendent:

$$a_{i,j} + a_{j,i} <= 1, \forall i, j \in \{1, \ldots, N\} \tag{15}$$
$$a_{i,i} = 0, \forall i \in \{1, \ldots, N\} \tag{16}$$

2.4.2 Subclone variant copy number constraints

We model the tree as describing somatic variations accumulating after descent from the root node and so assume that the root node has 0 breakpoints (Eq. 18) or SNVs and has a single copy of each allele in each genomic segment (Eq. 19). For algorithmic convenience, we also restrict copy numbers to a set maximum value c_{max} (Eq. 17).

$$c_{k,s} \leq c_{max}, \forall k \in \{1, \ldots, N\}, s \in \{1 \ldots, l+g+2r\} \tag{17}$$
$$c_{N,b} = 0, \forall b \in \{1, \ldots l+g\} \tag{18}$$
$$c_{N,l+g+s} = 1, \forall s \in \{1, \ldots 2r\} \tag{19}$$

In contrast to the original TUSV, we use two temporary variables $x^1_{i,j,s}$ and $x^2_{i,j,s}$ corresponding to two original alleles, instead of one variable $x_{i,j,s}$, to define the absolute changes in copy numbers of segment s on edge (v_i, v_j). We add these two copy numbers together to form a copy number evolutionary distance used in regularization term R.

$$0 \leq x^1_{i,j,s} \leq c_{max} e_{i,j}, s \in \{1, \ldots, r\} \tag{20}$$
$$x^1_{i,j,s} \geq c_{i,s+l+g} - c_{j,s+l+g} - c_{max}(1 - e_{i,j}) \tag{21}$$
$$x^1_{i,j,s} \geq -c_{i,s+l+g} + c_{j,s+l+g} - c_{max}(1 - e_{i,j}) \tag{22}$$
$$0 \leq x^2_{i,j,s} \leq c_{max} e_{i,j} \tag{23}$$
$$x^2_{i,j,s} \geq c_{i,s+l+g+r} - c_{j,s+l+g+r} - c_{max}(1 - e_{i,j}) \tag{24}$$
$$x^2_{i,j,s} \geq -c_{i,s+l+g+r} + c_{j,s+l+g+r} - c_{max}(1 - e_{i,j}) \tag{25}$$
$$\rho_{i,j} = \sum_{s=1}^{r} x^1_{i,j,s} + \sum_{s=1}^{r} x^2_{i,j,s} \tag{26}$$
$$R = \sum_{i=1}^{N} \sum_{j=1}^{N} \rho_{i,j} \tag{27}$$

2.4.3 Perfect phylogeny on appearance of breakpoints

We constrain paired breakpoints, so that they must occur in the same edge in the tree same as Eaton et al. (2018). We also impose a Dollo phylogeny constraint on SV breakpoints as described in Eaton et al. (2018)

ensuring that each variant can only appear once although it may be lost multiple times. We treat SNVs as a special case of breakpoints without pairing information and make the same Dollo assumption (Eq. 30). The assumption of no homoplasy is more dubious for SNVs than for SVs and we recognize that it may be a source of error that we permit out of algorithmic necessity. Therefore we define $W \in \{0,1\}^{N\times N\times(l+g)}$ to describe the edge on which each breakpoint or SNV occurs, where $w_{i,j,b} = 1$ if breakpoint or SNV b occur at edge from node i to j and 0 otherwise (Eq. 28-29).

In addition, we add constraint that every loss of breakpoint or SNV should co-occur with a allelic-specific segment copy number loss for the segment in which the breakpoint or SNV lies. Similarly, we also constrain breakpoints or SNVs to be duplicated only if their corresponding genome segment is duplicated.

We define a variable X to help define W where

$$x_{i,j,b} = 2 + \bar{c}_{i,b} - \bar{c}_{j,b} - e_{i,j} \tag{28}$$

$$w_{i,j,b} = 1 - \bar{x}_{i,j,b}, \forall i,j \in \{1,\dots,N\}, b \in \{1,\dots,l+g\} \tag{29}$$

$$\sum_{i=1}^{N}\sum_{j=1}^{N} w_{i,j,b} = 1, \forall b \in \{1,\dots,l+g\} \tag{30}$$

We further define a binarization operator as follows

$$\bar{x} = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \end{cases}$$

which can be linearly expressed by an additional variable y_b where

$$x = \sum_{u=0}^{\lfloor \log_2 x_{max} \rfloor + 1} 2^u y_u \tag{31}$$

$$0 \le y_u \le \bar{x} \le \sum_{v=0}^{\lfloor \log_2 x_{max} \rfloor + 1} y_v, \forall u \in \{0,\dots,\lfloor \log_2 x_{max} \rfloor + 1\} \tag{32}$$

as described in Zaccaria et al. (2018); Eaton et al. (2018).

In order to determine the co-occurence of the variants and corresponding allele, we introduced a new binary variable matrix $D \in \{0,1\}^{(l+g)\times 2}$ where $d_b = 1$ indicates that the breakpoint b is in the first allele and $d_b = 0$ indicates the second allele. During preprocessing, we compare the copy numbers of each allele and order alleles so that the major copy number is in the first allele and minor copy number is in the second allele. We further make an assumption that the major segment in all samples belong to the same allele. We therefore define matrix $\Gamma \in Z_{\ge 0}^{N\times(l+g)\times 2}$ where $\gamma_{i,b,0}$ is the segment copy number of node i at the position where breakpoint b locate in the first allele and $\gamma_{i,b,1}$ the second allele. We want to constrain that for any existing breakpoint in the clone (not newly developed breakpoint in this branch), the breakpoint number change should be smaller or equal to the corresponding copy number change at the breakpoint position.

$$\gamma_{j,b,0} - \gamma_{i,b,0} \ge c_{j,b} - c_{i,b} - (2 - e_{i,j} - d_b + w_{i,j,b})(2c_{max} + 1), \tag{33}$$

$$\gamma_{j,b,0} - \gamma_{i,b,0} \le c_{j,b} - c_{i,b} + (2 - e_{i,j} - d_b + w_{i,j,b})(2c_{max} + 2) \tag{34}$$

$$\gamma_{j,b,1} - \gamma_{i,b,1} \ge c_{j,b} - c_{i,b} - (1 - e_{i,j} + d_b + w_{i,j,b})(2c_{max} + 1), \tag{35}$$

$$\gamma_{j,b,1} - \gamma_{i,b,1} \le c_{j,b} - c_{i,b} + (1 - e_{i,j} + d_b + w_{i,j,b})(2c_{max} + 2) \tag{36}$$

$$\forall i,j \in \{1,\dots,N\}, b \in \{1,\dots,l+g\} \tag{37}$$

$$\tag{38}$$

2.4.4 Structural variant and segment consistency

The copy number of a breakpoint or SNV for each clone should not exceed the copy number of the segment in which it is found. However, in order to integrate the allelic information, we only constrain the copy number

of a breakpoint to be below the copy number of the segment of the corresponding allele.

$$c_{k,b} \leq \gamma_{k,b,0} + (1 - d_b)c_{max} = \sum_{s=1}^{r} q_{b,s} c_{k,s+l+g} + (1 - d_b)c_{max} \tag{39}$$

$$c_{k,b} \leq \gamma_{k,b,1} + d_b c_{max} = \sum_{s=1}^{r} q_{b,s} c_{k,s+l+g+r} + d_b c_{max} \tag{40}$$

For mixed sample copy numbers, we define $\Psi \in Z_{\geq 0}^{m \times (l+g)}$ so that $\psi_{p,b}$ to be the mixed copy number of segment for both alleles in sample p that contains the breakpoint or SNV b. We can calculate the ratio of mixed copy number of breakpoint or SNV and mixed copy number of segments as $\pi_{p,b}$ (Eq.41-42, also known as variant allele frequency VAF). We want to minimize the distance of the estimated ratio and the actual ratio (Eq.43) by adding a regularization term which sums up the absolute differences between the two ratios for sample p and variant b (breakpoint or SNV) among all samples and variants, represented as temporary variable $z_{p,b}$ (Eq.44-46).

$$\psi_{p,b} = \sum_{s=1}^{r} q_{b,s}(f_{p,s+l+g} + f_{p,s+l+g+r}) \tag{41}$$

$$\pi_{p,b} = \frac{f_{p,b}}{\psi_{p,b}} \tag{42}$$

$$\left| \pi_{p,b} - \frac{\sum_{k=1}^{N} u_{p,k} c_{k,b}}{\sum_{k=1}^{N} u_{p,k}(\gamma_{k,b,0} + \gamma_{k,b,1})} \right| \tag{43}$$

$$\iff z_{p,b} \geq \pi_{p,b} \sum_{k=1}^{N} u_{p,k}(\gamma_{k,b,0} + \gamma_{k,b,1}) - \sum_{k=1}^{N} u_{p,k} c_{k,b} \tag{44}$$

$$z_{p,b} \geq -\pi_{p,b} \sum_{k=1}^{N} u_{p,k}(\gamma_{k,b,0} + \gamma_{k,b,1}) + \sum_{k=1}^{N} u_{p,k} c_{k,b}, \forall b \in \{1, \ldots, l+g\} \tag{45}$$

$$S = \sum_{p=1}^{m} \sum_{b=1}^{l+g} z_{p,b} \tag{46}$$

3 Method: Mapping unsampled SNVs or breakpoints to identified nodes

The full algorithm in principle provides a problem formulation for which we can optimize for any data set. However, the ILP framework is limited in its ability to scale to large numbers of variants, as may occur in highly mutable cancers. We therefore consider a variation on the method in which we subsample a full set of variants for purposed of phylogeny inference and then map the remaining variants to identified subclonal populations derived from the subsample.

3.1 Problem Statement

Input: Mixed sample copy numbers matrix $\hat{F}^{m \times \hat{g}}$, Tree $\mathcal{T}$ (presented as ancestry matrix $A^{n \times n}$ where $A_{i,j} = 1$ if node i is the ancestor of node j and edge matrix $E^{n \times n}$ where $E_{i,j} = 1$ if node i is the direct parent of node j), mapping matrix $\hat{Q}^{\hat{g} \times r}$ which maps the unsample SNVs to segments, frequency matrix $U^{m \times n}$ and component matrix $C^{n \times (l+g+2r)}$ including g sampled SNVs.
Output: An assignment list $AssignList^{1 \times \hat{g}}$ which shows the optimal assignment of nodes for $\hat{g}$ unsampled

SNVs (or breakpoints).

$$
\begin{aligned}
\hat{C} =& C_{:,(l+g+d_j r):(l+g+(d_j+1)r)}\hat{Q}^T \\
\check{C} =& C_{:,(l+g+(1-d_j)r):(l+g+(2-d_j)r)}\hat{Q}^T \\
\hat{C}^{parent} =& E^\top \hat{C} \\
\bar{F}^{i,j,1} =& \begin{cases} U_{:,i} * \hat{C}_{i,j} + UA_{i,:} * \hat{C}_{:,j} & \text{if } \hat{C}_{i,j} = 1 \text{ or } \hat{C}_{i,j} - \hat{C}_{i,j}^{parent} > 1 \\ \infty & \text{otherwise} \end{cases} \\
\bar{F}^{i,j,2} =& \begin{cases} U_{:,i} * \hat{C}_{i,j} + UA_{i,:} * \hat{C}_{:,j}/\hat{C}_{i,j} & \text{if } \hat{C}_{i,j} > 1 \\ \infty & \text{otherwise} \end{cases} \\
d_j^{opt}, i_j^{opt} =& \underset{d_j \in \{0,1\}, i \in \{1,\ldots,n\}}{\operatorname{argmin}} \min\{\bar{F}^{i,j,1}, \bar{F}^{i,j,2}\} \\
AssignList_j =& i_j^{opt}
\end{aligned}
$$

3.2 Algorithm

Since the number of clones is generally small, we assign unsampled variants to clones by enumerating all possible clones to which a given variant can belong, choosing an assignment that is most parsimonious with respect to mutation of the mapped variant. We used factorization to speed up the algorithm for all variants which obey the conditions.

4 Results

4.1 Validation on simulated data

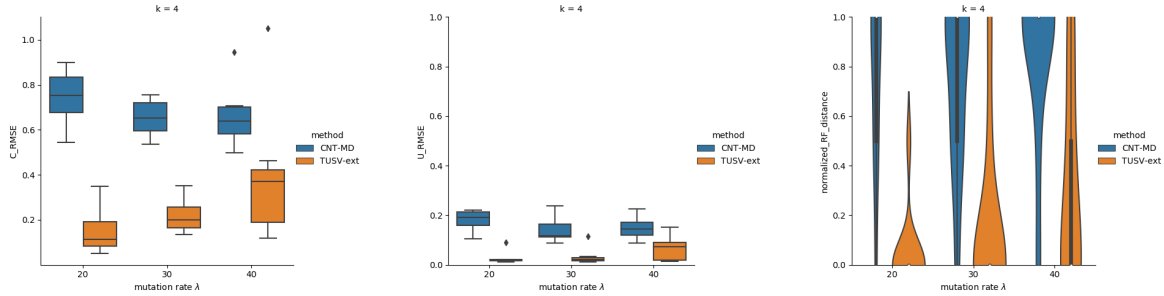
We consider a series of experiments to test robustness of the method to data parameter variants, simulating 10 cases for each experiment. We simulate with different structural mutation rate λ^{total}, which is the parameter of a Poisson distribution on the overall structural variants number. We simulate allele-specific duplications from 2 to 6, deletion, translocation and inversion with relative probabilities of 2:2:1:1 to ensure a relatively high rate of copy number changes. The SNV mutation rate is set to $100\lambda^{total}$. For each simulated cancer, we also simulate either 1, 5, or 10 samples. We simulate the frequencies of each clone for each sample randomly under the frequency constraint and simulate final mixed samples with the weighted sum of the profiles of each clone. We set a maximum run time of 1000 seconds for each iteration and with maximum of 10 iterations except where specially mentioned below. Each case was run with 5 random initializations to avoid local optima.

Since there is no other method to our knowledge that integrates all three types of variants, direct comparison to any one alternative methods is challenging. We therefore split validation into two tasks — one assessing performance on CNAs and one on how CNVs contribute to SVs and SNVs reconstruction — through which we can compare to other work in the literature.

4.1.1 Task 1: Validation on segment copy numbers, subclonal frequencies and tree estimation

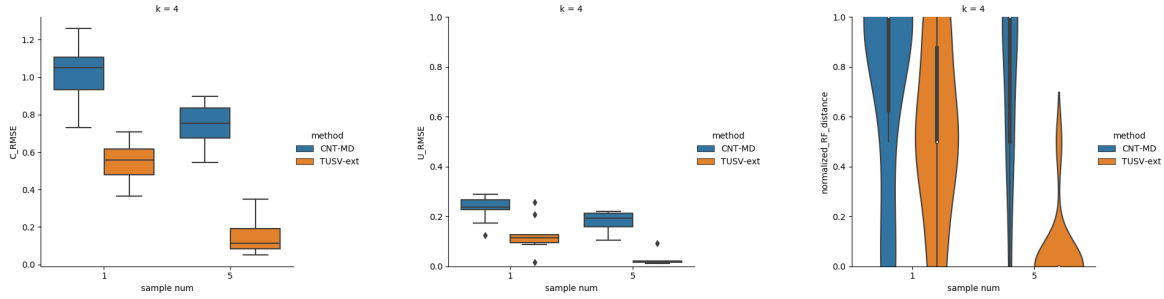
We assume that the clone number is known and compare our method with CNT-MD (Zaccaria et al., 2018). Since CNT-MD also uses linear programming for optimization but only utilizes segmental copy numbers, we only compare the estimated copy numbers and clonal frequencies. For fair comparison with CNT-MD, which assumed in our simulations that only leaf nodes and normal clones exist in the mixed samples. We set our method to be leaf-only mode to match the assumptions of CNT-MD.

We tested on different structural mutation rates from 20 to 40. The results show that our method can converge faster with the assistance of other variants like SVs and SNVs, showing higher accuracy in estimating both C and U matrices, as well as tree estimation. The advantage is larger when the mutation rate is small, which we expect may be more of an issue for CNT-MD because fewer mutations would leave it comparatively less signal from which to deconvolve.



(a) (b) (c)

Figure 1: Results on task 1 with varying mutation rates. (a) The root mean square error (RMSE) of estimate C. (b) The root mean square error (RMSE) of estimate U. (c) Normalized Roubinson-Foulds distance of estimated trees and true tree



(a) (b) (c)

Figure 2: Results on task 1 with varying sample numbers. (a) The root mean square error (RMSE) of estimate C. (b) The root mean square error (RMSE) of estimate U. (c) Normalized Roubinson-Foulds distance of estimated trees and true tree

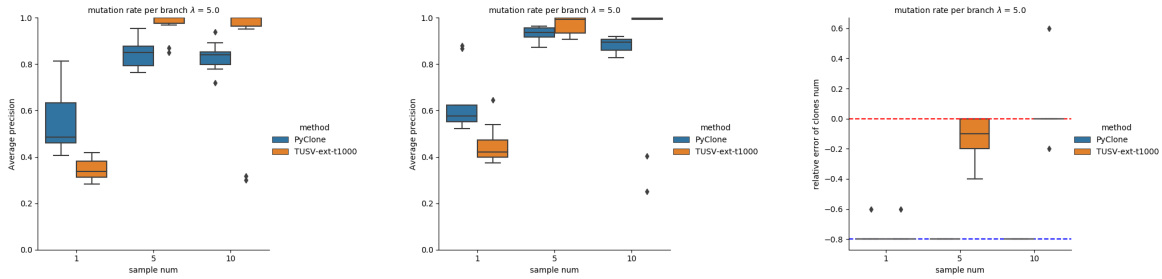
Both methods have relatively poor performance for sample size 1, although CNT-MD yields superior accuracy in this case. For 5 or 10 samples, both methods improve substantially although TUSV-ext yields notably higher accuracy than CNT-MD.

4.1.2 Task 2: Validation on SVs, SNVs and number of clones

In task 2 we validate the deconvolution of SNVs and SVs as well as determination of number of clones. We use the average precision of the co-clustering matrix to validate the deconvolution performance of both breakpoints and SNVs. To determine the correct number of clones, we collapse nodes with 0 frequency when they only have one child and child nodes with 0 branch length before counting the final number of uncollapsed clones. We use the relative distance $(k_{true} - k_{estimate})/k_{true}$ as evaluation metric. We compare our method with PyClone (Roth et al., 2014), which is a popular method for inferring subclonal population from single or multiple samples. We examined the performance among different sample numbers and clone numbers with a fixed mutation rate per branch.

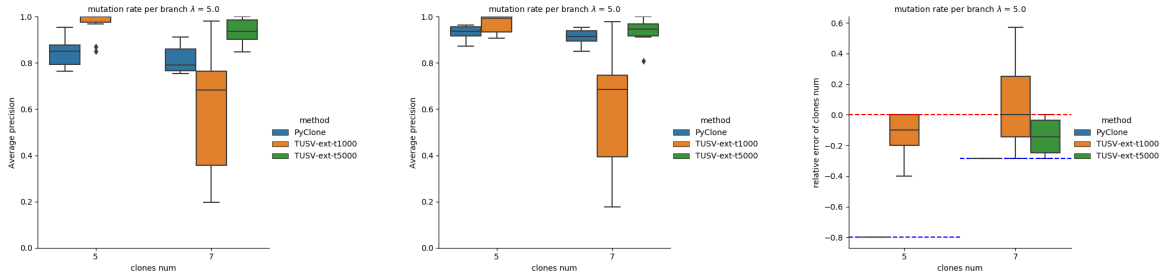
PyClone performs better than TUSV-ext when the sample number is small. When the sample number is 5 or 10, TUSV-ext shows better result in terms of the SNVs and breakpoints. For the clone number determination, PyClone is prone to overestimate the clone number, whereas our method is able to identify the clone number more accurately.

For larger clone numbers, we found that TUSV-ext performance notably degraded. We suspected that this was caused by a failure to converge adequately due to the larger number of variants. We therefore increased maximum run time per iteration to 5000 for this case, which resulted in improved performance relative to Pyclone.



(a) (b) (c)

Figure 3: Results on task 2 with different samples size (a) Average precision of breakpoints. (b) Average precision of SNVs. (c) Accuracy of determining number of clones



(a) (b) (c)

Figure 4: Results on task 2 with different clone numbers (or different total mutation rate) when the mutation rate per branch remains $\lambda = 5$ (a) Average precision of breakpoints. (b) Average precision of SNVs. (c) Accuracy of determining number of clones

4.2 Application to real prostate cancer data

Gundem et al. (2015) performed whole genome sequencing on 51 tumor samples from 10 patients and provided the variant calls of SVs, CNAs, SNVs and Indels. We selected one case from this study, patient A32, as a demonstration of our method on real data. For this patient, the study included samples from five tumor regions: A from right rib metastasis, C from primary tumor from prostate, D from left cervical lymph node metastasis, E from left subclavicular lymph node metastasis and F from right humerus metastasis.

Since there are many more rearrangements than SNVs, we subsampled both SVs and SNVs for eight runs (approximately 80 breakpoints and 40 SNVs for each run) and mapped the unsampled variants to nodes in the phylogenetic tree. For each run, we set the maximal node count to be 9 and set 10 iterations with time limit of 4000 seconds per iteration. Each run was performed with 2 random initializations. We also subsampled the copy number segments to only retain those with breakpoints or SNVs lying inside them to reduce the time complexity. We used the two most similar runs, according to the CASet and DISC metric (DInardo et al., 2020), to identify a consensus solution used for subsequent analysis.

We compare the tree inference (Fig. 5(a)) with that of Gundem et al. (2015). Both trees show branching evolution, yet with distinct trajectories. Clone 0 is identified in our method only exhibits in samples C and E with relatively low frequencies, while no comparable separate branch is identified in the prior work. Mutations of TP53 (K132N) and TBLIXR (essential splice) and chr8p deletion are inferred as early events in both inferences. Clone 6 was found to contain CTNNB1 substitution. Clones 5 and 7 were estimated to acquire PTEN and CDKN1B loss respectively, which is different from the estimation in prior work where both PTEN and CDKN1B loss of heterozygosity were estimated in early stages of clonal evolution.

We further examined estimated frequencies of inferred cell populations in each sample (Fig.5(b)). The primary tumor (sample C) shows notably higher clonal heterogeneity than the metastases, as we would expect, with evidence for most clones having been present in the primary tumor rather than evolving *de*

novo in the metastases. Furthermore, samples A and E are inferred to have small proportions of normal cells (clone 8) while samples C, D and F have no normal cell population. Internal nodes (clone 5, 6 and 7) exhibit low frequencies in all samples. Most leaf nodes (clone 0-5) exhibit higher frequencies in samples C and E and lower in samples A, D and F, or vice versa. Clone 0, the closest child to the diploid normal clone, is not inferred in Gundem et al. (2015). Only primary sample C and metastasis E contain this clone and it is expanded in sample C. Clone 4 shows higher proportion in samples C and E than A, D and E while clones 2 and 3 show the opposite case, perhaps suggestive of distinct polyclonal seeding events. Our model overall suggests a more complex pattern of polyclonal origins, whether through polyclonal seeding or subsequent cell migration, than is suggested in the prior work.

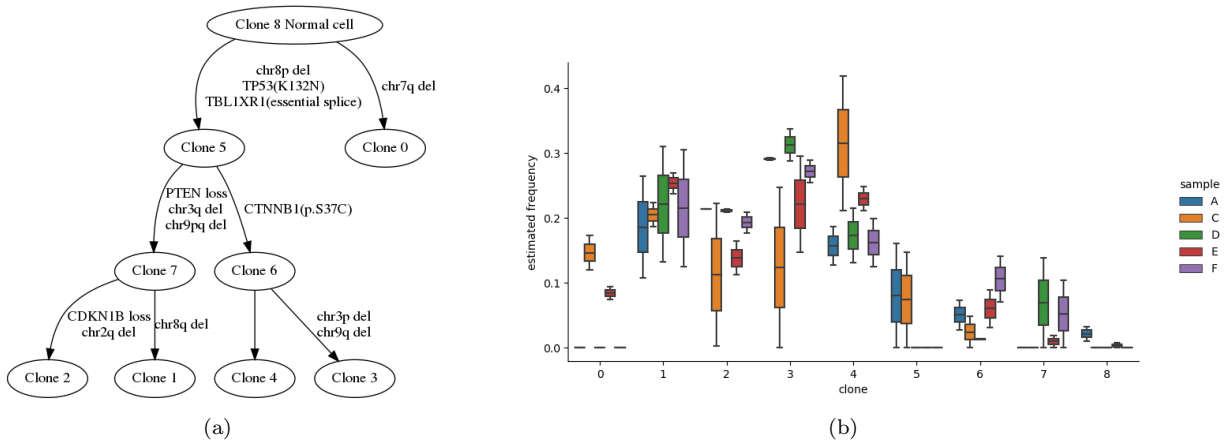


Figure 5: Results for prostate cancer A32 patient data. (a) Inferred consensus tree structure with a subset of variants/mutations (b) Comparison of proportions of all 9 clones across samples.

5 Conclusion

In this work, we develop a tumor phylogeny method, TUSV-ext, to incorporate SNVs, CNAs, and SVs into a single clonal lineage tree reconstruction. We show on simulated data how these variant types can be complementary, yielding superior accuracy to competing methods that make use of only subsets of these variants. We further demonstrate effectiveness of the method on a real prostate cancer data set involving primary and metastatic samples from a single patient. One of the main challenges to the inference is scalability, as the method so far can accommodate relatively limited numbers of variants and potential requires long convergence time. While we offer a heuristic solution for variant subsampling and subsequent mapping, the method would benefit from more principled solutions for truly handling large variant sets in a single reconstruction. Furthermore, even with more accurate clonal reconstruction and assignment, some aspects of tumor evolution are not fully resolvable from phylogenetics alone and extending alternative approaches that incorporate clonal migration (El-Kebir et al., 2018) to similar variant combinations might be necessary to supplement a purely phylogenetic approach such as presented here.

6 Acknowledgements

We thank the AWS ML Research Awards program for providing cloud computing support for this work. Portions of this work have been funded by US NIH award R21CA216452 and Pennsylvania Dept. of Health award FP00003273. Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award number R01HG010589. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The Pennsylvania Department of Health specifically disclaims responsibility for any analyses, interpretations or conclusions.

References

A. Davis, R. Gao, and N. Navin. Tumor evolution: Linear, branching, neutral or punctuated? *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1867(2):151–161, 4 2017. ISSN 0304419X. doi: 10.1016/j.bbcan.2017.01.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0304419X17300197.

A. G. Deshwar, S. Vembu, C. K. Yung, G. H. Jang, L. Stein, and Q. Morris. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 2015. ISSN 1474760X. doi: 10.1186/s13059-015-0602-8.

R. Desper, F. Jiang, O. P. Kallioniemi, H. Moch, C. H. Papadimitriou, and A. A. Schäffer. Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of Computational Biology*, 6(1):37–51, 1999. ISSN 10665277. doi: 10.1089/cmb.1999.6.37. URL www.liebertpub.com.

Z. DInardo, K. Tomlinson, A. Ritz, and L. Oesper. Distance measures for tumor evolutionary trees. *Bioinformatics*, 36(7):2090–2097, 4 2020. ISSN 14602059. doi: 10.1093/bioinformatics/btz869. URL https://bitbucket.org/.

J. Eaton, J. Wang, and R. Schwartz. Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics*, 34(13):i357–i365, 7 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty270. URL https://academic.oup.com/bioinformatics/article/34/13/i357/5045780.

M. El-Kebir, L. Oesper, H. Acheson-Field, and B. J. Raphael. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70, jun 2015. ISSN 14602059. doi: 10.1093/bioinformatics/btv261. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv261.

M. El-Kebir, G. Satas, L. Oesper, and B. J. Raphael. Inferring the Mutational History of a Tumor Using Multi-state Perfect Phylogeny Mixtures. *Cell Systems*, 3(1):43–53, 7 2016. ISSN 24054712. doi: 10.1016/j.cels.2016.07.004. URL https://linkinghub.elsevier.com/retrieve/pii/S2405471216302216.

M. El-Kebir, G. Satas, and B. J. Raphael. Inferring parsimonious migration histories for metastatic cancers. *Nature Genetics*, 50(5):718–726, apr 2018. ISSN 15461718. doi: 10.1038/s41588-018-0106-z. URL https://www.nature.com/articles/s41588-018-0106-z.

X. Fu, H. Lei, Y. Tao, K. Heselmeyer-Haddad, I. Torres, M. Dean, T. Ried, and R. Schwartz. Joint clustering of single-cell sequencing and fluorescence in situ hybridization data for reconstructing clonal heterogeneity in cancers. *Journal of Computational Biology*, 2021.

R. Gao, A. Davis, T. O. McDonald, E. Sei, X. Shi, Y. Wang, P.-C. Tsai, A. Casasent, J. Waters, H. Zhang, F. Meric-Bernstam, F. Michor, and N. E. Navin. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nature Genetics*, 48(10):1119–1130, 10 2016. ISSN 1061-4036. doi: 10.1038/ng.3641. URL http://www.nature.com/articles/ng.3641.

G. Gundem, P. Van Loo, B. Kremeyer, L. B. Alexandrov, J. M. Tubio, E. Papaemmanuil, D. S. Brewer, H. M. Kallio, G. Högnäs, M. Annala, K. Kivinummi, V. Goody, C. Latimer, S. O'Meara, K. J. Dawson, W. Isaacs, M. R. Emmert-Buck, M. Nykter, C. Foster, Z. Kote-Jarai, D. Easton, H. C. Whitaker, D. E. Neal, C. S. Cooper, R. A. Eeles, T. Visakorpi, P. J. Campbell, U. McDermott, D. C. Wedge, and G. S. Bova. The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547):353–357, 4 2015. ISSN 14764687. doi: 10.1038/nature14347. URL https://www.nature.com/articles/nature14347.

H. Lei, B. Lyu, E. M. Gertz, A. A. Schäffer, X. Shi, K. Wu, G. Li, L. Xu, Y. Hou, M. Dean, and R. Schwartz. Tumor Copy Number Deconvolution Integrating Bulk and Single-Cell Sequencing Data. *Journal of Computational Biology*, 27(4):565–598, 4 2020. ISSN 1557-8666. doi: 10.1089/cmb.2019.0302. URL https://www.liebertpub.com/doi/10.1089/cmb.2019.0302.

H. Lei, E. M. Gertz, A. A. Schäffer, X. Fu, Y. Tao, K. Heselmeyer-Haddad, I. Torres, G. Li, L. Xu, Y. Hou, K. Wu, X. Shi, M. Dean, T. Ried, and R. Schwartz. Tumor heterogeneity assessed by sequencing and fluorescence in situ hybridization (FISH) data. *Bioinformatics*, 07 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab504. URL https://doi.org/10.1093/bioinformatics/btab504. btab504.

Y. Li and X. Xie. MixClone: A mixture model for inferring tumor subclonal populations. *BMC Genomics*, 16(S2):S1, 1 2015. ISSN 14712164. doi: 10.1186/1471-2164-16-S2-S1. URL https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-16-S2-S1.

S. Malikic, F. R. Mehrabadi, S. Ciccolella, M. K. Rahman, C. Ricketts, E. Haghshenas, D. Seidman, F. Hach, I. Hajirasouliha, and S. C. Sahinalp. PhISCS: A combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. *Genome Research*, 29(11): 1860–1877, 2019. ISSN 15495469. doi: 10.1101/gr.234435.118.

N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W. R. McCombie, J. Hicks, and M. Wigler. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94, 4 2011. ISSN 0028-0836. doi: 10.1038/nature09807. URL http://www.nature.com/articles/nature09807.

P. Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 10 1976. ISSN 0036-8075. doi: 10.1126/science.959840. URL http://www.sciencemag.org/cgi/doi/10.1126/science.959840.

L. Oesper, A. Mahmoody, and B. J. Raphael. THetA: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biology*, 14(7), 2013. ISSN 1474760X. doi: 10.1186/gb-2013-14-7-r80. URL http://compbio.cs.brown.edu/software/.

A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté, and S. P. Shah. PyClone: Statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4): 396–398, 2014. ISSN 15487105. doi: 10.1038/nmeth.2883.

G. Satas, S. Zaccaria, G. Mon, and B. J. Raphael. SCARLET: Single-Cell Tumor Phylogeny Inference with Copy-Number Constrained Mutation Losses. *Cell Systems*, 10(4):323–332, 2020. ISSN 24054720. doi: 10.1016/j.cels.2020.04.001. URL https://doi.org/10.1016/j.cels.2020.04.001.

R. Schwartz and A. A. Schäffer. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, 18(4):213–229, 4 2017. ISSN 1471-0056. doi: 10.1038/nrg.2016.170. URL http://www.nature.com/articles/nrg.2016.170.

R. Schwartz and S. E. Shackney. Applying unmixing to gene expression data for tumor phylogeny inference. *BMC bioinformatics*, 11(1):1–20, 2010.

C. Wang, K. B. Schroeder, and N. A. Rosenberg. A maximum-likelihood method to correct for allelic dropout in microsatellite data with no replicate genotypes. *Genetics*, 2012. ISSN 00166731. doi: 10.1534/genetics.112.139519.

K. Yuan, T. Sakoparnig, F. Markowetz, and N. Beerenwinkel. BitPhylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biology 2015 16:1*, 16(1):1–16, 2 2015. ISSN 1465-6906. doi: 10.1186/S13059-015-0592-6. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0592-6.

S. Zaccaria and B. J. Raphael. Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. *Nature Communications*, 11(1):1–13, 12 2020. ISSN 20411723. doi: 10.1038/s41467-020-17967-y. URL https://doi.org/10.1038/s41467-020-17967-y.

S. Zaccaria, M. El-Kebir, G. W. Klau, and B. J. Raphael. Phylogenetic Copy-Number Factorization of Multiple Tumor Samples. *Journal of Computational Biology*, 25(7):689–708, 7 2018. ISSN 1557-8666. doi: 10.1089/cmb.2017.0253. URL http://www.liebertpub.com/doi/10.1089/cmb.2017.0253.

H. Zafar, N. Navin, K. Chen, and L. Nakhleh. SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome Research*, 29 (11):1847–1859, 11 2019. ISSN 1088-9051. doi: 10.1101/gr.243121.118. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.243121.118.