Journal Title Here, 2022, pp. 1–9

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Deconvolution and Phylogeny Inference of Diverse Variant Types Integrating Bulk DNA-seq with Single-cell RNA-seq

Nishat Anjum Bristy[1,*] and Russell Schwartz[1,2,*]

[1]Ray and Stephanie Lane Computational Biology Department, Carnegie Mellon University, Pittsburgh, 15213, PA, USA and [2]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, 15213, PA, USA

*Corresponding authors. nbristy@andrew.cmu.edu, russells@andrew.cmu.edu

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Motivation: Reconstructing clonal lineage trees ("tumor phylogenetics") has become a core tool of cancer genomics. Earlier approaches based on bulk DNA sequencing (DNA-seq) have largely given way to single-cell DNA-seq (scDNA-seq), which offers far greater resolution for clonal substructure. Available data has lagged behind computational theory, though. While single-cell RNA-seq (scRNA-seq) has become widely available, scDNA-seq is still sufficiently costly and technically challenging to preclude routine use on large cohorts. This forces difficult tradeoffs between the limited genome coverage of scRNA-seq, limited availability of scDNA-seq, and limited clonal resolution of bulk DNA-seq. These limitations are especially problematic for studying structural variations and focal copy number variations that are crucial to cancer progression but difficult to observe in RNA-seq.

Results: We develop a method, TUSV-INT, combining advantages of these various genomic technologies by integrating bulk DNA-seq and scRNA-seq data into a single deconvolution and phylogenetic inference computation while allowing for single nucleotide variant (SNV), copy number alteration (CNA) and structural variant (SV) data. We accomplish this by using integer linear programming (ILP) to deconvolve heterogeneous variant types and resolve them into a clonal lineage tree. We demonstrate improved deconvolution performance over comparative methods lacking scRNA-seq data or using more limited variant types. We further demonstrate the power of the method to better resolve clonal structure and mutational histories through application to a previously published DNA-seq/scRNA-seq breast cancer data set.

Availability: The source code for TUSV-INT is available at https://github.com/CMUSchwartzLab/TUSV-INT.git

Key words: tumor phylogeny, cancer, somatic evolution, sequencing, single-cell, integer linear programming

Introduction

Cancer develops through an evolutionary process by which an initially healthy cell population successively accumulates mutations, generating genetic diversity that in turn creates opportunities for selection for more aggressive cell populations (clones) over the course of time [31]. While most such mutations are likely selectively neutral [36], chance deleterious mutations can lead cancers to select for phenotypic changes promoting tumor growth and progression [15]. Mutations causing pathway dysregulation might act through changes in gene copy number, molecular functions, and gene expressions, among other mechanisms. This resulting process of clonal evolution, together with the response it provokes in the tumor microenvironment, ultimately lead to the phenotype of tumor growth that defines cancer. Unraveling the enormous complexity of the process, though, has been the work of many years of effort into gathering diverse data sources on tumor genetics, function, and morphology and resolving their data into into coherent models of the process of tumor evolution and its relationship to disease progression [28].

Computational tools for reconstructing the process of clonal evolution in cancers from genomic data sources ("tumor phylogenetics") have become a crucial part of cancer research, with their development emerging as its own vibrant subfield of research driven by and driving biotechnology development for profiling cancer progression [34] as well as helping to spur extensive study of somatic evolution beyond the cancer context [3]. The earliest methods for clonal lineage reconstruction relied on methods that predate the genomic era [25, 32]. Tumor phylogenetics only came into more widespread use as newer genomic technologies began to make cancer genetic variation data available at scale. The first widely used tools were designed to take advantage of bulk DNAseq (e.g., [4, 35]). The advent of scDNA-seq (scDNA-seq) for cancer studies [30] revolutionized clonal phylogenetics, with many methods today designed to us scDNA-seq data (e.g., [7, 33, 18]). However, scDNA-seq remains comparatively costly and technically challenging and so is still not widely used for large-scale cancer genomic studies or clinical practice [8]. Single-cell RNA-seq (scRNA-seq) is far more accessible and cost-effective [16], leading to efforts at scRNA-seq-based tumor phylogenetics [17, 20, 29], although such methods are challenged by the limited coverage of the genome they offer, poor resolution for fine-scale copy number changes, and limited ability to detect structural variations in particular.

Multiomic studies offer a potential path for leveraging advantages of distinct technology platforms. Multiomic methods have a long history in tumor phylogeny studies, including, for example, methods for combining bulk and scRNA-seq [22], bulk and scDNA-seq [26, 23], bulk DNA-seq and pregenomic fluorescent in situ hybridization (FISH) data [21], scDNA-seq and FISH [10], and scDNA-seq and scRNA-seq [6]. Particular value may be found in taking advantage of the comprehensive genomic coverage offered by bulk DNA-seq and the fine clonal resolution offered by scRNA-seq, both of which are now routinely available in low-cost commercial platforms, as recently adopted by some tumor phylogeny efforts [17, 40, 27]. Among these methods, PhylEx [17] and Canopy2 [40] use Bayesian methods to reconstruct tumor phylogenies from SNVs while Cardelino [27] is primarily an scRNA-seq method but can optionally use bulk sequence if provided a clonal guide tree. The benefits of this combination of technologies should be particularly pronounced as we move to develop more comprehensive models of how different variant types collectively contribute to tumor evolution, given the severe limitations of RNA-seq alone for profiling SNVs, CNAs, or SVs comprehensively. Recent work [11, 1] has shown the importance of looking at these classes of variants together, given their complementary value for accurately resolving clonal evolution and our growing appreciation for how they can each act in parallel in causing phenotypic changes that shape clonal evolution pathways [1, 39].

In the present work, we develop TUSV-INT, a platform for clonal evolution studies integrating bulk DNA-seq and scRNA-seq to produce clonal evolution models that are unprecendented in their comprehensive accounting for fine clonal structure, genomic coverage, and diverse variant types (SNV, CNA, and SV). The work uses a general integer linear programming (ILP) framework of clonal lineage reconstruction in the presence of SVs [5] that we have previously adapted to provide similarly comprehensive variant coverage from bulk DNA-seq [11] and scDNA-seq [1] data. We validate TUSV-INT in comparison to published alternative approaches, showing the value of considering multiple variant types and of combining bulk DNA-seq and sc-RNAseq in resolving clonal evolution accurately and comprehensively. We further demonstrate with application to a previously published DNA/RNA breast cancer data set [2] that TUSV-INT is practical on real data and can yield improvements in clonal deconvolution and lineage reconstruction capable of driving novel biological discovery.

Methods

We aim to reconstruct tumor clonal evolutionary histories, infer the SNV, SV, and CNA profiles of individual tumor clones, and determine their mixture fractions in samples of bulk DNA-seq using scRNA-seq data as a guide. We follow prior work of the original TUSV method [5] and variants [11, 1] in representing these variants for the formal problem statement. CNAs are represented as copy numbers for genomic regions, discretized into disjoint segments (bins). SVs are represented as pairs of genomic breakpoints that are adjacent in the cancer genome but non-adjacent in the reference genome. SNVs are represented by a single position in the genome. Allele fractions at SNV and SV positions are calculated as the ratio of reads supporting the alternate allele to those supporting the reference allele. We elaborate on these representations in defining problem variables in the following sections.

Problem Statement

TUSV-INT takes processed variant calls as input for both bulk DNA-seq (which may be WGS, WES, or targeted sequence data) and scRNA-seq. In bulk DNA-seq, we represent SVs with segmental mean copy numbers of paired breakpoint ends, CNAs with allele specific mean copy number for a set of discrete genomic segments, and SNVs with estimated mean copy numbers at the SNV positions. In scRNA-seq data, we represent CNAs with allele-specific mean copy numbers of genomic segments, which can be derived from existing copy number callers, such as Numbat [13] or SIGNALS [12]. Assume we are given: 1) m samples of bulk DNA-seq data covering r genomic segments and containing l SV breakpoints and g SNVs and 2) scRNA-seq data covering the same r genomic segments and corresponding to n tumor clones. Note that for computational efficiency, the variants used as input may be subsampled from the full variant set, a strategy for which we provide a more detailed quantitative justification in our prior work [1]. Given these inputs, we first construct an $m \times (l+g+2r)$ clonal variant matrix, F, from the bulk DNA-seq data. We then build an $n \times 2r$ clonal copy number matrix C^{RNA}, from the scRNA-seq data. Additionally, we build an $l \times r$ breakpoint to segment mapping matrix, Q, identifying the segment in which each breakpoint end is located and an $l \times l$ breakpoint pair mapping matrix G identifying which breakpoints ends are paired to one another. The notations are described in Table 1 in more detail. Given the F, C^{RNA}, Q, and G matrices, our goal is to estimate an $m \times n$ clonal mixture fraction matrix U, $n \times (l+g+2r)$ deconvolved clonal copy number matrix C, $n \times 2r$ estimated RNA copy number matrix C', and $n \times n$ single-cell RNA seq clone to bulk clone mapping matrix M. We do that by solving for a constrained optimization problem minimizing the following objective function:

$$\min_{U,C,M} |F - UC| + |C' - MC^{\text{RNA}}| + \lambda_1 R + \lambda_2 S \tag{1}$$

$|F - UC|$ provides a measure of the quality of the deconvolution of the bulk data. $|C' - MC^{\text{RNA}}|$ provides a measure of consistency between the copy number estimates from deconvolved DNA-seq versus scRNA-seq. R is a regularization terms used to penalize for $L1$ cost of copy number distances along edges of a clonal phylogeny built from the inferred variants, used to bias the optimization towards a clonal deconvolution that yields a plausible evolutionary tree as assessed by an approximate minimum evolution cost. S is a measure of consistency between CNA and SV copy numbers. These terms and the constraint sets by which they are defined are described in more detail below.

Algorithm

We formalize TUSV-INT as a constrained optimization problem and solve for it with a coordinate descent algorithm, iterating

Table 1. Key parameters and variables needed for TUSV-INT

Notation	Interpretation	Type
n	The number of clones	User defined
m	The number of samples	Data derived
l	The number of SV breakpoints	Data derived
g	The number of SNVs	Data derived
r	The number of DNA copy number segments	Data derived
$F \in \mathbb{R}_{\geq 0}^{m \times (l+g+2r)}$	$F_{i,j}$: Average copy number of variant j in sample i	Constant
$Q \in \{0,1\}^{(l+g) \times r}$	$Q_{i,j} = 1$ iff breakpoint or SNV i is in segment j j	Constant
$G \in \{0,1\}^{l \times l}$	$G_{i,j} = 1$ iff i and j are paired breakpoints, deriving from the same SV	Constant
$C^{\mathrm{RNA}} \in \mathcal{Z}_{\geq 0}^{n \times (2r)}$	$C^{\mathrm{RNA}}_{i,j}$: Clonal copy number of CNA j in clone i	Constant
$U \in \mathbb{R}_{\geq 0}^{m \times n}$	$U_{i,j}$: Fraction of clone j in sample i	Variable
$C \in \mathbb{Z}_{\geq 0}^{n \times (l+g+2r)}$	$C_{i,j}$: Clonal copy number of variant j in clone i	Variable
$M \in \{0,1\}^{n \times n}$	$M_{i,j} = 1$ iff clone i in the phylogeny represents RNA clone j	Variable
$E \in \{0,1\}^{n \times n}$	$E_{i,j} = 1$ iff there is a directed edge from clone i to clone j	Variable
$A \in \{0,1\}^{n \times n}$	$A_{i,j} = 1$ iff clone i is an ancestor of clone j	Variable
$W \in \{0,1\}^{n \times n \times (l+g)}$	$W_{i,j,k} = 1$ iff breakpoint or SNV k is mapped on the edge from clone i to clone j	Variable

between subproblems of the full optimization each of which we solve using integer linear programming (ILP). This is based on a strategy first developed for the original TUSV [5] and elaborated on in the more general TUSV-ext [11] for solving for tumor phylogenies integrating SNVs, CNAs and SVs from bulk-DNA sequencing data alone. The present work extends the TUSV-ext algorithm to accommodate the additional inputs, outputs, and constraints created by introducing scRNA-seq data.

We use coordinate descent to address the nonlinearity of our objective function in estimating the clonal mixture fraction matrix U and clonal copy number matrix C. The algorithm iteratively optimizes independently for U and C, starting with a random initialization of U. The variant copy number matrix C is estimated in two stages: 1) inference of the segmental CNAs from single-cell RNA sequencing data by estimating M and C' and 2) estimation of the clonal SNVs and SVs constrained by the ILP formulation. We describe these in more detail in the following sections, focusing particularly on the extensions beyond TUSV-ext. We refer readers to the prior works [5, 11] for portions of the optimization that are largely unchanged, except where needed for clarity of exposition.

Estimating clonal mixture fraction matrix U

We define a set of ILP constraints to estimate U given F and C. The objective function is set to minimize the L1 norm of $|F - UC|$. Additionally, for each sample, we impose constraints so that the mixture fractions sum to 1 (Eqn. 2).

$$\sum_{j=1}^{n} U_{i,j} = 1 \ \forall \ i \in \{1, \ldots, m\} \tag{2}$$

Estimating M and C′

Once U is estimated, we estimate the $n \times n$ bulk DNA-seq-to-scRNA-seq clone mapping matrix M and $n \times 2r$ estimated RNA copy number matrix C' by minimizing the L1 norm of $|C' - MC^{\mathrm{RNA}}|$. We impose constraints on M to ensure that each DNA clone corresponds to exactly one RNA clone (Eqn. 3, 4, 5).

$$M_{i,j} \in \{0,1\} \ \forall \ i,j \in \{1, \ldots, n\} \tag{3}$$

$$\sum_{j=1}^{n} M_{i,j} = 1 \ \forall \ i \in \{1, \ldots, n\} \tag{4}$$

$$\sum_{i=1}^{n} M_{i,j} = 1 \ \forall \ j \in \{1, \ldots, n\} \tag{5}$$

$$\tag{6}$$

We then define the RNA copy number distance portion of the objective in Eqn. 1 using L1 distance as the following equations.

$$\mathcal{C}^{\Delta}_{i,j} \geq C'_{i,j} - \sum_{k=1}^{n} M_{i,k}.C^{\mathrm{RNA}}_{k,j} \tag{7}$$

$$\mathcal{C}^{\Delta}_{i,j} \geq -C'_{i,j} + \sum_{k=1}^{n} M_{i,k}.C^{\mathrm{RNA}}_{k,j} \tag{8}$$

$$|C' - MC^{\mathrm{RNA}}| = \sum_{i=1}^{n} \sum_{j=1}^{2r} \mathcal{C}^{\Delta}_{i,j} \tag{9}$$

Simultaneously, we set the estimated C' as the segmental CNA values of C (Eqn. 10).

$$C_{i,l+g+j} = C'_{i,j} \forall \ i \in \{1, \cdots, n\}, j \in \{1, \cdots, 2r\} \tag{10}$$

Edge and ancestry constraints

We model the underlying evolutionary history associated with the variants as a rooted directed binary phylogenetic tree $\mathcal{T}$ with n nodes corresponding to n clones, represented by a binary adjacency matrix E, where $e_{i,j} = 1$ if clone i is the parent of clone j for $i, j \in \{1, \cdots, n\}$. To ensure that E represents a valid tree structure, we designate the first $n' = \frac{(n-1)}{2}$ nodes to represent the leaves of the tree, $n'+1$ through $n-1$ to represent the internal nodes, and node n to represent the root. n' represents the number of internal nodes in $\mathcal{T}$. We then impose the following constraints on the E matrix:

The root has no incoming edge and the leaves have no outgoing edge -

$$E_{i,n} = 0 \quad \forall\, i \in \{1, \cdots, n\} \tag{11}$$

$$E_{i,j} = 0 \quad \forall\, i \in \{0, \cdots, n'\}, j \in \{1, \cdots, n\} \tag{12}$$

Additionally, following TUSV-ext [11], we define an ancestor matrix $A \in \{0,1\}^{n \times n}$. $A_{i,j} = 1$ if clone i is an ancestor of clone j. We define the following two constraints to ensure that a node's parent is its ancestor and a child node will carry all of its parent's ancestors -

$$A_{i,j} \geq E_{i,j} \quad \forall\, i, j \in \{1, \cdots, n\} \tag{13}$$

$$A_{k,j} \geq A_{k,i} + E_{i,j} - 1 \quad \forall\, i, j \in \{0, \cdots, n\}, k \in \{1, \cdots, n\} \setminus \{i\} \tag{14}$$

$$A_{k,j} \geq A_{k,i} - E_{i,j} + 1 \quad \forall\, i, j \in \{0, \cdots, n\}, k \in \{1, \cdots, n\} \setminus \{i\} \tag{15}$$

Copy number constraints

We define similar constraints to TUSV-ext [11] to ensure a tree structure on our estimated variant copy number matrix C. We assume that the root is a normal clone and therefore has no SNV/SV mutations, as well as normal diploid copy number at each genomic segment.

$$C_{n,j} = 0 \quad \forall\, j \in \{1, \cdots, (l+g)\} \tag{16}$$

$$C_{n,j} = 1 \quad \forall\, j \in \{(l+g+1), \cdots, (l+g+2r)\} \tag{17}$$

For algorithmic tractability, we constrain the maximum copy number to be a user-specifiable parameter $c_{\max}$.

$$C_{i,j} \leq c_{max} \; \forall\, i \in \{1, \cdots, n\}, j \in \{(l+g+1), \cdots, (l+g+2r)\} \tag{18}$$

Additional constraints define the phylogenetic cost R as the sum of L1 distances of copy number differences across the edges of the tree. These are defined as in our prior work [11] and so are not reproduced here.

Dollo phylogeny on breakpoints and SNVs and breakpoint consistency

We apply constraints for the loss-supported Dollo phylogeny model, where a mutation can be gained only once, but lost multiple times. However, to favor consistency between CNA and SNV/SV data, we assume that SNV and SV losses are coupled with a loss of their corresponding genomic segments. Similarly, SNV and SV amplifications are constrained by their corresponding segmental amplifications, as described in TUSV-ext [11]. Additionally, we assign each breakpoint and SNV to an allele. This is achieved by introducing a binary matrix $D \in 0, 1^{(l+g)}$, where $D_b = 1$ if and only if breakpoint or SNV b belongs to the first allele, and 0 otherwise.

Since each breakpoint belongs to exactly one segment, a breakpoint copy number should not exceed the copy number of its corresponding segment. For a breakpoint b, we impose the following constraints:

$$\begin{gathered} C_{i,b} \leq \sum_{s=1}^{r} Q_{b,s} C_{i,l+g+s} + (1 - D_b) c_{\max} \\ C_{i,b} \leq \sum_{s=1}^{r} Q_{b,s} C_{i,l+g+r+s} + D_b c_{\max} \\ \forall\, i \in \{1, \cdots, n\}, b \in \{1, \cdots, l\} \end{gathered} \tag{19}$$

Finally, our second regularization term S in the objective function (Eqn. 1) enforces a relationship between the CNA segmental copy numbers and the breakpoint allele copy numbers. Since breakpoint copy numbers can be at most as large as the segmental copy numbers, the mixture copy number at the breakpoint position should also be at most the mixture segmental copy numbers. For a breakpoint b in sample i, we denote the ratio of the mixture copy numbers at the breakpoint or SNV position and the corresponding segment as $\pi_{i,b}$.

$$\pi_{i,b} = \frac{F_{i,b}}{F_{i,l+g+b} + F_{i,l+g+r+b}} \tag{20}$$

This ratio represents the sum of the amplification or deletion events of the mutation-containing allele in comparison to the genomic segment as a whole. By minimizing the difference between the observed and estimated ratios, we align the estimated data with the observed mixture fractions. Following TUSV-ext [11], we include this minimization as a regularization term in the objective function to determine S.

Choice of the regularization parameters

We set the regularization parameters to scale R and S so that they are comparable to the larger of the first two terms in the objective function. Specifically, we choose

$$\lambda_1 = \frac{1}{2} \frac{(l+g+2r)m}{2rN},$$

which is the ratio of the size of the first term to R, and

$$\lambda_2 = \frac{1}{2} \frac{l+g+2r}{l+g},$$

as in TUSV-ext [11], based on empirical observation that these formulas perform well in practice in the original TUSV and that solutions are minimally sensitive in the vicinity of these values.

Availability

TUSV-INT is implemented in Python 2 and utilizes the licensed Gurobi optimizer for ILP solution. The code is available at https://github.com/CMUSchwartzLab/TUSV-INT.git.

Results

We first validated the accuracy and robustness of TUSV-INT using simulated data and assessed its performance relative to TUSV-ext and Canopy2 [40]. While there are many tools available for tumor phylogenetics, there are no accepted standards for their specification and differences in inputs and outputs render direct head-to-head comparisons difficult. We compare to TUSV-ext to provide a direct assessment of the value of scRNA-seq in deconvolving bulk DNA-seq. A so-far unique feature of our TUSV-family methods is their simultaneous use of SNV, CNA, and SV data, and we therefore use Canopy2 to assess the comparative advantage TUSV-INT gets from including additional variant types relative to a leading tool also using bulk DNA-seq and scRNA-seq.

Additionally, we evaluated TUSV-INT's effectiveness on real cancer patient samples. A full direct comparison of TUSV-INT with other existing methods is again not feasible, as none of them utilize single-cell RNA sequences for clonal deconvolution from bulk DNA sequencing data using all of the variant types TUSV-INT can access. We therefore again compare it with TUSV-ext, which reconstructs tumor phylogenies with bulk DNA sequences only with all of the same variant types, to assess what specifically we gain by adding scRNA-seq to the analysis.

Canopy2 [40] reconstructs tumor phylogenies from bulk DNA-seq and scRNA-seq based only on SNVs. While Canopy2 utilizes reference and alternate reads from bulk DNA-seq and scRNA-seq data, we use the copy numbers from scRNA-seq to guide our method. Canopy2 relies on assumptions of different mutation "activation" and "deactivation" rates in relating scRNA-seq data to clonal mixture fractions and mutation frequencies. In contrast, we do not explicitly model these rates. Instead, to ensure a fair comparison, we generate scRNA-seq reference and alternate read counts for Canopy2 assuming uniform activation and deactivation rates across all mutations and no sequencing errors. We achieve this by keeping the scRNA-seq reference and alternate read counts consistent with the clonal B-allele frequencies (BAFs) observed in the bulk mixture. For this purpose, we first extract clonal BAFs from the bulk mixture to represent the proportion of alternate allele reads for each clone. Next, for each single cell, we simulate reference and alternate read counts at each site by scaling the total read depth to 120x. These comparisons are meant only to assess the value of the unique combination of data sources (sequence types and variant types) used by TUSV-INT and so we do not explore other scenarios where the variant types assumed by TUSV-INT are not available.

Evaluation with simulated data

To evaluate the performance of TUSV-INT, we simulated bulk WGS data with SVs, SNVs, and CNAs based on known ground-truth tumor phylogenies, mutation rates, clonal populations, and their frequencies. Each simulation represented subclones from a single patient with two samples, using genomic segments from chromosomes 1 and 2. We began by assigning random clonal frequencies to the subclones. We then introduce genetic alterations, where we generate SNVs at uniformly random positions, and SVs with lengths drawn from a Poisson distribution with a mean of 5,745,000 bp, corresponding to the average SV length observed in the TCGA-BRCA cohort [5]. CNAs are generated using a relative probability distribution of 2:1:2:1 for amplifications, inversions, deletions, and translocations, respectively. To model sequencing errors, we generate segment-specific read counts using a Poisson distribution with a mean depth of 50. These read counts are used as the trial numbers for a binomial density that generates variant-associated read counts based on variant allele frequencies (VAFs). For each variant, we compute theoretical bulk DNA-seq copy numbers as a weighted sum across subclones based on their respective clonal frequencies. Using these copy numbers, we calculate variant allele frequencies (VAFs) for SNVs and breakpoints, as well as BAFs for CNAs. For transcriptomic simulations, we generate transcript counts using a Poisson density with a mean of 50 for at least 3 and at most 6 genes per genomic segment from the bulk DNA sequences for each clone. These counts are averaged and normalized for allele-specific single-cell RNA sequencing segment-wise CNA profiles.

We ran TUSV-INT with the following parameter sets — (1) varying numbers of clones, $n \in \{3, 5, 7, 9\}$ and (2) varying numbers of SNV mutations, $g \in \{25, 50, 100\}$ — with each scenario having 3 iterations, 3 random restarts, 5000 seconds maximum compute time per iteration, 80 SVs, 40 SNVs (with the exception of 20 SNV for $g = 25$) and with the corresponding genomic segments for CNAs. Figure 1 shows the root mean square error (RMSE) of the true versus estimated clonal fraction matrices in these simulation settings.

In all cases, TUSV-INT outperformed our previous method, TUSV-ext, which relies solely on bulk DNA-seq. We observed a decrease in accuracy for TUSV-INT as the number of clones increases (Figure 1 (a), (d), (e)), indicating that deconvolution becomes harder for a larger dataset. Inference of clonal SVs suffers more from increasing clone numbers than SNVs. Clonal frequencies show a smaller loss in quality with increasing numbers of clones than variant inferences. This is in contrast to TUSV-ext, which shows less sensitivity to clone numbers, although still worse performance than TUSV-INT even under the most favorable conditions. This may correspond to the greater difficulty of accurately mapping scRNA-seq data to distinct DNA-seq clones as the clone number increases.

Increasing numbers of variants leads to less consistent trends between the two methods (Figure 1 (b), (e), (h)). Both methods show some loss of accuracy in SVs with increasing variant numbers, although more stable results for SNVs. Clonal frequencies are essentially stable for TUSV-INT with increasing numbers of variants. TUSV-INT remains consistently the better performer and the performance gap between the two methods is relatively stable across changes in variant numbers for the most part.

Increasing number of samples (Figure 1 (c), (f), (i)) generally led to improved solution quality by all measures, with the exception of the outlier of clonal frequency accuracy degrading for ten samples. We would generally expect more samples to lead to more effective deconvolution despite potentially increasing the size and difficulty of the optimization problem. There is a slight increase in separation between the two methods within increasing sample numbers, although not clearly distinguishable from chance. TUSV-ext did not yield results for ten samples within the allotted compute time. While improved runtime was not the goal of TUSV-INT, the results may indicate an unanticipated value of scRNA-seq in more tightly constraining the problem and pruning the search space, thus accelerating ILP-type optimization.

We also evaluated Canopy2's [40] performance on our simulated datasets, comparing to TUSV-INT and TUSV-ext with respect to clonal frequency inference and SNV placement onto clones. For each simulation instance, we provided Canopy2 with the true number of clones and ran it using its default parameters. Fig. 1 shows Canopy2's performance in green. Since Canopy2 does not make use of SVs, there are no Canopy2 results for Fig. 1(d,e,f). Root mean squared errors (RMSE) for clonal frequency inference and SNV placement with Canopy2 are generally comparable to those of TUSV-ext, with both notably worse than TUSV-INT. The results confirm observations from prior TUSV variants that the addition of SV data leads to improved accuracy by other measures than just SVs, as SVs provide complementary information to other variant types

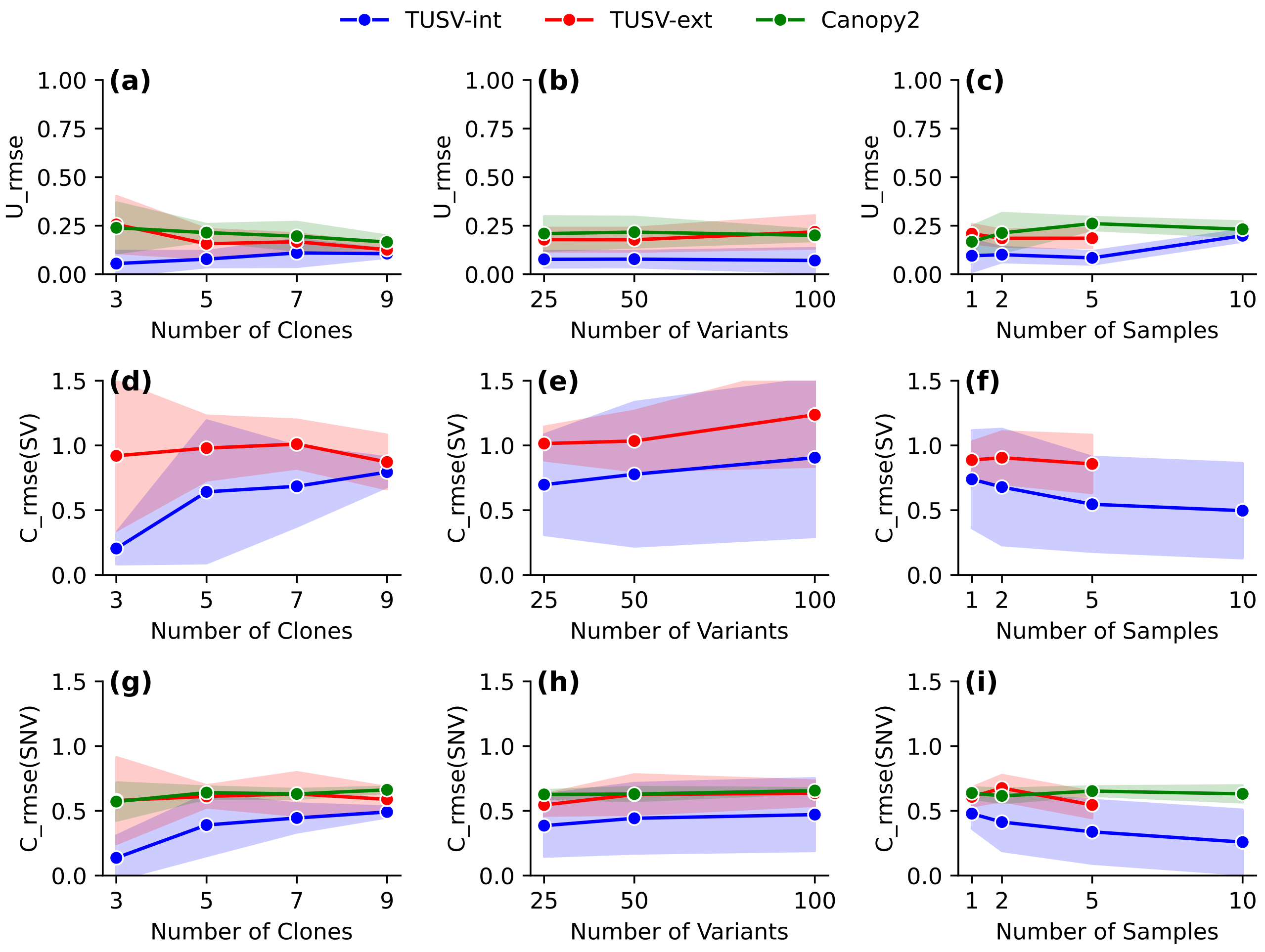

Fig. 1. Results on simulation with different number of clones, samples and mutations. (a) - (c) shows mean squared error of the true vs estimated clonal mixture fractions (U) for TUSV-INT , TUSV-ext and Canopy2. (d) - (f) shows the mean squared error of the estimated SVs for TUSV-INT and TUSV-ext. (g) - (h) shows the mean squared error of the estimated SNVs for TUSV-INT , TUSV-ext and Canopy2. We ran TUSV-INT and TUSV-ext with m = 5000, t = 3, r = 3, sv_ub = 80 and C = 120 (except for 25 SNVs, where C = 100). The solid lines represent the means of the simulation instances and shaded areas represent one standard deviation on either side of the mean. TUSV-INTis evaluated at all data points. TUSV-ext and Canopy2 results were feasible for only for subsets of the data points, as described in the accompanying text. Note that clone 0 is assigned a small frequency in (c) despite not appearing in the TUSV-INT phylogeny of (a) because it was inferred in one of the ten TUSV-INT runs.

that helps in correctly resolving the tree topology and inferring placement of all variants.

Real breast cancer patient samples

To demonstrate TUSV-INT's performance on a true example of joint bulk DNA-seq and scRNA-seq data, we selected one dataset from an ER/HER2-positive breast cancer patient (BC03) among 11 patients studied by Chung et al. [2]. This dataset included three bulk WES samples collected from the primary tumor, metastatic lymph node, and blood. scRNA-seq was performed on 92 single cells from the primary tumor site and metastatic lymph nodes, and thus we included these two samples in our analyses while omitting the blood sequence for which no matched scRNA-seq is available. The bulk WES coverage was 100x for the tumor and metastatic lymph node, and 50x for the paired blood sample.

We preprocessed and called variants from the bulk WES data using the Nfcore/sarek pipeline [9, 14], aligning paired-end FASTQ files to the GRCh38 human reference genome. Raw sequencing reads were then trimmed and quality-controlled with fastp and FastQC, mapped using BWA-MEM2, deduplicated with GATK MarkDuplicates, and the base quality scores were recalibrated. Strelka2 [19], and CNVkit [38] were then used to call SNVs and allele-specific CNVs, respectively. We derived average allele-specific copy numbers from the total copy numbers reported by CNVkit, using the Strelka2 genotypes. We preprocessed the scRNA-seq data in a similar manner from the paired-end raw sequencing reads. Allele-specific CNA calls were then performed using Numbat [13]. Since Chung et al. [2] generated the RNA sequencing data using the SMARTer Ultra Low RNA Kit [2], we ran Numbat with the "SmartSeq" setting to accommodate the non-barcoded single-cell RNA sequences.

After obtaining SNVs and CNAs from bulk WES and allele-specific CNAs from scRNA-seq, we ran TUSV-INT using 5 clones, 3 random restarts, 3 iterations, a time limit of 1500 seconds per iteration, 120 SNVs, and the corresponding average bulk CNAs alongside scRNA-seq-derived CNAs. For comparison, we also ran TUSV-ext using the same settings. We used the five scRNA-seq clones derived from Numbat, and calculated the average segmental copy number within each cluster to derive representative values as an input to TUSV-INT.

Fig. 2 summarizes the results, comparing those between TUSV-INT and the DNA-only TUSV-ext. TUSV-INT inferred four clones (Fig. 2(a)) on 9 of its 10 runs, consistent with the results reported by Canopy2 for an optimal run on patient BC03's data, and places the SNVs and SVs across all branches of the phylogenetic tree (Figure 2 (a)). On the other hand, TUSV-ext inferred five clones from the two samples (Fig. 2(b)), although its clone 0 and clone 1 have low support (Figure 2 (b)). The consistency between TUSV-INT and Canopy2 results provides indirect support for the value of scRNA-seq to better resolve clonal structure. Clone 2 in both the trees have common mutations, however, they are placed differently in the two trees. We would expect the single-cell data available to TUSV-INT to better identify which variants co-occur in the same clone and correct an error in placement in TUSV-ext. This leads to some notable changes in inference of mutation orders, for example in the observation that SNV mutations hitting the TP53 and Ras pathways (TP53BP1 and RASGRF2) and immune-related CNA mutation in TOLLIP occur at a late stage of progression following earlier mutations such as SNV mutation of SOX5 and CNA mutation of CD53. This is in contrast to the TUSV-ext result, which placed the mutations in clones 2 and 3 all as early mutations on two parallel clonal lineages. The distinction may be significant in interpreting the meaning of the results. For example, mutations affecting SOX5 and TP53BP1 can both promote proliferation in breast cancers through very different mechanisms [24, 37]. The TUSV-ext result would suggest these could be independently evolved ways of driving a similar phenotype in distinct clones while the improved TUSV-INT result instead suggests they may be acting synergistically in the same cell lineage. Such a distinction can have important implications for identifying potential treatment options or predicting mechanisms of therapy resistance.

The two methods also yield fairly different inferences of clonal frequencies averaged across ten replicates (Fig. 2(c) and Fig. 2(d)), although the differences in the inferred clones themselves and their phylogenetic placements complicate interpretation. We would expect the scRNA-seq data to give TUSV-INT a significant advantage in inferring accurate clonal frequencies, likely in addition to correcting errors due to misinferred clonal identities. TUSV-INT and TUSV-ext both infer clone 1 to be the dominant clone in the primary tumor, although TUSV-INT infers it to be substantially more diverged from the rarer clone 3 from which it descends. Both also infer clone 2 to be the dominant clone in the metastasis, making the revised placement of the clone potentially important in correctly characterizing the pathway to metastatic progression in this patient. Note that clone 0 is assigned a small frequency in the TUSV-INT results despite not appearing in the TUSV-INT phylogeny of Fig. 2(a) because it was inferred in one of the ten TUSV-INT runs. There is little we can conclude about mechanism from a single patient, but the results illustrate how scRNA-seq data can lead to a more accurate model of the clonal evolution process and clonal dynamics across stages of progression, while taking account of how distinct types of genetic variants contribute to both phenomena.

Discussion

We have developed a method, TUSV-INT, for integrating bulk DNA-seq and scRNA-seq data for clonal phylogenies accommodating SNV, CNA, and SV variants. The work makes use of similar strategies for combining bulk DNA-seq and scRNA-seq to other recent works [17, 40] while handling a broader combination of variant types than any competing method. We accomplish this with a model built around an ILP framework for phylogenetic inference on heterogeneous data that offers comparatively easy extensibility to new data sources, variant types, and features of the biology. Validation on simulated data shows the advantages of combining both the two sequencing types and the three variant types in producing more accurate clonal deconvolution and lineage trees than do prior methods. Application to real data reinforces the importance of this combination in accurately resolving clonal evolution and revealing how different forms of variation can each contribute in parallel to the biology of cancer progression.

The work nonetheless leaves a number of avenues for improvement. Our method currently does not account for clones present in bulk DNA-seq but absent from single-cell RNA sequencing, a question we leave for future exploration. Improving computational efficiency is also a future concern, as ILP can be a computationally costly technique, in some cases necessitating subsampling of large data sets. ILP, despite its power for efficiently solving hard optimization problems, also has limitations compared to more common maximum likelihood and Bayesian approaches to phylogenetics. For example, combinatorial optimization methods like ILP provide less direct support for considering uncertainty in model inferences. It would have great value to the clonal lineage problems studied here and to and the field more generally to find new methods that can bring together the speed and versatility of ILP-like combinatorial optimization with the more principled handling of uncertainty of probabilistic models. There is also likely more advantage to be gained by considering ways to incorporate other data sources into the analysis, such as scDNA-seq, long-read sequencing, or data from blood-based sequencing ("liquid biopsy"). Finally, we have focused here primarily on methods development and the question is largely unexplored yet what the improved power of these methods can tell us about somatic variation and clonal evolution in cancer progression as well as in other diseases and normal aging.

Competing interests

No competing interest is declared.

Author contributions statement

N.B. and R.S. conceived the method and study and wrote and reviewed the manuscript. N.B. implemented the code, executed the experiments, and analysed the results.

Acknowledgments

We thank the Pittsburgh Supercomputing Center (PSC) and its ACCESS resource for providing compute resources used in the analyses described here. N.B. was supported in part by UPMC Enterprises through a Fellowship in Digital Health Innovation from Carnegie Mellon's Center for Machine Learning and Health (CMLH). Research reported in this publication was

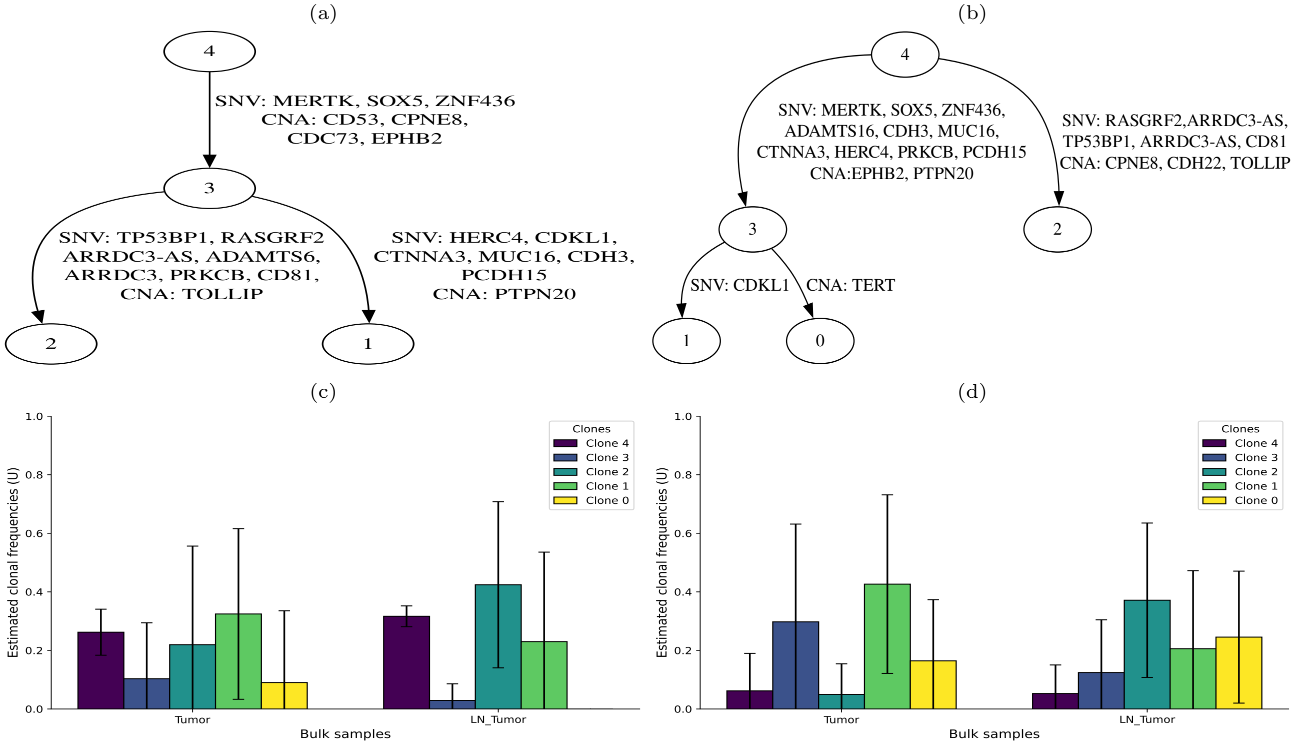

Fig. 2. Results on ER/HER2 positive breast cancer patient BC03's data. (a) Clonal tree inferred by TUSV-INT with edges annotated by single-nucleotide variants (SNVs) and copy number alterations (CNAs). (b) Clonal tree inferred by TUSV-ext [11], also annotated with SNVs and CNAs. (c) Clonal mixture fractions in tumor and metastatic lymph node tumor (LN_Tumor) samples inferred by TUSV-INT. (d) Clonal mixture fractions in tumor and LN_Tumor samples inferred by TUSV-ext. We ran TUSV-INT and TUSV-ext with 120 subsampled SNVs, 3 iterations, 3 random restarts and 1500 second per iteration.

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.

References

1. N. A. Bristy, X. Fu, and R. Schwartz. Sc-tusv-ext: Single-cell clonal lineage inference from single nucleotide variants (snv), copy number alterations (cna) and structural variants (sv). *bioRxiv*, pages 2023–12, 2023.
2. W. Chung, H. H. Eum, H.-O. Lee, K.-M. Lee, H.-B. Lee, K.-T. Kim, H. S. Ryu, S. Kim, J. E. Lee, Y. H. Park, et al. Single-cell rna-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature communications*, 8(1):15081, 2017.
3. S. De. Somatic mosaicism in healthy human tissues. *Trends in Genetics*, 27(6):217–223, 2011.
4. 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*, 16(1):1–20, 2015.
5. J. Eaton, J. Wang, and R. Schwartz. Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics*, 34(13):i357–i365, 2018.
6. M. Edrisi, X. Huang, H. A. Ogilvie, and L. Nakhleh. Accurate integration of single-cell dna and rna for analyzing intratumor heterogeneity using macrodna. *Nature communications*, 14(1):8262, 2023.
7. M. El-Kebir. Sphyr: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics*, 34(17):i671–i679, 2018.
8. G. D. Evrony, A. G. Hinch, and C. Luo. Applications of single-cell dna sequencing. *Annual review of genomics and human genetics*, 22(1):171–197, 2021.
9. P. A. Ewels, A. Peltzer, S. Fillinger, H. Patel, J. Alneberg, A. Wilm, M. U. Garcia, P. Di Tommaso, and S. Nahnsen. The nf-core framework for community-curated bioinformatics pipelines. *Nature biotechnology*, 38(3):276–278, 2020.
10. 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*, 28(11):1035–1051, 2021.
11. X. Fu, H. Lei, Y. Tao, and R. Schwartz. Reconstructing tumor clonal lineage trees incorporating single-nucleotide variants, copy number alterations and structural variations. *Bioinformatics*, 38(Supplement_1):i125–i133, 2022.
12. T. Funnell, C. H. O'Flanagan, M. J. Williams, A. McPherson, S. McKinney, F. Kabeer, H. Lee, S. Salehi, I. Vázquez-García, H. Shi, et al. Single-cell genomic variation induced by mutational processes in cancer. *Nature*, 612(7938):106–115, 2022.
13. T. Gao, R. Soldatov, H. Sarkar, A. Kurkiewicz, E. Biederstedt, P.-R. Loh, and P. V. Kharchenko. Haplotype-aware analysis of somatic copy number

variations from single-cell transcriptomes. *Nature Biotechnology*, 41(3):417–426, 2023.

14. M. Garcia, S. Juhos, M. Larsson, P. I. Olason, M. Martin, J. Eisfeldt, S. DiLorenzo, J. Sandgren, T. D. De Ståhl, P. Ewels, et al. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Research*, 9, 2020.
15. D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.
16. D. Jovic, X. Liang, H. Zeng, L. Lin, F. Xu, and Y. Luo. Single-cell rna sequencing technologies and applications: A brief overview. *Clinical and translational medicine*, 12(3):e694, 2022.
17. S.-H. Jun, H. Toosi, J. Mold, C. Engblom, X. Chen, C. O'Flanagan, M. Hagemann-Jensen, R. Sandberg, S. Aparicio, J. Hartman, et al. Reconstructing clonal tree for phylo-phenotypic characterization of cancer using single-cell transcriptomics. *Nature Communications*, 14(1):982, 2023.
18. T. L. Kaufmann, M. Petkovic, T. B. Watkins, E. C. Colliver, S. Laskina, N. Thapa, D. C. Minussi, N. Navin, C. Swanton, P. Van Loo, et al. Medicc2: whole-genome doubling aware copy-number phylogenies for cancer evolution. *Genome biology*, 23(1):241, 2022.
19. S. Kim, K. Scheffler, A. L. Halpern, M. A. Bekritsky, E. Noh, M. Källberg, X. Chen, Y. Kim, D. Beyter, P. Krusche, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nature methods*, 15(8):591–594, 2018.
20. S. Kurtenbach, A. M. Cruz, D. A. Rodriguez, M. A. Durante, and J. W. Harbour. Uphyloplot2: visualizing phylogenetic trees from single-cell rna-seq data. *BMC genomics*, 22(1):419, 2021.
21. H. Lei, E. M. Gertz, A. A. Schäffer, X. Fu, Y. Tao, K. Heselmeyer-Haddad, I. Torres, G. Li, L. Xu, Y. Hou, et al. Tumor heterogeneity assessed by sequencing and fluorescence in situ hybridization (fish) data. *Bioinformatics*, 37(24):4704–4711, 2021.
22. H. Lei, X. A. Guo, Y. Tao, K. Ding, X. Fu, S. Oesterreich, A. V. Lee, and R. Schwartz. Semi-deconvolution of bulk and single-cell rna-seq data with application to metastatic progression in breast cancer. *Bioinformatics*, 38(Supplement_1):i386–i394, 2022.
23. H. Lei, B. Lyu, E. M. Gertz, A. A. Schäffer, X. Shi, K. Wu, G. Li, L. Xu, Y. Hou, M. Dean, et al. Tumor copy number deconvolution integrating bulk and single-cell sequencing data. *Journal of Computational Biology*, 27(4):565–598, 2020.
24. X. Li, B. Xu, M. S. Moran, Y. Zhao, P. Su, B. G. Haffty, C. Shao, and Q. Yang. 53bp1 functions as a tumor suppressor in breast cancer via the inhibition of nf-κb through mir-146a. *Carcinogenesis*, 33(12):2593–2600, 2012.
25. C. C. Maley. Multistage carcinogenesis in barrett's esophagus. *Cancer letters*, 245(1-2):22–32, 2007.
26. S. Malikic, K. Jahn, J. Kuipers, S. C. Sahinalp, and N. Beerenwinkel. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nature communications*, 10(1):2750, 2019.
27. D. J. McCarthy, R. Rostom, Y. Huang, D. J. Kunz, P. Danecek, M. J. Bonder, T. Hagai, R. Lyu, W. Wang, et al. Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes. *Nature methods*, 17(4):414–421, 2020.
28. N. McGranahan and C. Swanton. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*, 168(4):613–628, 2017.
29. J. C. Moravec, R. Lanfear, D. L. Spector, S. D. Diermeier, and A. Gavryushkin. Cancer phylogenetics using single-cell rna-seq data. *bioRxiv*, pages 2021–01, 2021.
30. N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, et al. Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341):90–94, 2011.
31. P. C. Nowell. The clonal evolution of tumor cell populations: Acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression. *Science*, 194(4260):23–28, 1976.
32. G. Pennington, C. A. Smith, S. Shackney, and R. Schwartz. Reconstructing tumor phylogenies from heterogeneous single-cell data. *Journal of Bioinformatics and Computational Biology*, 5(02a):407–427, 2007.
33. 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.
34. R. Schwartz and A. A. Schäffer. The evolution of tumour phylogenetics: principles and practice. *Nature Reviews Genetics*, 18(4):213–229, 2017.
35. R. F. Schwarz, A. Trinh, B. Sipos, J. D. Brenton, N. Goldman, and F. Markowetz. Phylogenetic quantification of intra-tumour heterogeneity. *PLoS computational biology*, 10(4):e1003535, 2014.
36. A. Sottoriva, H. Kang, Z. Ma, T. A. Graham, M. P. Salomon, J. Zhao, P. Marjoram, K. Siegmund, M. F. Press, D. Shibata, et al. A big bang model of human colorectal tumor growth. *Nature genetics*, 47(3):209–216, 2015.
37. C. Sun, Y. Ban, K. Wang, Y. Sun, and Z. Zhao. Sox5 promotes breast cancer proliferation and invasion by transactivation of ezh2. *Oncology Letters*, 17(3):2754–2762, 2019.
38. E. Talevich, A. H. Shain, T. Botton, and B. C. Bastian. Cnvkit: genome-wide copy number detection and visualization from targeted dna sequencing. *PLoS computational biology*, 12(4):e1004873, 2016.
39. L. L. Weber, A. Hart, I. Ochoa, and M. El-Kebir. Pharming: Joint clonal tree reconstruction of snv and cna evolution from single-cell dna sequencing of tumors. *bioRxiv*, pages 2024–11, 2024.
40. A. M. K. Weideman, R. Wang, J. G. Ibrahim, and Y. Jiang. Canopy2: tumor phylogeny inference by bulk dna and single-cell rna sequencing. *bioRxiv*, 2024.