Bioinformatics Advances, 2025, **00**, vbaf234
https://doi.org/10.1093/bioadv/vbaf234
Advance Access Publication Date: 29 September 2025
Original Article

OXFORD

Phylogenetics

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, PA 15213, United States
[2]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, United States

*Corresponding authors. Nishat Anjum Bristy, Ray and Stephanie Lane Computational Biology Department, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, United States. E-mail: nbristy@andrew.cmu.edu; Russell Schwartz, Department of Biological Sciences and Ray and Stephanie Lane Computational Biology Department, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, United States. E-mail: russells@andrew.cmu.edu.

Associate Editor: Thomas Lengauer

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 remains costly and technically challenging, precluding routine use on large cohorts. This forces 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 present TUSV-INT, integrating bulk DNA-seq and scRNA-seq into a single deconvolution and phylogenetic inference framework while accommodating single nucleotide variants (SNV), copy number alterations (CNA), and structural variants (SV). Using integer linear programming (ILP), we deconvolve heterogeneous variant types and resolve them into a clonal lineage tree. We demonstrate improved deconvolution performance over methods lacking scRNA-seq or using more limited variant types. We further demonstrate its ability to better resolve clonal structure and mutational histories on a published DNA-seq/scRNA-seq breast cancer dataset.
Source code is available at https://github.com/CMUSchwartzLab/TUSV-INT (https://doi.org/10.5281/zenodo.16884120).

Availability and implementation: Source code is available at https://github.com/CMUSchwartzLab/TUSV-INT (https://doi.org/10.5281/zenodo.16884120).

1 Introduction

Cancer develops through an evolutionary process by which a 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 (Nowell 1976). While most such mutations are likely selectively neutral (Sottoriva *et al.* 2015), chance deleterious mutations can lead cancers to select for phenotypic changes promoting tumor growth and progression (Hanahan and Weinberg 2000). 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 coherent models of tumor evolution and its relationship to disease progression (McGranahan and Swanton 2017). 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 (Schwartz and Schäffer 2017) as well as helping to spur extensive study of somatic evolution beyond the cancer context (De 2011). The earliest methods for clonal lineage reconstruction relied on technologies that predate the genomic era (Maley 2007, Pennington *et al.* 2007). 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. Schwarz *et al.* 2014, Deshwar *et al.* 2015). The advent of scDNA-seq for cancer studies (Navin *et al.* 2011) revolutionized clonal phylogenetics, with many methods today designed to use scDNA-seq data (e.g. El-Kebir 2018, Satas *et al.* 2020, Kaufmann *et al.* 2022). 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 (Evrony *et al.* 2021).

Received: July 3, 2025; **Revised:** September 12, 2025; **Accepted:** September 16, 2025
© The Author(s) 2025. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Single-cell RNA-seq (scRNA-seq) is far more accessible and cost-effective (Jovic *et al.* 2022), leading to efforts at scRNA-seq-based tumor phylogenetics (Kurtenbach *et al.* 2021, Moravec *et al.* 2021, Jun *et al.* 2023), 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, e.g. methods for combining bulk and scRNA-seq (Lei *et al.* 2022), bulk and scDNA-seq (Malikic *et al.* 2019, Lei *et al.* 2020), bulk DNA-seq and pregenomic fluorescent in situ hybridization (FISH) data (Lei *et al.* 2021), scDNA-seq and FISH (Fu *et al.* 2021), and scDNA-seq and scRNA-seq (Edrisi *et al.* 2023). 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 (McCarthy *et al.* 2020, Jun *et al.* 2023, Weideman *et al.* 2025). Among these methods, PhylEx (Jun *et al.* 2023) and Canopy2 (Weideman *et al.* 2025) use Bayesian methods to reconstruct tumor phylogenies from SNVs, while Cardelino (McCarthy *et al.* 2020) is primarily an scRNA-seq method but can optionally use bulk data given 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 has shown the importance of looking at these classes of variants together (Fu *et al.* 2022, Bristy *et al.* 2025), given their complementary value to 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 (Weber *et al.* 2025, Bristy *et al.* 2025).

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 unprecedented 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 (Eaton *et al.* 2018) that we have previously adapted to provide similarly comprehensive variant coverage from bulk DNA-seq (Fu *et al.* 2022) and scDNA-seq (Bristy *et al.* 2025) 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 scRNA-seq in resolving clonal evolution accurately and comprehensively. We further demonstrate with application to a previously published DNA/RNA breast cancer data set (Chung *et al.* 2017) that TUSV-INT is practical on real data and can yield improvements in clonal deconvolution and lineage reconstruction capable of driving novel biological discovery.

2 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 (Eaton *et al.* 2018) and variants (Fu *et al.* 2022, Bristy *et al.* 2025) 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.

2.1 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 (Gao *et al.* 2023) or SIGNALS (Funnell *et al.* 2022). Assume we are given: (i) m samples of bulk DNA-seq data covering r genomic segments, l SV breakpoints and g SNVs originating from n tumor clones and (ii) scRNA-seq data covering the same r genomic segments, originating from n_r tumor clones. Note that for computational efficiency, the variants used as input may be randomly subsampled from the full variant set (see Supplementary Methods, available as supplementary data at *Bioinformatics Advances* online for more details), a strategy for which we provide a more detailed quantitative justification in our prior work (Bristy *et al.* 2025). 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 $n \times n_r$ bulk clone to single-cell RNA seq clone mapping matrix M. We do that by solving for a constrained optimization problem minimizing the following objective function:

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

$|F - UC|$ 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 term 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

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

Notation	Interpretation	Type
n	The number of DNA clones	User defined
n_r	The number of RNA clones	Data derived
m	The number of samples	Data derived
l, g, r	The number of SV breakpoints, SNVs, and DNA copy number segments, respectively	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	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
$C^{\text{RNA}} \in \mathcal{Z}_{\geq 0}^{n_r \times (2r)}$	$C_{i,j}^{\text{RNA}}$: Clonal copy number of segment j in RNA clone i	Constant
$C' \in \mathcal{Z}_{\geq 0}^{n \times (2r)}$	$C'_{i,j}$: Clonal copy number of segment j in clone i of the phylogeny	Variable
$U \in \mathbb{R}_{\geq 0}^{m \times n}$	$U_{i,j}$: 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
$M \in \{0,1\}^{n \times n_r}$	$M_{i,j} = 1$ 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
$W \in \{0,1\}^{n \times n \times (l+g)}$	$W_{i,j,b} = 1$ iff breakpoint or SNV b is mapped on the edge from clone i to clone j	Variable
$D \in \{0,1\}^{(l+g)}$	$D_b = 1$ iff breakpoint or SNV b belongs to the first allele	Variable

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. Figure 1, available as supplementary data at *Bioinformatics Advances* online illustrates the overall problem using a toy example.

2.2 Algorithm

We formalize TUSV-INT as a constrained optimization problem and solve for it with a coordinate descent algorithm, iterating 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 (Eaton *et al.* 2018) and elaborated on in the more general TUSV-ext (Fu *et al.* 2022) 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. The overall analysis pipeline is illustrated in Fig. 2, available as supplementary data at *Bioinformatics Advances* online.

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: (i) inference of the segmental CNAs from single-cell RNA sequencing data by estimating M and C' (ii) estimation of the SNVs and SVs associated with each clone constrained by the ILP formulation. We describe these in more detail in the following sections, focusing particularly on the extensions beyond TUSV-ext. For a more complete exposition of portions of the optimization that are largely unchanged, we refer readers to the Supplementary Methods, available as supplementary data at *Bioinformatics Advances* online or to the prior works (Eaton *et al.* 2018, Fu *et al.* 2022).

2.2.1 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 [Equation (2)].

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

2.2.2 Estimating M and C'

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

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

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

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

We then define the RNA copy number distance portion of the objective in Equation (1) using L1 distance as [Equations (6–8)]. Here, Equation (8) holds because the ILP is minimizing $|C' - MC^{\text{RNA}}|$, allowing each $\mathcal{C}_{i,j}^{\Delta}$ to be equal to $|C'_{i,j} - (MC^{\text{RNA}})_{i,j}|$ for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, 2r\}$.

$$\mathcal{C}_{i,j}^{\Delta} \geq C'_{i,j} - \sum_{k=1}^{n_r} M_{i,k} . C_{k,j}^{\text{RNA}} \tag{6}$$

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

$$|C' - MC^{\mathrm{RNA}}| = \sum_{i=1}^{n}\sum_{j=1}^{2r} C^{\Delta}_{i,j} \tag{8}$$

Simultaneously, we set the estimated C' as the segmental CNA values of C [Equation (9)].

$$C_{i,l+g+j} = C'_{i,j} \forall i \in \{1,\ldots,n\}, j \in \{1,\ldots,2r\} \tag{9}$$

2.2.3 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 $i,j \in \{1,\ldots,n\}$. 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. We then impose the following constraints on the E matrix:

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

$$E_{i,n} = 0 \quad \forall i \in \{1,\ldots,n\} \tag{10}$$

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

Each internal node and the root has two outgoing edges, ensuring that the tree is binary and each non-root node has exactly one incoming edge:

$$\sum_{j=1}^{n} E_{i,j} = 2 \quad \forall i \in \{n'+1,\ldots,n\} \tag{12}$$

$$\sum_{i=n'+1}^{n} E_{i,j} = 1 \quad \forall j \in \{1,\ldots,n-1\} \tag{13}$$

There are no two-node cycles:

$$E_{i,j} + E_{j,i} \leq 1 \quad \forall i,j \in \{n'+1,\ldots,n\} \tag{14}$$

Additionally, following TUSV-ext (Fu *et al.* 2022), 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:

$$A_{i,j} \geq E_{i,j} \quad \forall i,j \in \{1,\ldots,n\} \tag{15}$$

$$A_{k,j} \geq A_{k,i} + E_{i,j} - 1 \quad \forall i,j \in \{1,\ldots,n\}, k \in \{1,\ldots,n\} \setminus \{i\} \tag{16}$$

$$A_{k,j} \leq A_{k,i} - E_{i,j} + 1 \quad \forall i,j \in \{1,\ldots,n\}, k \in \{1,\ldots,n\} \setminus \{i\} \tag{17}$$

2.2.4 Copy number constraints

We define similar constraints to TUSV-ext (Fu *et al.* 2022) 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.

$$C_{n,j} = 0 \quad \forall j \in \{1,\ldots,(l+g)\} \tag{18}$$

$$C_{n,j} = 1 \quad \forall j \in \{(l+g+1),\ldots,(l+g+2r)\} \tag{19}$$

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

$$C_{i,j} \leq c_{max} \forall i \in \{1,\ldots,n\}, j \in \{(l+g+1),\ldots,(l+g+2r)\} \tag{20}$$

Additional constraints define the phylogenetic cost R as the sum of L1 distances of copy number differences across the edges of the tree according to Equations (S3–S10), available as supplementary data at *Bioinformatics Advances* online.

2.2.5 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 (Fu *et al.* 2022). 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:

$$\begin{aligned} C_{i,b} &\leq \sum_{s=1}^{r} Q_{b,s} C_{i,l+g+s} + (1-D_b)c_{\max} \\ C_{i,b} &\leq \sum_{s=1}^{r} Q_{b,s} C_{i,l+g+r+s} + D_b c_{\max} \\ &\forall i \in \{1,\ldots,n\}, b \in \{1,\ldots,l\} \end{aligned} \tag{21}$$

Finally, our second regularization term S in the objective function [Equation (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 $\pi_{i,b}$.

$$\pi_{i,b} = \frac{F_{i,b}}{\sum_{j=1}^{r} Q_{b,j}(F_{i,l+g+j} + F_{i,l+g+r+j})} \tag{22}$$

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 (Fu *et al.* 2022), we include this minimization as a regularization term in the objective function to determine S, which are described in detail at the Supplementary Methods, available as supplementary data at *Bioinformatics Advances* online section.

2.3 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 (Fu *et al.* 2022), 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.

2.4 Availability

TUSV-INT is implemented in Python 2 and Python 3. It utilizes the licensed Gurobi optimizer for ILP solution. The code is available at https://github.com/CMUSchwartzLab/TUSV-INT.git.

3 Results

We first validated the accuracy and robustness of TUSV-INT using simulated data and assessed its performance relative to TUSV-ext and Canopy2 (Weideman *et al.* 2025). 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 (Weideman *et al.* 2025) 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 120×. 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.

3.1 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. The Supplementary Methods, available as supplementary data at *Bioinformatics Advances* online provide more detail on the procedure for generating simulated data. We ran TUSV-INT with the following parameter sets—(i) varying numbers of clones, $n \in \{3,5,7,9\}$, (ii) varying numbers of SNV mutations, $g \in \{25,50,100\}$ and (iii) varying numbers of bulk samples, $m \in \{1,2,5,10\}$—with each scenario having three iterations, three random restarts, 5000 s 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 (Fig. 1a, d, and 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 (Fig. 1b, e, and 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 (Fig. 1c, f, and i) generally led to improved solution quality by all measures, with the

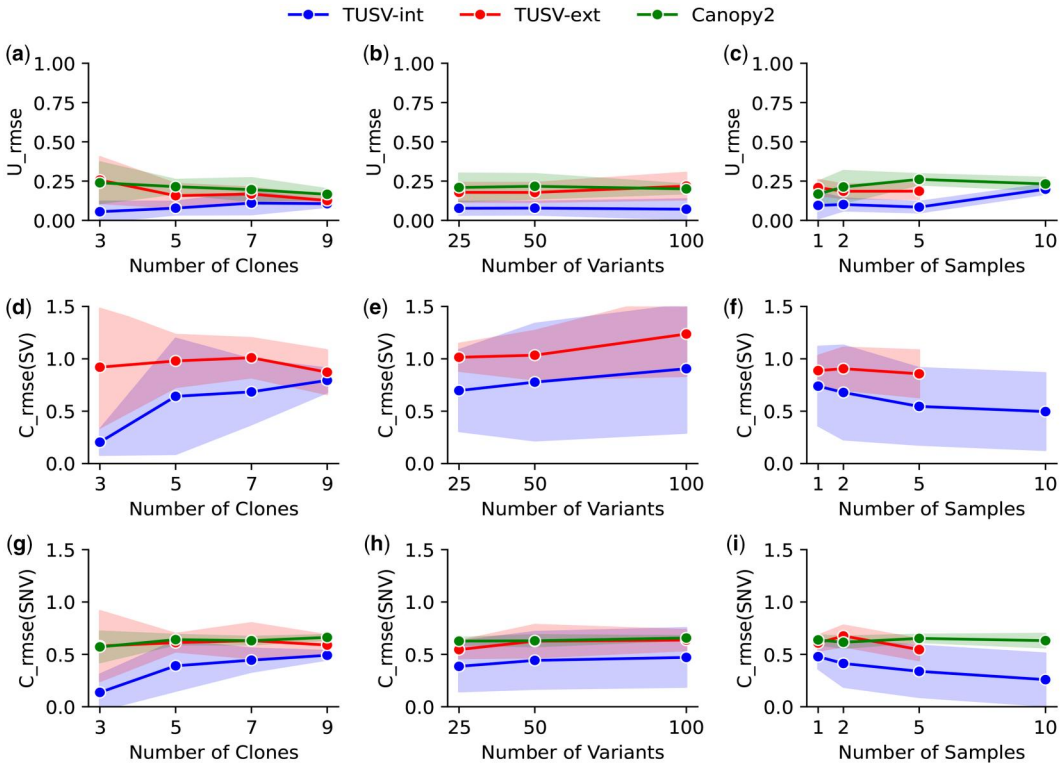


Figure 1. Results on simulation with different number of clones, samples and mutations. (a)–(c) Mean squared error of the true vs. estimated clonal mixture fractions (U) for TUSV-INT, TUSV-ext and Canopy2. (d)–(f) The mean squared error of the estimated SVs for TUSV-INT and TUSV-ext. (g)–(h) Mean squared 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 1 SD on either side of the mean. TUSV-INT is 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.

exception of the outlier of clonal frequency accuracy degrading for 10 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 with increasing sample numbers, although not clearly distinguishable from chance. TUSV-ext did not yield results for 10 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 (Weideman *et al.* 2025) 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. Figure 1 shows Canopy2's performance in green. Since Canopy2 does not make use of SVs, there are no Canopy2 results for Fig. 1d–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 that helps in correctly resolving the tree topology and inferring placement of all variants.

The Supplementary Results, available as supplementary data at *Bioinformatics Advances* online cover additional experiments testing functionality for inferring unobserved ancestral clones (Figs 3 and 4, available as supplementary data at *Bioinformatics Advances* online) and analyzing how runtime (Fig. 5, available as supplementary data at *Bioinformatics Advances* online) and convergence of the ILP (Fig. 6, available as supplementary data at *Bioinformatics Advances* online) vary with problem size.

3.2 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.* (2017). This dataset included three bulk WES samples collected from the primary tumor (Tumor), metastatic lymph node (LN_Tumor), 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 100× for the tumor and metastatic lymph node, and 50× for the paired blood sample.

We preprocessed and called variants from the bulk WES data using the Nfcore/sarek pipeline (Ewels *et al.* 2020, Garcia *et al.* 2020), 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 (Kim *et al.* 2018), and CNVkit (Talevich *et al.* 2016) 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 (Gao *et al.* 2023). Since Chung *et al.* (2017) generated the RNA sequencing data using the SMARTer Ultra Low RNA Kit (Chung *et al.* 2017), 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 five clones, three random restarts, three iterations, a time limit of 1500 s 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.

Figure 2 summarizes the results, comparing those between TUSV-INT and the DNA-only TUSV-ext. TUSV-INT inferred four clones (Fig. 2a and combining two samples of Fig. 2c) 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 (Fig. 2a). On the other hand, TUSV-ext inferred five clones from the two samples (Fig. 2b). The consistency between TUSV-INT and Canopy2 results provides indirect support for the value of scRNA-seq to better resolve clonal structure. Clone 0 in TUSV-INT tree and Clone 2 in TUSV-ext tree 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, e.g. 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 (Li *et al.* 2012, Sun *et al.* 2019). 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. 2c and d), although the differences in the inferred clones themselves and their phylogenetic placements complicate interpretation. The lower standard deviations suggest that the scRNA-seq data provides 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 0 to be the dominant clone in the metastasis, making the placement of clone 2 potentially

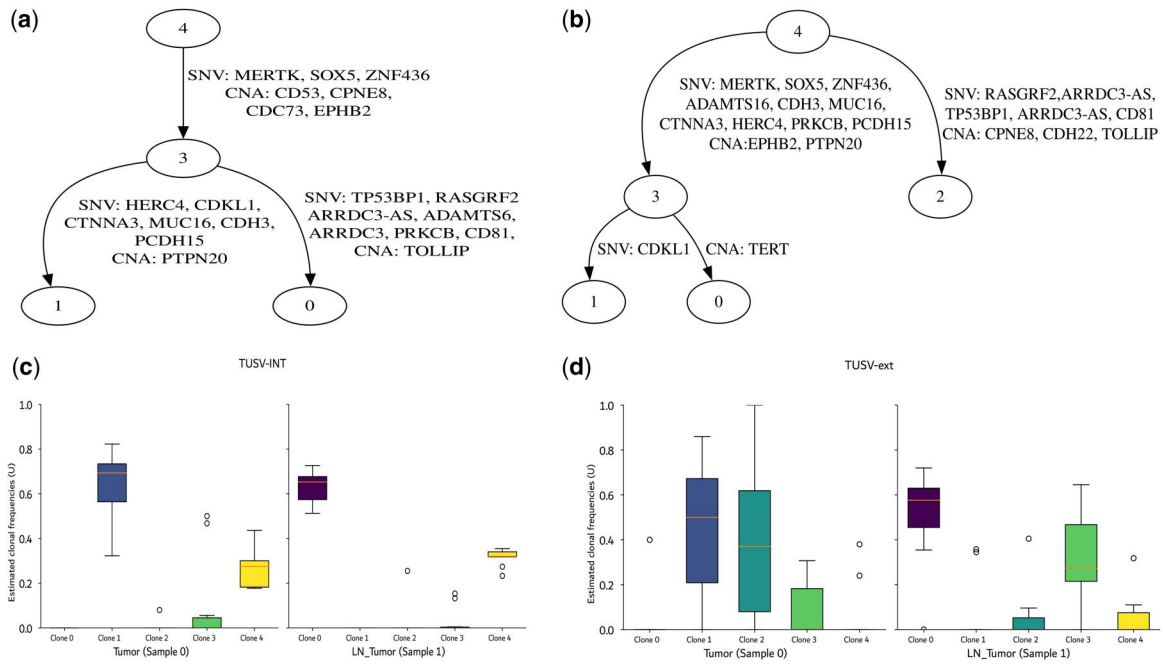


Figure 2. Results on ER/HER2 positive breast cancer patient BC03's data. (a) Clonal tree inferred by TUSV-INT with edges annotated by single-nucleotide variants (SNVs) and copy number alterations (CNAs). (b) Clonal tree inferred by TUSV-ext (Fu *et al.* 2022), also annotated with SNVs and CNAs. (c) Clonal 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, three iterations, three random restarts, and 1500 s per iteration.

important in correctly characterizing the pathway to metastatic progression in this patient. Note that clone 2 is assigned a small frequency in the TUSV-INT results despite not appearing in the TUSV-INT phylogeny of Fig. 2a because it was inferred in 1 of the 10 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.

4 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 (Jun *et al.* 2023, Weideman *et al.* 2025) 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. While TUSV-INT is designed to integrate bulk DNA-seq and scRNA-seq data, it is applicable to other multiomics settings as well. In particular, it can be readily applied to datasets combining bulk DNA-seq with scDNA-seq (Gawad *et al.* 2014), where clonal copy number profiles are derived from the scDNA-seq data. Our focus on scRNA-seq in this work is due to both its availability and the complexity of clonal inference from RNA-seq.

The work nonetheless leaves a number of avenues for improvement. Improving computational efficiency is 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 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.

Acknowledgements

We thank the Pittsburgh Supercomputing Center (PSC) and its ACCESS resource for providing compute resources used in the analyses described here. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

Nishat Anjum Bristy (Conceptualization [equal], Data curation [equal], Methodology [equal], Software [lead], Validation [equal], Writing—original draft [lead], Writing—review & editing [equal]) and Russell Schwartz (Conceptualization [equal], Funding acquisition [lead], Investigation [equal], Methodology [equal], Project administration [lead], Supervision [lead], Validation [Supporting], Visualization [Supporting], Writing—review & editing [equal])

Supplementary data

Supplementary data are available at *Bioinformatics Advances* online.

Conflict of interest

No competing interest is declared.

Funding

This work was supported by the National Human Genome Research Institute of the National Institutes of Health (R01HG010589). N.A.B. was supported in part by University of Pittsburgh Medical Center (UPMC) Enterprises through a Fellowship in Digital Health Innovation from Carnegie Mellon's Center for Machine Learning and Health (CMLH).

Data availability

Data original to this article are available along with the method source code at https://github.com/CMUSchwartzLab/TUSV-INT.

References

Bristy NA, Fu X, Schwartz R *et al.* Sc-TUSV-Ext: single-cell clonal lineage inference from single nucleotide variants, copy number alterations, and structural variants. *J Comput Biol* 2025;**32**:723–37.

Chung W, Eum HH, Lee H-O *et al.* Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat Commun* 2017;**8**:15081.

De S. Somatic mosaicism in healthy human tissues. *Trends Genet* 2011;**27**:217–23.

Deshwar AG, Vembu S, Yung CK *et al.* PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol* 2015;**16**:35–20.

Eaton J, Wang J, Schwartz R *et al.* Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics* 2018;**34**:i357–65.

Edrisi M, Huang X, Ogilvie HA *et al.* Accurate integration of single-cell DNA and RNA for analyzing intratumor heterogeneity using MaCroDNA. *Nat Commun* 2023;**14**:8262.

Evrony GD, Hinch AG, Luo C *et al.* Applications of single-cell DNA sequencing. *Annu Rev Genomics Hum Genet* 2021;**22**:171–97.

Ewels PA, Peltzer A, Fillinger S *et al.* The nf-core framework for community curated bioinformatics pipelines. *Nat Biotechnol* 2020;**38**:276–8.

Fu X, Lei H, Tao Y *et al.* Joint clustering of single-cell sequencing and fluorescence in situ hybridization data for reconstructing clonal heterogeneity in cancers. *J Comput Biol* 2021;**28**:1035–51.
Fu X, Lei H, Tao Y *et al.* Reconstructing tumor clonal lineage trees incorporating single-nucleotide variants, copy number alterations and structural variations. *Bioinformatics* 2022;**38**:i125–33.
Funnell T, O'Flanagan CH, Williams MJ *et al.*; IMAXT Consortium. Single-cell genomic variation induced by mutational processes in cancer. *Nature* 2022;**612**:106–15.
Gao T, Soldatov R, Sarkar H *et al.* Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes. *Nat Biotechnol* 2023;**41**:417–26.
Garcia M, Juhos S, Larsson M *et al.* Sarek: a portable workflow for whole genome sequencing analysis of germline and somatic variants. *F1000Res* 2020;**9**:63.
Gawad C, Koh W, Quake SR *et al.* Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc Natl Acad Sci USA* 2014;**111**:17947–52.
Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell* 2000; **100**:57–70.
Jovic D, Liang X, Zeng H *et al.* Single-cell RNA sequencing technologies and applications: a brief overview. *Clin Transl Med* 2022;**12**:e694.
Jun SH *et al.* Reconstructing clonal tree for phylophenotypic characterization of cancer using single-cell transcriptomics. *Nat Commun* 2023;**14**:982.
Kaufmann TL, Petkovic M, Watkins TBK *et al.* MEDICC2: whole-genome doubling aware copy-number phylogenies for cancer evolution. *Genome Biol* 2022;**23**:241.
El-Kebir M. SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics* 2018;**34**:i671–9.
Kim S, Scheffler K, Halpern AL *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* 2018;**15**:591–4.
Kurtenbach S, Cruz AM, Rodriguez DA *et al.* Uphyloplot2: visualizing phylogenetic trees from single-cell RNA-seq data. *BMC Genomics* 2021;**22**:419.
Lei H, Gertz EM, Schäffer AA *et al.* Tumor heterogeneity assessed by sequencing and fluorescence in situ hybridization (FISH) data. *Bioinformatics* 2021;**37**:4704–11.
Lei H, Guo XA, Tao Y *et al.* Semi-deconvolution of bulk and single-cell RNA-seq data with application to metastatic progression in breast cancer. *Bioinformatics* 2022;**38**:i386–94.
Lei H, Lyu B, Gertz EM *et al.* Tumor copy number deconvolution integrating bulk and single-cell sequencing data. *J Comput Biol* 2020; **27**:565–98.
Li X, Xu B, Moran MS *et al.* 53BP1 functions as a tumor suppressor in breast cancer via the inhibition of NF-κB through miR-146a. *Carcinogenesis* 2012;**33**:2593–600.
Maley CC. Multistage carcinogenesis in Barrett's esophagus. *Cancer Lett* 2007;**245**:22–32.
Malikic S, Jahn K, Kuipers J *et al.* Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nat Commun* 2019;**10**:2750.
McCarthy DJ, Rostom R, Huang Y *et al.* Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes. *Nat Methods* 2020;**17**:414–21.
McGranahan N, Swanton C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* 2017;**168**:613–28.
Moravec JC, Lanfear R, Spector D *et al.* Cancer phylogenetics using single cell RNA-seq data. *bioRxiv*, https://doi.org/10.1101/2021.01.07.425804, 2021, preprint: not peer reviewed.
Navin N, Kendall J, Troge J *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* 2011;**472**:90–4.
Nowell PC. The clonal evolution of tumor cell populations: acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression. *Science* 1976;**194**:23–8.
Pennington G, Smith CA, Shackney S *et al.* Reconstructing tumor phylogenies from heterogeneous single-cell data. *J Bioinform Comput Biol* 2007;**5**:407–27.
Satas G, Zaccaria S, Mon G *et al.* SCARLET: single-cell tumor phylogeny inference with copy-number constrained mutation losses. *Cell Syst* 2020;**10**:323–32.e8.
Schwartz R, Schäffer AA. The evolution of tumour phylogenetics: principles and practice. *Nat Rev Genet* 2017;**18**:213–29.
Schwarz RF, Trinh A, Sipos B *et al.* Phylogenetic quantification of intra tumour heterogeneity. *PLoS Comput Biol* 2014;**10**:e1003535.
Sottoriva A, Kang H, Ma Z *et al.* A big bang model of human colorectal tumor growth. *Nat Genet* 2015;**47**:209–16.
Sun C, Ban Y, Wang K *et al.* SOX5 promotes breast cancer proliferation and invasion by transactivation of EZH2. *Oncol Lett* 2019; **17**:2754–62.
Talevich E, Shain AH, Botton T *et al.* CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput Biol* 2016;**12**:e1004873.
Weber LL, Hart A, Ochoa H *et al.* Pharming joint clonal tree reconstruction of SNV and CNA evolution from single-cell DNA sequencing of tumors. *Lect Notes Comput Sci* 2025;**15647**:294–8.
Weideman AMK, Wang R, Ibrahim JG *et al.* Canopy2: tumor phylogeny inference by bulk DNA and single-cell RNA sequencing. *Stat Biosci* 2025;1–43.

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Bioinformatics Advances, 2025, 00, 1–9
https://doi.org/10.1093/bioadv/vbaf234
Original Article