JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 27, Number 4, 2020

Pp. 1–34
DOI: 10.1089/cmb.2019.0302

Research Article

Tumor Copy Number Deconvolution Integrating Bulk and Single-Cell Sequencing Data

HAOYUN LEI,[1] BOCHUAN LYU,[2] E. MICHAEL GERTZ,[3,4]
ALEJANDRO A. SCHÄFFER,[3,4] XULIAN SHI,[5] KUI WU,[5] GUIBO LI,[5]
LIQIN XU,[5] YONG HOU,[5] MICHAEL DEAN,[6] and RUSSELL SCHWARTZ[1,7]

ABSTRACT

Characterizing intratumor heterogeneity (ITH) is crucial to understanding cancer development, but it is hampered by limits of available data sources. Bulk DNA sequencing is the most common technology to assess ITH, but involves the analysis of a mixture of many genetically distinct cells in each sample, which must then be computationally deconvolved. Single-cell sequencing is a promising alternative, but its limitations—for example, high noise, difficulty scaling to large populations, technical artifacts, and large data sets—have so far made it impractical for studying cohorts of sufficient size to identify statistically robust features of tumor evolution. We have developed strategies for deconvolution and tumor phylogenetics combining limited amounts of bulk and single-cell data to gain some advantages of single-cell resolution with much lower cost, with specific focus on deconvolving genomic copy number data. We developed a mixed membership model for clonal deconvolution via non-negative matrix factorization balancing deconvolution quality with similarity to single-cell samples via an associated efficient coordinate descent algorithm. We then improve on that algorithm by integrating deconvolution with clonal phylogeny inference, using a mixed integer linear programming model to incorporate a minimum evolution phylogenetic tree cost in the problem objective. We demonstrate the effectiveness of these methods on semisimulated data of known ground truth, showing improved deconvolution accuracy relative to bulk data alone.

Keywords: cancer, copy number alteration (CNA), genomic deconvolution, heterogeneity, non-negative matrix factorization (NMF).

[1]Computational Biology Department, Carnegie Mellon University, Pittsburgh, Pennsylvania.
[2]Department of Mathematics, Rose-Hulman Institute of Technology, Terre Haute, Indiana.
[3]National Center for Biotechnology Information, U.S. National Institutes of Health, Bethesda, Maryland.
[4]Cancer Data Science Laboratory, National Cancer Institute, U.S. National Institutes of Health, Bethesda, Maryland.
[5]BGI-Shenzhen, Shenzhen, China.
[6]Laboratory of Translational Genomics, Division of Cancer Epidemiology & Genetics, National Cancer Institute, U.S. National Institutes of Health, Gaithersburg, Maryland.
[7]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, Pennsylvania.

1. INTRODUCTION

Cancer is one of the most lethal terminal diseases in the world, resulting, for example, in ~600,000 deaths in the United States in the past year (Siegel et al., 2017). Nevertheless, the age-adjusted rate of cancer deaths in the United States has been declining, partly due to the development of new cancer treatments. Recent work in cancer therapeutics is based on the notion of personalized or precision medicine (Coyne et al., 2017) to target driver alterations in specific cancer genes. Such targeted treatments have shown success in prolonging life but rarely lead to durable cures (Fisher et al., 2013), largely because tumors are not normally static or homogeneous entities (Dexter and Leith, 1986).

Most cancers exhibit phenotypes of hypermutability (Loeb, 2001) that result in a process of continuing evolution of clonal populations of tumor cells (Nowell, 1976), creating the opportunity for continuing acquisition of adaptive mutations as well as putatively selectively neutral genetic variants (Williams et al., 2016). As a consequence, different cells in the same tumor may acquire distinct sets of somatic alterations, including single-nucleotide variants (SNVs), copy number alterations (CNAs), and structural variations such as gene fusions or chromosomal rearrangements. This phenomenon, called intratumor heterogeneity (ITH) (Marusyk and Polyak, 2010), allows tumors to develop resistance to targeted treatments, as treatment-resistant subclones emerge within the tumor (Nowell, 1976; Fisher et al., 2013) or expand from initially rare subpopulations within the tumor's clonal diversity. Considerable recent research into the molecular mechanisms of cancer has concentrated on characterizing ITH and reconstructing the processes of clonal evolution by which it develops across tumor progression (Schwartz and Schäffer, 2017).

Currently, the most common technology to profile ITH is bulk DNA sequencing, which allows one to observe aggregate genetic variation in tumors and, when available, matched normal tissue from the same patients. Bulk DNA sequencing allows one to identify reasonably common genetic lesions and estimate their variant allele fractions (VAFs). Resolving these VAFs into models of clonal heterogeneity, however, requires solving a challenging computational inference problem, known as *genomic deconvolution*, which strives to explain VAFs as mixtures of unobserved clonal sequences occurring at varying frequencies within the tumor. These methods have limited accuracy and resolution, particularly with respect to rare clonal subpopulations (Barber et al., 2015), and reveal far less clonal heterogeneity than is evident from direct single-cell analysis (Navin et al., 2011; Heselmeyer-Haddad et al., 2012). Genomic deconvolution is particularly challenging in cancers exhibiting CNAs (Tolliver et al., 2010), a significant limitation given that CNAs are the primary mechanism of functional adaptation in at least some cancer types (Zack et al., 2013; Macintyre et al., 2018) and that CNAs at specific loci can have important consequences for treatment outcome (McGranahan et al., 2017).

Single-cell sequencing (SCS) has emerged as an alternative allowing for the direct inference of clonal genotypes (Navin et al., 2011). SCS itself is limited by difficult technical artifacts, such as the phenomenon of allelic dropout (Hou et al., 2012) and distortion of copy numbers due to the amplification steps used in most SCS methods to date (Lei et al., 2015). Moreover, SCS is relatively costly in comparison with bulk sequencing. As a result, SCS studies to date have involved only small sample sizes (Ortega et al., 2017).

The trade-offs between bulk sequencing and SCS have recently led to the idea that they might be combined to reconstruct ITH with both accuracy and scale (Malikic et al., 2017, 2018), yielding improved performance in bulk data deconvolution and relative to using SCS data alone. To date, though, such work has focused on SNVs specifically. There is substantial value in developing comparable methods for CNAs, given their biological importance, the greater difficulty of CNA deconvolution, and their suitability for phylogenetics from low-coverage SCS (Navin et al., 2011).

In this work, we develop methods for combining bulk and single-cell data to characterize ITH by CNAs specifically, both as a stand-alone inference and together with phylogenetic inference on clonal subpopulations. We pose the problem of inferring the tumor subpopulations and their representation across genomic samples using a variant of non-negative matrix factorization (NMF). We seek solutions that deconvolve bulk data while achieving consistency between inferred single cells and limited SCS data. We consider two problem variants, one minimizing genomic distance between SCS-observed single cells and inferred clones and the other explicitly incorporating a tumor phylogeny model to favor solutions that yield parsimonious evolution models relating observed cells and inferred clones. We characterize performance of the methods on semisimulated data generated from low-coverage SCS. We show that both methods are effective at improving clonal deconvolution of CNAs with limited amounts of SCS data, with increasing

accuracy as the number of genomic samples grows. We further show that explicitly modeling clonal evolution notably improves accuracy, suggesting the value of accounting for the process of tumor evolution in characterizing clonal structure.

2. METHODS

2.1. NMF deconvolution model

As in previous work of Schwartz and Shackney (2010), we formalize the generic problem of genomic deconvolution in terms of a mixed membership model, but here relating bulk and SCS data. We focus here specifically on deconvolution of copy number data, as in Tolliver et al. (2010) and Zaccaria et al. (2018), which we assume is profiled on a set of m genomic regions. In the pure deconvolution problem, we assume a set of n bulk samples, which might correspond to measurements from distinct tumor sites or regions in one patient. These bulk samples are collectively encoded in an $m \times n$ matrix $\boldsymbol{B}$, where element b_{ij} corresponds to the mean copy number of locus i in sample j. Our goal is to identify an $m \times k$ matrix of mixture components $\boldsymbol{C}$, representing copy numbers of inferred common clones, and a $k \times n$ matrix of mixture fractions, $\boldsymbol{F}$, describing the degree to which each column of $\boldsymbol{C}$ is represented in each column of $\boldsymbol{B}$. $\boldsymbol{B}$ is presumed to be approximated by the product of $\boldsymbol{C}$ and $\boldsymbol{F}$. We seek to minimize the deviation between $\boldsymbol{B}$ and $C \times \boldsymbol{F}$ by some measure, such as the Frobenius norm. With the additional constraints that $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{F}$ are non-negative, the problem is known as NMF (Wang and Zhang, 2013). More formally, we seek the following:

$$\min_{C,F} \quad ||\boldsymbol{B}-\boldsymbol{CF}||_{\mathrm{Fr}}^2 \tag{1}$$

where $||\cdot||_{\mathrm{Fr}}$ is the Frobenius norm of the matrix, $||\boldsymbol{B}-\boldsymbol{CF}||_{\mathrm{Fr}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(b_{ij}-\sum_{\ell=1}^{k} c_{i\ell}\cdot f_{\ell j}\right)^2}$, subject to the constraints $f_{\ell j} \geq 0,\ \forall \ell \in \{1,\ \ldots, k\},\ j \in \{1,\ \ldots, n\};\ \sum_{\ell=1}^{k} f_{\ell j} = 1,\ \forall j \in \{1,\ \ldots, n\};\ \ c_{i\ell} \in \mathbb{N}_0,\ \forall i \in \{1,\ \ldots, m\},\ \ell \in \{1,\ \ldots, k\}$.

This optimization problem is nonconvex, but prior work showed that the Euclidean distance between $\boldsymbol{B}$ and $\boldsymbol{CF}$ is nonincreasing under the following multiplicative update rules in Lee and Seung (2001):

$$f_{\ell j} \leftarrow f_{\ell j}\frac{(\boldsymbol{C}^T\boldsymbol{B})_{\ell j}}{(\boldsymbol{C}^T\boldsymbol{CF})_{\ell j}}, \qquad c_{i\ell} \leftarrow c_{i\ell}\frac{(\boldsymbol{BF}^T)_{i\ell}}{(\boldsymbol{CFF}^T)_{i\ell}},$$

providing formulas for iterative local optimization by fixing $\boldsymbol{C}$ or $\boldsymbol{F}$ on alternate steps. In practice, we modify this process heuristically to renormalize columns of F after each iteration to ensure they add to 1. Since this heuristic might undermine the guarantee of monotonicity, we manually verify that $||\boldsymbol{B}-\boldsymbol{CF}||_{\mathrm{Fr}}$ decreases on each iteration, terminating the optimization if it fails to yield continuing improvements. More details are provided in Section 5.

Figure 1 provides an illustrative example of the deconvolution model. Suppose we have a possible $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{F}$. The two data points B_1 and B_2 represent bulk tumor samples combining three mixture components C_1, C_2, and C_3. For ease of illustration, we assume data are assayed on the copy numbers of just two genomic loci, G_1 and G_2. The matrix $\boldsymbol{B}$ represents the average copy numbers of G_1 and G_2 in the bulk tumor samples B_1 and B_2. In each component of C, the copy numbers should be integers, but since the bulk tumors are weighted mixtures of components, the values in $\boldsymbol{B}$ need not be integers. The matrix $\boldsymbol{F}$ represents the fractional weights used to generate B_1 and B_2 from the pure components in $\boldsymbol{C}$. For example, the first column of $\boldsymbol{F}$ indicates that B_1 is a mixture of equal parts of C_1 and C_2. This relationship can be expressed via matrix multiplication, $\boldsymbol{B} = \boldsymbol{CF}$, as shown in the right part of Figure 1.

2.2. Extending NMF with SCS data

The multiplicative update algorithm is a standard method for the pure NMF optimization problem, provided the number of samples n is large compared with the intrinsic dimension k of the mixture. We would, however, expect it to perform poorly for our problem, in part, because real tumor data generally include few samples per patient and, in part, because deconvolution of copy numbers is an underdetermined problem. We sought to improve the optimization by biasing the objective function to favor inferred clones

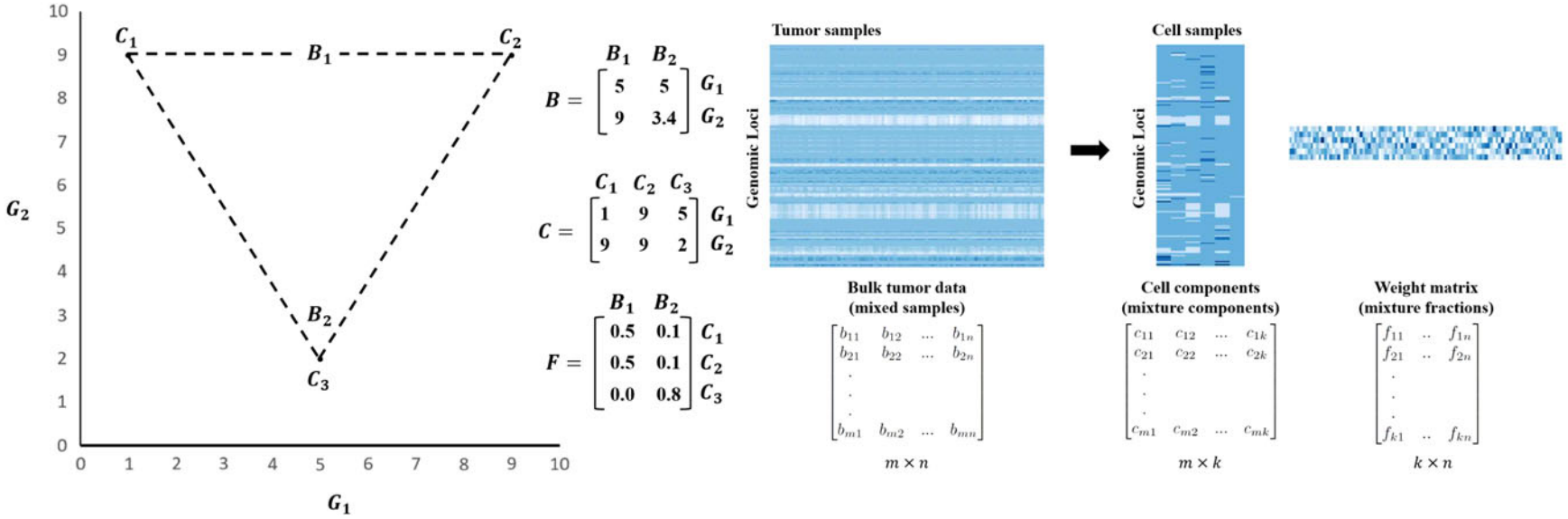

FIG. 1. Illustration of the mathematical formulation for the mixed membership modeling problem. The model implies that each entry of $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{F}$ is non-negative, each entry of $\boldsymbol{C}$ is integer, and each column of $\boldsymbol{F}$ must sum up to 1.

similar to the observed SCS data via an auxiliary penalty in the objective function, similar to the approach of Malikic et al. (2017, 2018). Intuitively, we assume that the inferred clones ($\boldsymbol{C}$) should be closely related to one or more of the observed single cells, which we call observed cell components ($\boldsymbol{C}^{(\text{observed})}$). While any given single cell may not exactly match a consensus clone, we propose that the method will be able to approximately infer mixture components reflecting dominant clones by balancing quality of deconvolution against similarity to observed single cells. We quantify this intuition using the Euclidean distance between the inferred clones and observed cells, introducing a regularization parameter α to balance the weight of this penalty relative to the prior cost based on deconvolution quality. The resulting combined objective appears as Equation (2):

$$\min_{C,F} \quad ||\boldsymbol{B}-\boldsymbol{C}\boldsymbol{F}||_{Fr}^2+\frac{1}{2}\alpha||\boldsymbol{C}-\boldsymbol{C}^{(\text{observed})}||_{Fr}^2 \tag{2}$$

which we optimize subject to the constraints $f_{\ell j}\geq 0,\ \forall \ell\in\{1,\ \ldots,k\},\ j\in\{1,\ \ldots,n\};\ \sum_{\ell=1}^{k} f_{\ell j}=1,\ \forall j\in\{1,\ \ldots,n\};\ c_{i\ell}\in\mathbb{N}_0,\ \forall i\in\{1,\ \ldots,m\},\ \ell\in\{1,\ \ldots,k\}$.

We solve for the revised model through an extension of the iterative update algorithm in Lee and Seung (2001) and Berry et al. (2007):

$$f_{\ell j}\leftarrow f_{\ell j}\frac{(\boldsymbol{C}^T\boldsymbol{B})_{\ell j}}{(\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{F})_{\ell j}},\qquad c_{i\ell}\leftarrow c_{i\ell}\frac{(\boldsymbol{B}\boldsymbol{F}^T)_{i\ell}}{(\boldsymbol{C}\boldsymbol{F}\boldsymbol{F}^T+\alpha(\boldsymbol{C}-\boldsymbol{C}^{(\text{observed})}))_{i\ell}},$$

adding the constraints on $\boldsymbol{C}$ and $\boldsymbol{F}$ to the update rules in Thurau et al. (2009). We further heuristically improve on the standard practice of random initialization by initializing the cell component matrix C with true SCS data. Pseudocode for the complete algorithm is provided in Section 5 as Algorithm 1. Collectively, these additions to the pure NMF iterative update algorithm constitute our first approach to integrating SCS data for improved deconvolution of CNAs from bulk DNA-seq, which we dub our phylogeny-free method.

2.3. Extending the NMF model with a single-cell phylogeny objective

We next developed an alternative phylogeny-based approach, seeking to deconvolve the bulk data into clonal subpopulations while simultaneously inferring a phylogeny on those deconvolved clones, similar to the SNV PHiSCS method of Malikic et al. (2018). Intuitively, evolutionary distance provides a more biologically motivated measure of what we mean in asserting that inferred single cells should be similar to observed single cells. As with the phylogeny-free method, we would expect that any small sample of single cells will not exactly reflect the spectrum of dominant clones, but that the method will be able to approximately infer dominant clones by balancing deconvolution quality against evolutionary distance of mixture components to observed single cells. This approach trades off a more principled measure of solution quality for a harder optimization problem.

We quantify phylogenetic distance as the minimum over evolutionary trees incorporating both observed single cells and inferred clones of the L_1 distance between copy number vectors describing each tree edge. Let $\boldsymbol{C}^*=[\boldsymbol{C},\boldsymbol{C}^{(\text{observed})}]$ be an $m\times k^*$ matrix consisting of columns representing inferred clonal copy numbers followed by columns representing the copy numbers of the observed cells. Let c_u^* denote column u of $\boldsymbol{C}^*$. We introduce a $k^*\times k^*$ matrix of binary variables $\boldsymbol{S}$. A value of $s_{uv}=1$ indicates the existence of a directed edge from node u to node v, and a value $s_{uv}=0$ indicates the absence of such a edge; we set $s_{uu}=0$ to avoid self-loops. In other words, $\boldsymbol{S}$ is an adjacency matrix for a directed graph; in the full formulation [Appendix 1 (Supplementary Methods)], we introduce constraints that ensure the graph is a tree. We define our measure of tree cost to be

$$J(\boldsymbol{S},\boldsymbol{C},\boldsymbol{C}^{(\text{observed})})=\sum_{u=1}^{k^*}\sum_{v=1}^{k^*}s_{uv}\cdot\ \|\ c_u^*-c_v^*\|_1. \tag{3}$$

Intuitively, $J(\boldsymbol{S},\boldsymbol{C},\boldsymbol{C}^{(\text{observed})})$ is a form of minimum evolution model on a phylogeny defined by $\boldsymbol{S}$. While there are more sophisticated and realistic models for CNA distance (Chowdhury et al., 2014, 2015; El-Kebir et al. 2017), we favored L_1 distance here as a tractable approximation easily incorporated into the overall integer linear programming (ILP) framework. Similarly, while there are now a number of sophisticated methods available specifically for phylogenetics of single-cell sequences (c.f., Kuipers et al., 2017), these are

largely focused on SNV rather than CNA phylogenetics (Jahn et al., 2016; Ross and Markowetz, 2016; Zafar et al., 2017) with limited exceptions (Wang et al., 2014; Subramanian and Schwartz, 2015).

More specifically, we modify the NMF objective function as follows:

$$\min_{C,F,S}\left(||\boldsymbol{B}-\boldsymbol{CF}||_1+\beta\cdot J(\boldsymbol{S},\boldsymbol{C},\boldsymbol{C}^{(\text{observed})})\right), \tag{4}$$

where $||\boldsymbol{B}-\boldsymbol{CF}||_1=\sum_{i=1}^{m}\sum_{j=1}^{n}\left|b_{ij}-\sum_{\ell=1}^{k}c_{i\ell}\cdot f_{\ell j}\right|$, and β is a regularization parameter to balance deconvolution quality against parsimony of the evolutionary model. The norm $||\cdot||_1$ is the element-wise L_1 matrix norm, that is, the sum of the absolute values of matrix elements, rather than the induced L_1 matrix norm for which the same notation is sometimes used. These are optimized subject to the same constraints as in the previous formulations: $f_{\ell j}\geq 0,\ \forall\ell\in\{1,\ \ldots,k\},\ j\in\{1,\ \ldots,n\};\ \sum_{\ell=1}^{k}f_{\ell j}=1,\ \forall j\in\{1,\ \ldots,n\};$ $c_{i\ell}\in\mathbb{N},\ \forall i\in\{1,\ \ldots,m\},\ \ell\in\{1,\ \ldots,k\}$.

The discrete tree optimization term lacks an analytic expression and hence does not lend itself to the prior iterative update strategy. We therefore use a different computational strategy based on ILP to replace the linear algebra steps of the method in Lee and Seung (2001), similar to other recent works in joint deconvolution and phylogenetics in Zaccaria et al. (2018) and Eaton et al. (2018).

For this optimization problem, we use an iterative coordinate descent approach. There are three sets of variables over which to optimize: the weight matrix $\boldsymbol{F}$, the tree structure $\boldsymbol{S}$, and the inferred copy numbers $\boldsymbol{C}$. We solve for variables $\boldsymbol{F}$, $\boldsymbol{S}$, and $\boldsymbol{C}$ alternately, in this order, while holding all other variables as constant. The iterative coordinate descent continues until the decrease between successive values of $\boldsymbol{C}$ falls below some threshold. To initialize $\boldsymbol{C}$, we used observed single-cell data. Whenever two of the three sets of variables are held constant, the resulting optimization problems can each be expressed as either a linear program (LP) or an ILP.

When certain subsets of the variables are fixed, the resulting LP or ILP may be simplified. When solving for $\boldsymbol{F}$ with fixed values of $\boldsymbol{S}$ and $\boldsymbol{C}$, the term $J(\boldsymbol{S},\boldsymbol{C},\boldsymbol{C}^{(\text{observed})})$ is constant and the value of $\boldsymbol{S}$ is irrelevant. Similarly, when solving for $\boldsymbol{S}$ for fixed values of $\boldsymbol{F}$ and $\boldsymbol{C}$, the term $||\boldsymbol{B}-\boldsymbol{CF}||_1$ is constant and therefore $\boldsymbol{F}$ is irrelevant. The optimal value of $\boldsymbol{C}$ for fixed values of $\boldsymbol{F}$ and $\boldsymbol{S}$, however, depends both on $\boldsymbol{F}$ and $\boldsymbol{S}$. In the limit of using no single-cell data, our problem statement and method are similar to that of Zaccaria et al. (2018) for incorporating tree mixtures into purely bulk CNA deconvolution.

We solve for $\boldsymbol{S}$ via an ILP that uses a flow model to constrain solutions to a minimum evolution tree, adapting a similar ILP method originally developed for finding maximum parsimony character-based phylogenies (Sridhar et al., 2007). Intuitively, the model forces a tree structure by setting up a flow from an arbitrary root to each other clone in the tree and minimizing the cost of edges needed to accommodate all such flows. The full ILP is described in Section 5.

2.4. Validation via observed single-cell data

To validate the method, we require bulk data for which clone copy number vectors and frequencies are known. As this is unavailable for any real data set, we use semisimulated data generated from copy number calls (Baslan et al., 2012) from real SCS data from two human glioblastoma cases of Wu et al. (2016). The full single-cell data set consists of low-depth SCS DNA-seq used to establish mean copy numbers at 9934 genomic positions throughout the genome, at intervals of $\sim$40 kbp. Each tumor was subdivided into three regions (i.e., samples), with each single cell labeled by its region (1, 2, or 3) of origin. We used these true SCS CNA data to generate a series of synthetic bulk data sets, simulating either one, two, or three bulk samples from each region for a total of three, six, or nine bulk samples per trial.

Each simulated sample is generated by sampling two dominant cells from a region to represent major clones, 23 other cells from the same region to represent minor clones, and 50 cells from the other regions to represent contamination, which are mixed with Dirichlet-sampled proportions with weight parameters for dominant, minor, and contaminant clones in the ratio 10 to 0.1 to 0.01. We then assessed our ability to deconvolve the bulk data across a range of regularization parameter values and random replicates of the chosen single cells. We assessed accuracy by the fraction of genomic positions assigned correct copy number and by the root mean square deviation (RMSD) between true and inferred clones and mixture fractions. Figure 2 summarizes the overall experimental design, which is described in more detail in Section 5.4. This design treats observed SCS as the ground truth, allowing us to ignore the problem of doublet cells that typically must be addressed with SCS data. We would normally require that likely

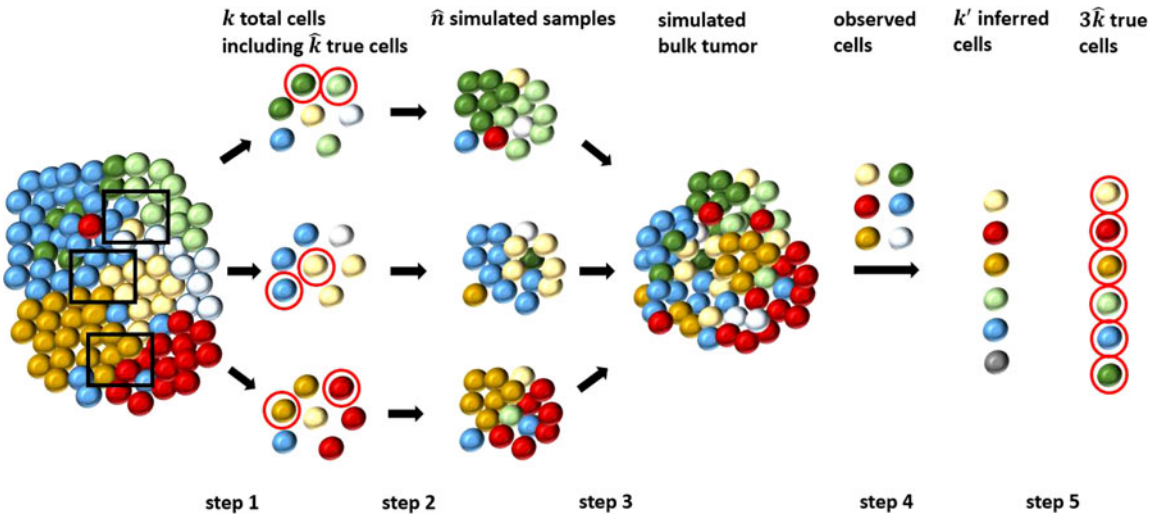

FIG. 2. Workflow for the simulation and validation. We separate the whole process into five main steps: in step 1, we randomly chose k total single cells from each region (indicated by the black frames), where we can pick $\hat{k}$ dominant clones (indicated by red circles, also called true cells); in step 2, we simulated $\hat{n}$ tumor samples from each region using the k cells; in step 3, we combined the $\hat{n}$ tumor samples to get a simulated bulk tumor; in step 4, we deconvolved the bulk tumor integrating observed cells to get $k' = 3\hat{k}$ inferred clones; and in step 5, we assessed the performance using the k' inferred clones and $3\hat{k}$ true cells.

doublets be removed from SCS data in preprocessing before applying our method. This design also does not explicitly include calling CNA markers on bulk data, itself a hard problem that would need to be performed in preprocessing before applying our method.

We were unable to identify any competitive tool for bulk deconvolution of purely CNA data applicable to small numbers of bulk samples and for which software is publicly available. We therefore compare our methods to standard NMF, as implemented by our code with zero regularization parameters.

2.5. *Real bulk sequencing data*

In addition to the semisynthetic data, we applied our methods to true bulk tumor data from the same human glioblastoma cases described in Section 2.4. One bulk sample was available for each region for each case. Copy number estimation was performed by the Complete Genomics Standard pipeline, developed by the Beijing Genomics Institute (BGI), to establish mean copy numbers for 931 regions throughout the genome. Single-cell CNA estimation had been performed on the same tumor regions to obtain estimated copy numbers at 9934 genomic locations. We treated the most extreme locations on each of the 24 chromosomes (includes Y) as chromosome ends, and hence, we interpret the data as copy numbers at 9910 (=9934–24) intervals. These SCS copy number intervals were always smaller than the genomic intervals obtained from the bulk data.

Because the intervals in the SCS and bulk data did not have their endpoints in common, we subdivided the bulk data into 9910 intervals aligned to the same endpoints as in the SCS intervals by applying two rules: (1) for each SCS interval that was wholly contained in a bulk interval, we created a new interval that had the same coordinates as the SCS interval, and assigned this new interval the copy number of the bulk interval; (2) for each SCS interval that overlapped two bulk intervals, A and B, we created a new interval for the bulk data, again having the same endpoints as the SCS interval, but with a copy number that is a weighted sum of the copy number of interval A and the copy number of interval B. The weights were chosen to be proportional to the number of nucleotides in the overlap between the new interval and A and B. This process yielded SCS and bulk data with copy numbers assigned to common genomic intervals (Section 5.7). We applied our deconvolution and phylogeny methods as with the semisynthetic data to the collection of three bulk genomic samples corresponding to the three tumor regions.

2.6. *Implementation*

The methods described in Section 2 (Methods) and refined below were all implemented in Python3, using Gurobi. One practical change from the formulation above is that we replaced the theoretical $f_{\ell j} \geq 0$ with

$f_{\ell j} > 10^{-4}$ to avoid having the f values trapped at 0. The observed human subjects' data cannot be redistributed, but code for the methods is available along with artificial data on GitHub (https://github.com/CMUSchwartzLab/SCS_deconvolution).

3. RESULTS

3.1. Phylogeny-free method

We first assessed the accuracy of the phylogeny-free method relative to pure NMF and simple heuristic improvements. Figure 3 provides a summary of accuracy and RMSD for inference of true SCS components via the method of Section 2.2 for variations in the number of tumor samples (3, 6, 9) and regularization parameter $\alpha(0-1)$ over 40 replicates per condition. To provide a baseline for comparison, each plot provides equivalent accuracy measures for NMF (Lee and Seung, 2001, i.e., Algorithm 1 with $\alpha = 0$) with random initial integer valued $\boldsymbol{C}$ (red dashed line in Fig. 3) and with the proposed solution that all copy numbers have the normal value of 2, which we call the ''all-diploid baseline'' (black dashed line in Figs. 3 and 12). In each case, the bulk data are simulated from $k' = 6$ fundamental cell components (2 out of a random 25 cells selected in each region).

Pure NMF with random initialization performed poorly, which is unsurprising since NMF on CNA data is an underdetermined problem, although the simple heuristic of biasing the search toward biologically plausible solutions by initializing with real SCS data improves accuracy. Bringing true SCS data into the objective function yielded modest improvements in accuracy over using SCS data solely for initialization for at least some values of the regularization parameter. The phylogeny-free method with $\alpha = 0$ corresponds to pure NMF initialized with true SCS data, and this performed slightly worse than the all-diploid baseline solution. Modestly increasing α led to some improvement in accuracy, but above some value, α put too much weight on similarity to the observed SCS data and too little weight on quality of the deconvolution, giving worse overall results. The best value of α depended on sample size, which we attribute again to NMF being underdetermined if the number of desired components is larger than the number of samples. The plots suggest that the method is fairly robust to α if the number of samples exceeds the intrinsic dimension of the data (six), but that SCS data can overcome that limit for small numbers of samples with a well-tuned regularization term. Additional results in Section 6 show minimal additional improvement even with unrealistically large sample sizes (Fig. 13), and also show that the performance is consistent across individual inferred clones (Fig. 14).

Figure 4 provides an illustrative example of performance for a single selected clone inferred from three, six, or nine samples, intended to demonstrate kinds of errors the method tends to produce. We chose the one inferred clone with smallest RMSD for each sample size to simplify visual inspection. We see that at least in these high-quality cases, the distributions of copy numbers are similar for the inferred and true clones. For loci at or just above diploid, the modified NMF can usually infer the exact copy number. Where errors occur, they tend to be in loci with large (5–10) or smaller copy numbers (0–1).

3.2. Phylogeny-based method

We next examined results of the phylogeny-based method of Section 2.3 under the same conditions used to assess the phylogeny-free method. Figure 5 summarizes average accuracy and RMSD as a function of regularization parameter β. The figure compares the results of pure NMF with the all-diploid baseline. Setting $\beta = 0$ provides poor performance, substantially below the all-diploid baseline solution. Making $\beta = 0$ for Figure 5 represents the same optimization problem as $\alpha = 0$ for Figure 3, but solved by the coordinate descent method we developed to accommodate the ILP phylogeny objective rather than by the modified iterative update algorithm with the simpler L_2 objective. Figure 5 thus suggests that the new coordinate descent method is less effective at pure NMF than is the prior iterative update algorithm. Despite that observation, the results on $\beta \geq 0.2$ show substantially better accuracy than was achieved by pure NMF or the phylogeny-free algorithm. Furthermore, the results appear robust to variation in β across the range examined. Results in Section 6 distinguishing accuracy across cells (Fig. 15) support the robustness of the phylogeny-based method to a range of β values in cell-to-cell inferences.

Figure 6 shows copy numbers for a single minimum-RMSD pair for inferred and true clones for each number of samples, again to visualize the nature of inference errors. The results again show exact fitting for

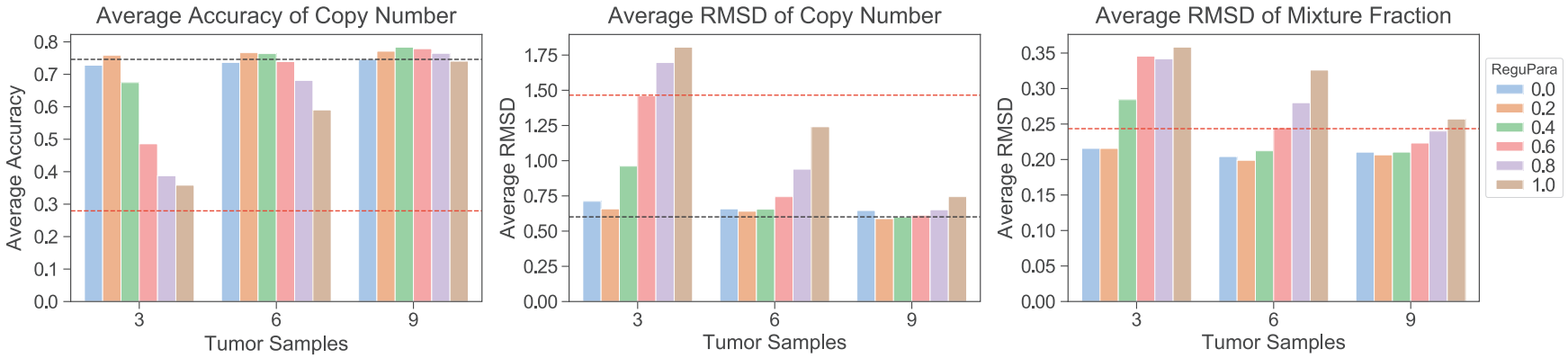

FIG. 3. Accuracy and RMSD of the phylogeny-free method as functions of tumor samples and regularization parameter. The red dashed line shows average overall accuracy (left panel) or RMSD (center and right panels) of NMF with random initialization. The black dashed line shows the performance of the all-diploid baseline solution. Since we cannot resolve mixture fractions for an all-diploid solution, we omit it from the mixture fraction results. Different bars show performance as a function of regularization parameter α of Equation (2) from 0.0 to 1.0 in increments of 0.2. The *X*-axis shows the number of tumor samples, and the *Y*-axis the average accuracy or RMSD. NMF, non-negative matrix factorization; RMSD, root mean square deviation.

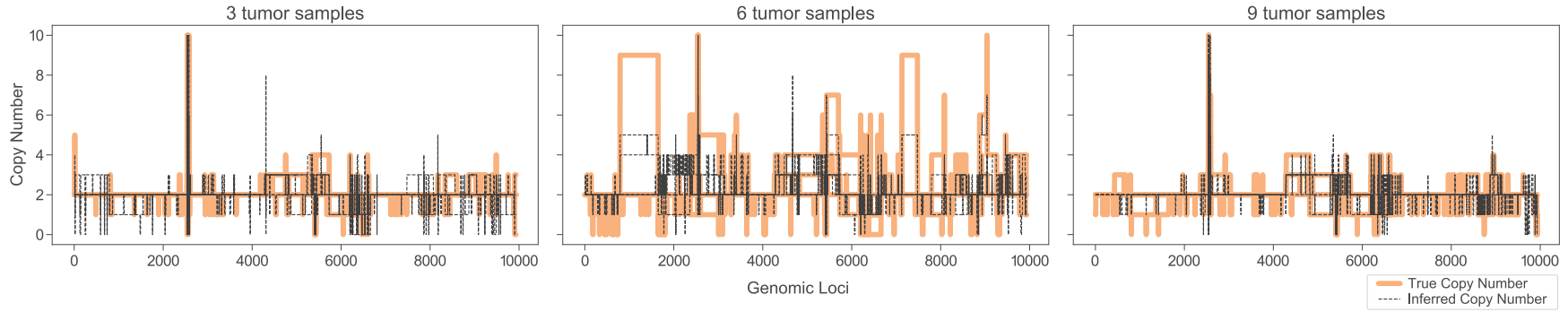

FIG. 4. Visualization of copy number as a function of genomic locus for single examples of inferred and true clones for the phylogeny-free method for three, six, and nine samples. The figure uses the minimum-RMSD pair for each case. The black dashed line shows the copy number inferred by modified NMF, and the orange bar shows the true copy number in that position.

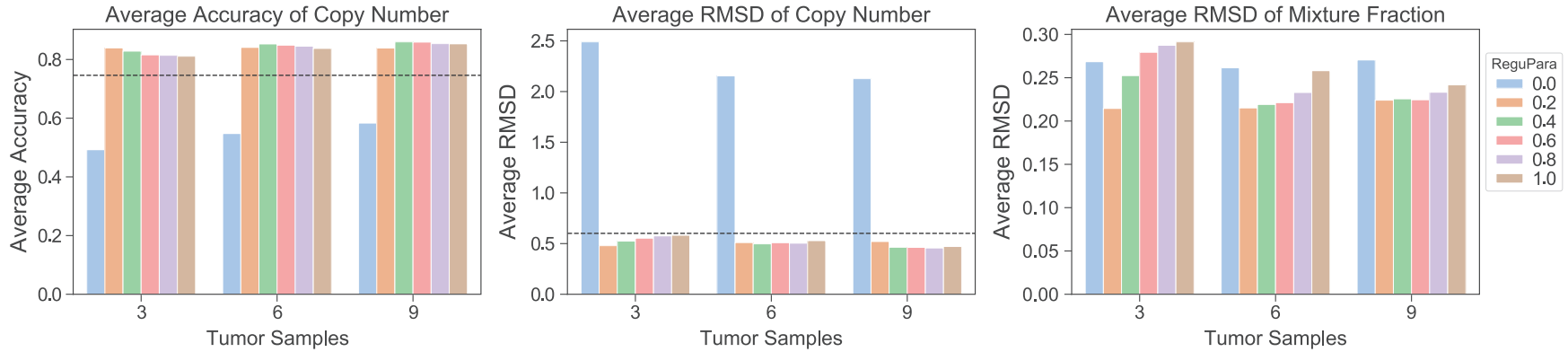

FIG. 5. Average accuracy and RMSD for the phylogeny-based method as functions of tumor samples and regularization parameter. The left panel shows the average accuracy of inferred copy numbers, the center panel average RMSD between inferred and true copy numbers, and the right panel average RMSD between the inferred and true mixture fractions. The black dashed line shows the performance of the all-diploid baseline solution. Since we cannot resolve mixture fractions for an all-diploid baseline, we omit it from the mixture fraction results. Bar plots show performance with different regularization parameters β of Equation (3) from 0.0 to 1.0 with increment of 0.2. The X-axis shows the number of tumor samples, and the Y-axis the average accuracy or RMSD.

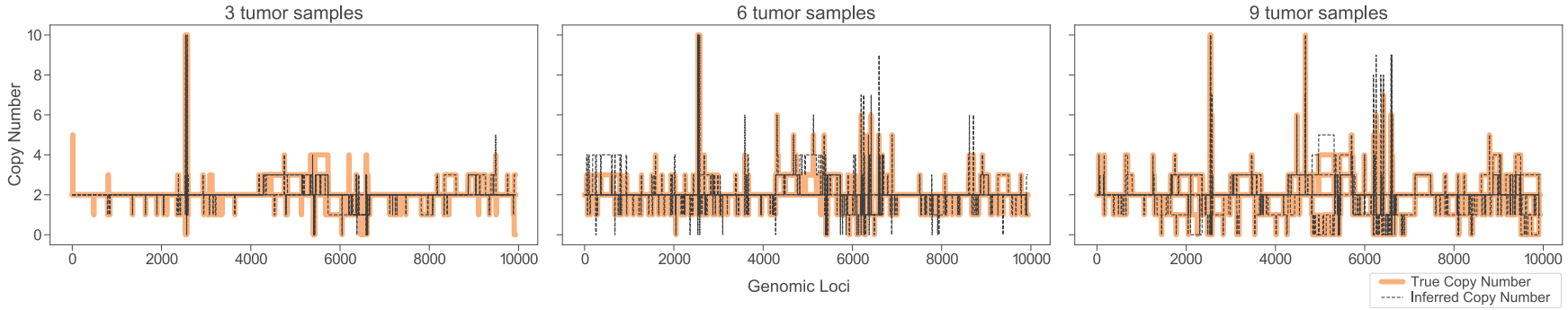

FIG. 6. Visualization of copy number as a function of genomic locus for single examples of inferred and true clones for the phylogeny-based method for three, six, and nine samples. The figure uses the minimum-RMSD pair for each case. The black dashed line shows the inferred copy number, and the orange bars show the true copy number in each position.

most loci, as well as better fitting for both large (5–10) and small (0–1) copy numbers than the phylogeny-free method of Figure 4. There is no evident pattern to the smaller number of errors that do occur for the phylogeny-based versus phylogeny-free method, which are observed for a range of low and high copy number values.

Figure 7 compares the two methods at their optimal regularization parameters for three, six, and nine tumor samples. The phylogeny-based method outperforms the phylogeny-free method in accuracy and copy number RMSD in all cases. It is slightly better in mixture fraction RMSD for three samples, but worse for six and nine samples. Figure 16 in Section 6 shows comparative performance of the two methods in individual cell components. Given the poorer performance at pure NMF of the phylogeny-based method's algorithm versus the phylogeny-free method's algorithm, we tentatively attribute the phylogeny-based method's better overall performance to better evolutionary distance estimates and not to a better optimization algorithm.

The phylogeny-based method also provides as output the phylogeny. While we cannot exhaustively show trees across all replicates, we provide three representative examples in Figure 8. Because we use true SCS data to generate our synthetic mixtures, we do not know the full ground truth trees for the data and do not attach any biological meaning to the inferred trees. We can partially validate correctness of the trees using the fact that the cells were gathered from distinct tumor regions, and while we would not expect clonal ancestry to segregate perfectly by region, we should see a trend toward closer evolutionary relationship among cells in spatial proximity. We tested whether pairs of cells from distinct regions cluster together in disjoint subtrees (a kind of partial-information quartet distance); we found that a significant majority of pairs-of-pairs do (79% for three-sample data, 74% each for six- or nine-sample data) providing some support for the biological relevance of the trees.

3.3. Application to real data

We then applied our method to true bulk data from the same glioblastoma data set used in the semi-simulated validation. For both patients (GBM07 and GBM33) and each region of the tumor, a single run of bulk sequencing and separate SCS were performed to estimate the copy number profile. We focus on the analysis on the patient GBM07 by using the phylogeny-based method since that method performed best with semisimulated data (Section 3.2).

For bulk data, we do not know the ground truth composition of clones, but may compare our deconvolution results to genomic alterations that have previously been reported in glioblastoma (Fig. 9, detailed analysis methods in Section 5.7). A total of 92 genes have been reported to be frequently altered in glioblastoma (Brennan et al., 2013; Körber et al., 2019), and 89 of these 92 genes were found to have altered copy numbers in the mixture components (clones) inferred by our methods. Among these known mutated genes, *EGFR*, *CDK4*, and *PDGFRA* have been reported to have significant recurrent amplifications, while *CDKN2A/B* and *PTEN* have been reported to have significant recurrent deletions according to the Cancer Genome Atlas Research Network et al. (2008).

Our analysis showed similar results. We observed copy number gains of *PDGFRA* in all but two components (Fig. 9C, blue text), and this gene generally exhibited substantial copy number gain in bulk sequencing (Section 6.6). *EGFR* was gained in all but one component. *PTEN* showed copy number loss in eight components and *CDKNA2/B* was lost in 10 components (Fig. 9C, orange text). We also found *ERBB2* gained in three components and *RB1* hemizygous or homozygous deletion in at least some glioblastoma samples (Cancer Genome Atlas Research Network et al., 2008; Abou-El-Ardat et al., 2017). These results are consistent with CNAs influencing the RTK/PI3K and RB signaling pathways frequently altered in glioblastoma (Fig. 9B). However, we could not identify CNAs affecting the p53 signaling pathway. We cannot tell from CNA inferences alone whether or not TP53 activity was normal in this particular patient; there are several common TP53-inactivating SNVs that would not show up in a CNA analysis (Magali et al., 2010). Figure 9C shows a more complete profile of inferred gains and losses in potential glioblastoma driver genes.

At the chromosome level, chromosomes 7, 9p, and 10 (10q) frequently display alterations in glioblastoma (Crespo et al., 2011; Davis et al., 2015; Abou-El-Ardat et al., 2017). We therefore looked at the copy numbers at a coarser scale for each inferred component, following the rules in Section 5.7 to decide whether or not the chromosome has a gain or loss. We found that in 11 components, almost all of the genomic loci showed gains in chromosome 7, although some specific loci did not display gain in some components, for example, component 10. All components show loss in chromosome 9p and eight

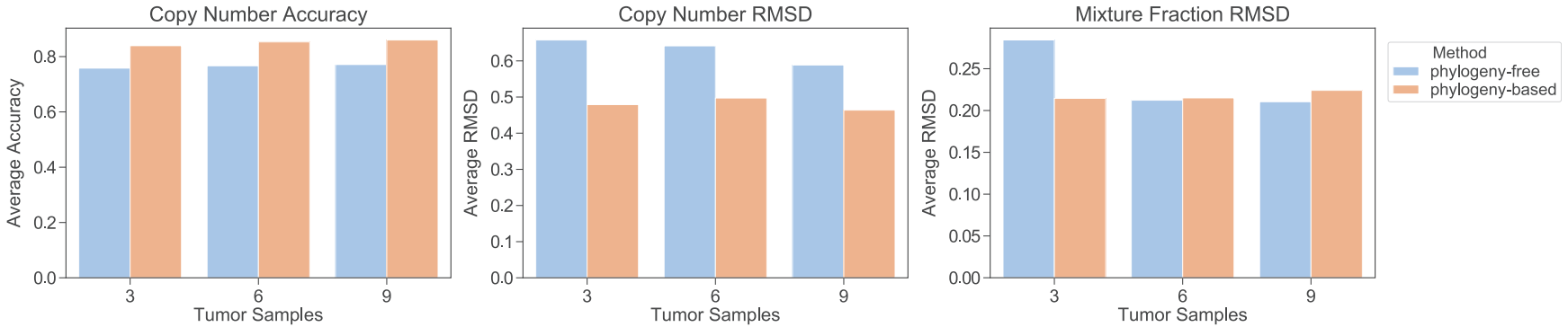

FIG. 7. Comparison between phylogeny-free and phylogeny-based methods. Bar graphs show average accuracy and RMSD over all cell components and replicates using the optimal regularization parameter for the given method, measure, and number of samples. The left panel shows accuracy in copy numbers for $\alpha = 0.2, 0.2, 0.4$ for the phylogeny-free method and $\beta = 0.2, 0.4, 0.4$ for the phylogeny-based method for 3, 6, and 9 tumor samples, respectively. The center panel shows RMSD of copy numbers for $\alpha = 0.2, 0.2, 0.2$ for the phylogeny-free method and $\beta = 0.2, 0.4, 0.4$ for the phylogeny-based method for 3, 6, and 9 tumor samples, respectively. The right panel shows RMSD of mixture fractions for $\alpha = 0.2, 0.2, 0.2$ for the phylogeny-free method and $\beta = 0.2, 0.2, 0.2$ for the phylogeny-based method for 3, 6, and 9 tumor samples, respectively. The *X*-axis shows the number of tumor samples, and the *Y*-axis the average accuracy or RMSD.

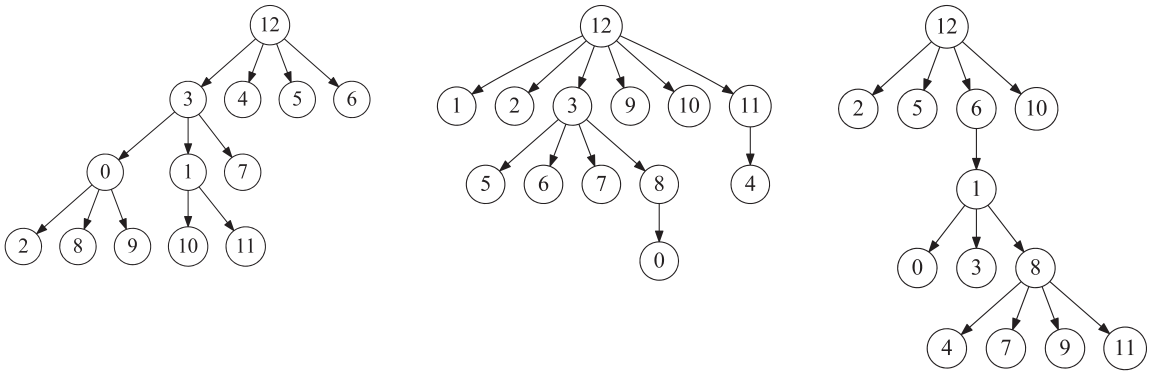

FIG. 8. Tree structure inferred via the phylogeny-based method ILP method for three problem instances. The examples come from the same instances used to pick the representative copy number profiles in Figure 6. In each tree, nodes 0–5 are inferred cells, nodes 6–11 are observed cells, and node 12 is the diploid root. ILP, integer linear programming.

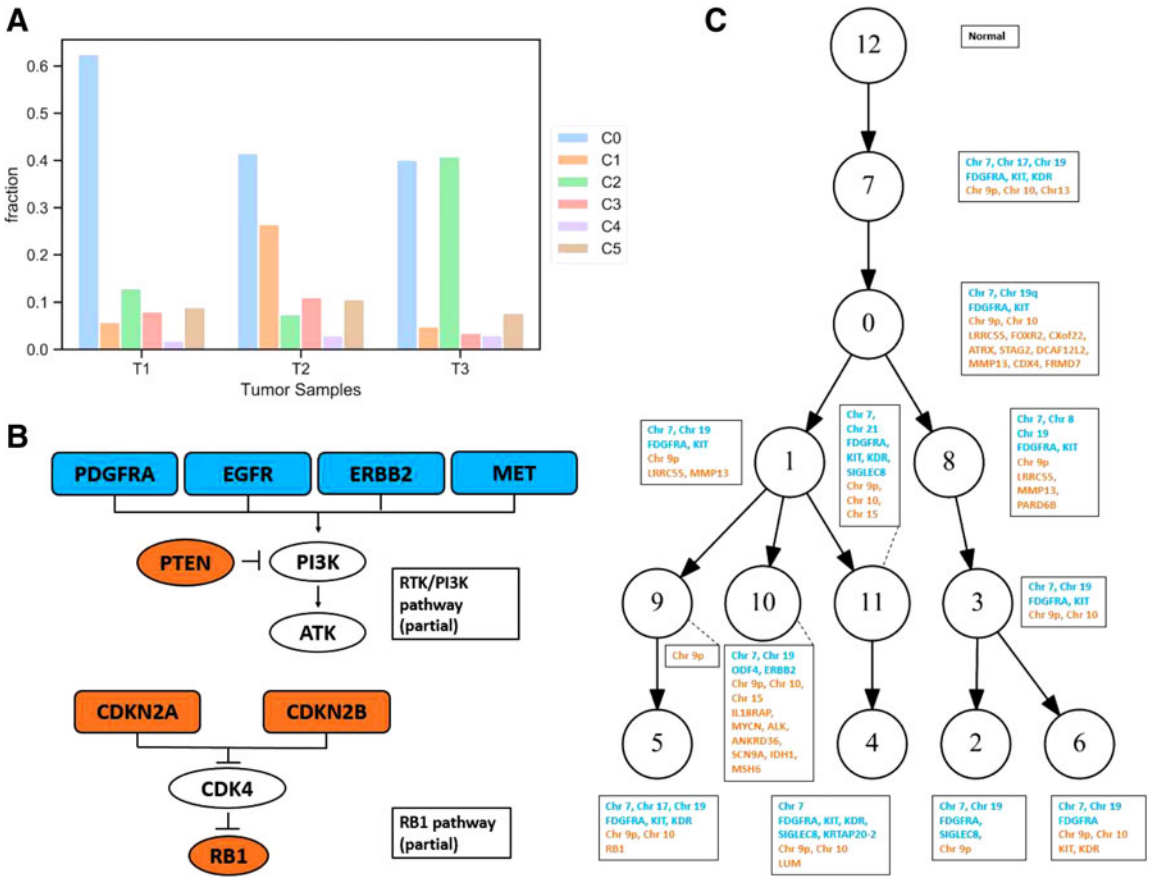

FIG. 9. Results for real data deconvolution. **(A)** The fraction of inferred components in three tumor samples. **(B)** Partial pathways of RTK/PI3K and RB, the blue boxes indicate gain of genes, orange boxes indicate loss of genes, and white boxes indicate no difference was detected. **(C)** The phylogenetic relationship among inferred clones (nodes 0–5), observed single cells (nodes 6–11), and a diploid normal cell (node 12), where blue text indicates gain of genes/chromosomes, and orange text indicates loss of genes/chromosomes.

components show loss in chromosome 10. Again, these observations are consistent with aneuploidy or large structural variants in these chromosomes as has been previously observed. Those alterations contain some key genes, for example, *EFGR* and *MET* are located on chromosome 7, *CDKNA2/B* on chromosome 9p, and *PTEN* on chromosome 10.

Abou-El-Ardat et al. (2017) and Körber et al. (2019) suggested that gain of chromosome 7, loss of chromosome 9p, and loss of chromosome 10 are likely to be early events in glioblastoma, followed by the focal loss of *CDKN2A/B*, and then later events such as *EGFR* and *PDGFR* amplification. Our inferred phylogeny (Fig. 9C) is largely consistent with, although more complicated than, this coarse model. The inferred phylogeny shows substantial branching rather than a simple linear structure, but the relative order of different components reflects early gain of chromosome 7 and loss of chromosome 9p. The loss of chromosome 10 is also inferred to be an early event, but this event could not be detected in all components. Loss of chromosome 10 is not apparent in all bulk samples, either, and the average copy number from the raw SCS data only shows slight loss at chromosome 10 (Section 6.6). The phylogeny suggests reversion of the chromosome 10 loss in some lineages, although that might be an error due to the difficulty of accurately calling CNAs that involve small, subclonal changes in copy number. Nevertheless, our inferred clonal mixture fractions (Fig. 9A) indicate that the first component (C0) has a large proportion in all three regions of the tumor sample, again consistent with the notion that gain of chromosome 7, and loss of chromosome 9p and chromosome 10 are early events in glioblastoma evolution.

We would expect that the model of early versus late events suggested by the prior work is a coarse consensus of what is a stochastic process for any individual tumor. It would require larger cohorts to test whether deviations reflect the variability of the process patient-to-patient or some systematic discrepancy in the overall model.

Besides the qualitative concordance with previous work, we also find some potential driver genes are gained in subsets of components (e.g., *KIT*, *KDR*, and *SIGLEC8*). Such subclonal events may support tumor survival or progression and thus are possible therapeutic targets for inhibitors (Gomes et al., 2007; Nobusawa et al., 2011; Kiwamoto et al., 2012; Pearson and Regad, 2017). Gain of chromosome 19 was found in two components, an event, previously identified in brain tumor-initiating cells (Davis et al., 2015), that in combination with chromosome 7 gain may help a tumor survive radiation therapy (Huhn et al., 1999). We did not see evidence of loss of heterozygosity on chromosome 19, which has been frequently found in secondary glioblastoma in Nakamura et al. (2000) and Ohgaki and Kleihues (2007). The gain of chromosome 17 was observed in two components, potentially a mechanism for amplification of the *ERBB2* oncogene. Another interesting gene is *ATRX*, which is located on chromosome X. This gene has a loss in some components. Mutations of *ATRX* are mostly found in younger patients diagnosed with glioblastoma but rare in adults, and loss of the orthologous *Atrx* accelerated glioblastoma growth in mouse (Koschmann et al., 2016; Nandakumar et al., 2017).

The mixture fraction reconstructions together with the phylogeny suggest an overall picture of early evolution of clone 0, which seeded all three tumor regions, followed by diversification most prominently into clones 1 and 2, which are dominant in two different tumor regions and represent two distinct early divergence events in the phylogeny. Most other clones are inferred to be comparatively rare. Given the uncertainty in mixture fraction analysis, we would not attach any significance to inferences of relatively rare clones in a given tumor region. We would thus suggest that the model is largely consistent with distinct tumor regions evolving independently without clear evidence for migration or reseeding between regions.

4. CONCLUSIONS AND DISCUSSION

We presented two novel methods for deconvolving clonal copy number variation from bulk tumor genomic data assisted by small amounts of SCS data. The work is intended to provide a practical strategy for producing high-quality clonal CNA deconvolution scalable to large tumor cohorts in the face of still high costs of single-cell DNA sequencing. Validation on semisimulated data shows that limited amounts of SCS copy number data can be productively used to improve upon pure bulk deconvolution, as assessed by accuracy in inferring clonal copy number profiles and their proportions in single- or multisample tumor genomic data. We showed substantial improvement by explicitly constructing clonal phylogenies jointly with deconvolution, suggesting the value of a principled evolutionary model in inferring accurate clonal structure.

While this work provides a proof-of-principle demonstration for combining bulk and SCS data for CNA deconvolution, it also suggests a need for future work. Data of the kind needed by this study remain rare,

largely because current SCS studies have not been designed for such a hybrid approach. Most studies to date have profiled many single cells from few patients rather than few cells from larger cohorts, as the current work proposes. We hope that demonstrating the effectiveness of the strategy will promote its use in future study designs, and stimulate new thinking on how most effectively to use SCS technologies to solve the underlying data science problems, in turn creating more data on which similar algorithms can be improved.

The framework might also be improved in a variety of ways, including more realistic tree models and consideration of other constraints one can extract from SCS data. For example, we considered only penalty terms on $\boldsymbol{C}$ but might also use SCS to improve estimates of the clonal frequency matrix $\boldsymbol{F}$, as given in Shackleton et al. (2009). The method might also be improved by replacing L1 distance with measures reflecting more sophisticated models of CNA-driven evolution such as in Chowdhury et al. (2014); Schwarz et al. (2015); Chowdhury et al. (2015); and El-Kebir et al. (2017). It could be useful to identify minor clones that have likely loss of heterozygosity events, since these may influence clinical outcomes, and to automate inference of the number of dominant clones. It would also be useful to combine the CNAs of this work with the SNVs of Malikic et al. (2017, 2018), as is commonly done now for bulk deconvolution (Deshwar et al., 2015; El-Kebir et al., 2016; Jiang et al., 2016), and to leverage more effectively data from new low-coverage SCS DNA-seq methods in Zahn et al. (2017) or long-read sequencing. In addition, our algorithms for solving for these models are heuristic and we might productively consider alternative methods to approach true global optima or to improve scalability to larger data sets.

5. APPENDIX 1 (SUPPLEMENTARY METHODS)

5.1. Phylogeny-free method for integrating single-cell sequencing data into non-negative matrix factorization

We solve for the phylogeny-free variant of single-cell sequencing (SCS)-assisted non-negative matrix factorization (NMF) [Eq. (2)] via an implementation of the iterative update algorithm, for which we present pseudocode as Algorithm 1. The core of the algorithm consists of the modified update rules from Section 2.2. Additional heuristic modifications are used to enforce non-negativity and integrality of solutions and make use of SCS data to bias initialization toward biologically plausible solutions, as well as to better handle possible numerical errors arising from finite machine precision.

Algorithm 1: Modified Multiplicative Update Algorithm for NMF

$\boldsymbol{C_0}$ = real single cell
$\boldsymbol{F_0}$ = rand(k, n);
normalize $\boldsymbol{F_0}$ to make each column sum up to 1;
distance = $+\infty$;
$i = 0$;
dnorm = 1;
$\text{dnorm}_0 = 0$;
while distance > threshold **do**
  $\text{dnorm} = ||\boldsymbol{B} - \boldsymbol{CF}||_{Fr}^2$;
  if dnorm_0 − dnorm >0 *or* $i >$ Maxiter **then**
    quit the loop
  end
  numerator = $\boldsymbol{C_0}^T\boldsymbol{B}$;
  $\boldsymbol{F} = \max(0,\ \boldsymbol{F_0} \cdot * (\text{numerator}./(\boldsymbol{C_0}^T\boldsymbol{C_0}\boldsymbol{F_0} + 10^{-9})))$;
  normalize $\boldsymbol{F}$ to make each column sum up to 1;
  numerator = $\boldsymbol{B}\boldsymbol{F}^T$;
  $\boldsymbol{C} = max(0, \boldsymbol{C_0} \cdot * (\text{numerator}./(\boldsymbol{C_0}\boldsymbol{F}\boldsymbol{F}^T + \alpha(\boldsymbol{C_0} - \boldsymbol{C}^{(\text{observed})}) + 10^{-9})))$;
  round entries of $\boldsymbol{C}$ to the nearest integers;
  $\boldsymbol{C_0} = \boldsymbol{C}$;
  $\boldsymbol{F_0} = \boldsymbol{F}$;
  $\text{dnorm}_0 = ||\boldsymbol{B} - \boldsymbol{C_0}\boldsymbol{F_0}||_{Fr}^2$;
  distance = dnorm_0
  $i \leftarrow i + 1$.
end

5.2. Integer linear programming for phylogeny inference in the phylogeny-based method

The major change in the phylogeny-based method of Section 2.3 is the introduction of a minimum-evolution phylogeny cost, $J(\boldsymbol{S}, \boldsymbol{C}, \boldsymbol{C}^{(\text{observed})})$, to the objective function [Eq. (4)]. We solve for the phylogeny on each pass of the coordinate descent algorithm via a multicommodity flow integer linear programming (ILP) in Sridhar et al. (2007). Let the vertex set $T = \{1, \ldots, k^*\}$ represent the set of all cells in $\boldsymbol{C}^*$. Let r be the unique predetermined root of T, which in the present practice is a purely diploid node. Furthermore, let $w_{u,v}$ be the L_1 distance between the copy number vectors corresponding to nodes $u, v \in T$. For $t, u, v \in T$, introduce the binary variables $g^t_{v,u}$ representing the amount of flow along edge (u, v) with destination $t \in T$. The full ILP is then as follows:

$$\min \sum_{u,v} s_{uv} w_{uv} \text{ s.t.} \tag{5}$$

$$\sum_v g^t_{uv} = \sum_v g^t_{vu}, \forall u \in T,\ u \neq t,\ u \neq r$$

$$\sum_v g^t_{vt} = 1,\ \forall t \in T,\ t \neq r$$

$$g^t_{vr} = 0,\ \forall v$$

$$\sum_v g^t_{tv} = 0,\ \sum_v g^t_{rv} = 1,\ \forall t \in T$$

$$0 \leq g^t_{uv} \leq s_{uv},\ \forall t \in T$$

$$s_{uu} = 0,\ \forall u$$

$$s_{uv} \in \{0, 1\}$$

Intuitively, the method defines a flow from a single root to every other vertex and requires the graph T to contain edges (u, v), indicated by $s_{uv} = 1$, such that all such flows can be accommodated. This forces the graph to be connected. We can further establish that the resulting graph is acyclic. For purposes of contradiction, assume the optimal T contains a cycle. We can then remove any edge (u, v) on the cycle and reroute any flow using that edge through the cycle in the other direction, reducing the cost of the tree by w_{uv}. Since w_{uv} is a non-negative L_1 distance, then this must reduce the cost of the tree provided $u \neq v$, showing that T was nonoptimal. This establishes by contradiction that the optimal T is connected and acyclic, that is, a tree. It is specifically a tree of minimum cost over the complete graph of observed single cells and inferred clones implied by $\boldsymbol{C}$ and $\boldsymbol{C}^{(\text{observed})}$.

At any stage of the algorithm, if there are multiple minimum-cost solutions for $\boldsymbol{F}$, $\boldsymbol{S}$, or $\boldsymbol{C}$, any solution might be chosen.

5.3. Generation of single-cell sequence glioblastoma data

Data for this analysis are provided from a study of single-cell genomics in two glioblastoma patients in Wu et al. (2016). Each patient's primary tumor was divided into three tumor regions, with 59–82 single cells extracted from each region for sequencing, for a total of 448 cells. Nuclei of 432 of these cells were amplified by multiple displacement amplification and sequenced to a coverage of $0.17\times$. We used 393 among the 410 cells that passed quality control. These cells were called for CNVs by modified variable binning in Baslan et al. (2012). The resulting calls provided the input for single-cell analysis and for construction of semisynthetic bulk data in the present work.

5.4. Generating semisynthetic data from SCS samples

This section provides additional details on the generation of semisynthetic data from SCS samples for use in validating the methods. For each simulated tumor, we generate a set of experiments in which we simulated one, two, or three bulk samples from each of three tumor regions, for a total of three, six, or nine bulk samples, respectively. To generate the simulated bulk samples, we randomly chose 25 single cells from each region, for a total of 75 cells. These 75 cells define the copy number vectors that make a nonzero contribution to any of the simulated bulk samples. For each region, we chose two cells from among the 25 to represent codominant clones. We refer to the two chosen cells as *dominant cells*. To model the noisy nature of bulk tumor data, the two dominant cells in the region, the 23 remaining cells from the same

region, and the 50 cells from the other two regions all contribute to the copy number of each bulk sample, but at different mixture fractions. The two dominant cells from each region make by far the greatest expected contribution to the simulated bulk samples for that region. This design is intended to approximate clonal structure of real tumor samples, where one might see a small number of dominant clones and a long tail of rarer cell populations (Heselmeyer-Haddad et al., 2012). We assess the methods on their ability to infer the dominant clones, with the rare clones effectively serving as a source of noise in the analysis.

To generate random mixture fractions for the sampled cells, we sampled from Dirichlet distributions. Dirichlet distributions are the conjugate priors of multinomial distributions. Thus, a Dirichlet distribution is a distribution, with vector valued parameter γ, of probabilities for the multinomial distribution—in other words, of mixture fractions for copy numbers obtained by sampling. For each region, each cell i of the two dominant cells was assigned $\gamma_i = 10$, each cell j of the 23 other selected cells from the same region was assigned $\gamma_j = 0.1$, and each cell ℓ of the other 50 cells was assigned $\gamma_\ell = 0.01$.

Figure 10 shows the overall experimental design, including different regions from which single cells are collected and DNA sequenced. We extracted the copy number for each genomic locus from the SCS results to compose the cell matrix, as shown in the heatmap in Figure 10.

More formally, we chose k total single-cell samples, $\hat{k}$ dominant cell samples per region, and $\hat{n}$ simulated bulk samples per region. Here, $k = 75$, $\hat{k} = 2$, and $\hat{n}$ is 1, 2, or 3. We drew $\hat{n}$ column vectors of mixture fractions for each region from a Dirichlet distribution using the parameters γ, assigned as described above, for a total of $n = 3\hat{n}$ columns of mixture fractions. We use these n columns to form the $k \times n$ matrix $\boldsymbol{F}^{(\mathrm{sel})}$. The $m = 9934$ and $k = 75$ columns of copy numbers from the single-cell data are used to form the $m \times k$ matrix $\boldsymbol{C}^{(\mathrm{sel})}$, and the $m \times n$ simulated matrix $\boldsymbol{B}$ is the product $\boldsymbol{B} = \boldsymbol{C}^{(\mathrm{sel})}\boldsymbol{F}^{(\mathrm{sel})}$. The entries of $\boldsymbol{B}$ are real values, not rounded to integers.

We also selected a number $k_{\mathrm{observed}} = 3\hat{k}$ of single cells to form the matrix $\boldsymbol{C}^{(\mathrm{observed})}$ for use in some algorithms. The number of single cells selected varied, but in each case, the same number of cells was selected from each region, and the cells were selected from those that were not chosen to be among the 75 used to simulate the bulk tumor.

Our goal is to recover the $k' = 3\hat{k}$ (here $k' = 6$) dominant cells, representing the major clones from the simulated bulk samples. With this notation, the problem can be formally stated as follows:

Input: A matrix of bulk tumor $\boldsymbol{B}_{m \times n}$ and a matrix of observed SCS samples $\boldsymbol{C}^{(\mathrm{observed})}_{m \times k_{\mathrm{observed}}}$.

Output: A set of inferred fundamental cell components $\boldsymbol{C}^{(\mathrm{inferred})}_{m \times k'}$ and corresponding set of mixture fraction $\boldsymbol{F}^{(\mathrm{inferred})}_{k' \times n}$.

5.5. Assessing solution quality

Given $\boldsymbol{B}_{m \times n}$, $\boldsymbol{C}^{(\mathrm{observed})}_{m \times k_{\mathrm{observed}}}$, and k', we would like to find $\boldsymbol{C}^{(\mathrm{inferred})}_{m \times k'}$ and $\boldsymbol{F}^{(\mathrm{inferred})}_{k' \times n}$ such that:

$$\boldsymbol{B}_{m \times n} \approx \boldsymbol{C}^{(\mathrm{inferred})}_{m \times k'} \boldsymbol{F}^{(\mathrm{inferred})}_{k' \times n.}$$

We repeated the above sampling, simulating, and deconvolving procedure for $N = 40$ experiments to assess the performance. Figure 2 summarizes the design by which we simulated bulk tumor from the existing SCS data and how we integrated the SCS data into bulk tumor deconvolution.

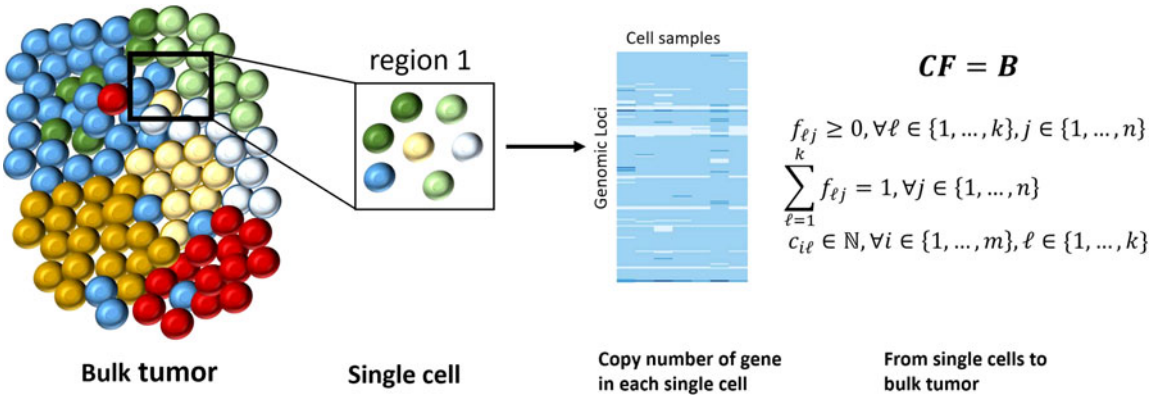

FIG. 10. The hierarchical structure of data. The illustration of the bulk tumor and SCS data used for validation, and a basic example to compose a bulk tumor from single-cell data. SCS, single-cell sequencing.

We used two ways to estimate the performance. We calculated accuracy, measured as the fraction of genomic positions (among the 40 replicates), with a correctly inferred copy number relative to the true cell components. We also measured the accuracy of inferred clones and mixture fractions by the root mean square deviation (RMSD) between true and inferred data.

For each genomic position in each inferred clone, if the copy number is equal to the copy number in the true clone, we consider the value to be accurately inferred. Then, we calculate the copy number error for each genomic position as follows:

$$\boldsymbol{C}^{(\text{acc})}_{m\times k'} = \boldsymbol{C}^{(\text{inferred})}_{m\times k'} - \boldsymbol{C}^{(\text{true})}_{m\times k'.}$$

We define the accuracy to be the fraction of 0's in $\boldsymbol{C}^{(\text{acc})}$.

Following Schwartz and Shackney (2010), we also calculate the RMSD of $\boldsymbol{C}^{(\text{inferred})}_{m\times k'}$ and $\boldsymbol{C}^{(\text{true})}_{m\times k'}$, specified as the root mean square distance over all entries of all clones between the two matrices:

$$\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{k'}\left(c_{ij}^{(\text{true})} - c_{ij}^{(\text{inferred})}\right)^2 / mk'.}$$

Similarly, we can measure the RMSD between $\boldsymbol{F}^{(\text{inferred})}_{k'\times n}$ and $\boldsymbol{F}^{(\text{true})}_{k'\times n}$ over all the loci and mixture fractions:

$$\sqrt{\sum_{i=1}^{k'}\sum_{j=1}^{n}\left(f_{ij}^{(\text{true})} - f_{ij}^{(\text{inferred})}\right)^2 / k'n.}$$

Following the same idea, when we assess the RMSD between inferred results and true data in each clone, we calculate the RMSD in the pairwise columns of the matrices:

$$\sqrt{\sum_{i=1}^{m}\left(c_{ij}^{(\text{true})} - c_{ij}^{(\text{inferred})}\right)^2 / m,} \text{ for } \forall j \in \{1, \ldots, k'\}.$$

$$\sqrt{\sum_{j=1}^{n}\left(f_{ij}^{(\text{true})} - f_{ij}^{(\text{inferred})}\right)^2 / n,} \text{ for } \forall i \in \{1, \ldots, k'\}.$$

In the inferred trees of 13 nodes each (six inferred, six observed, and a diploid root), we tested for clustering by tumor region as follows. Let the two-element sets $\{x_1, x_2\}, \{y_1, y_2\}, \{z_1, z_2\}$ be the nodes representing the two observed cells selected from the first region, second region, and third region, respectively. Considering the inferred tree as an undirected tree, we can define unique undirected paths between x_1 and x_2, between y_1 and y_2, and between z_1 and z_2. A pair of nodes/cells representing one region is considered to ''cluster together'' relative to another region if the path between the two nodes from the region of interest does not pass through either of the two nodes from the other region.

5.6. *Fully simulated SCS data*

We conducted additional tests on fully simulated data to provide some test case for which the ground truth is known and for which data could be distributed without restriction. We modeled these fully simulated data on the real data to match the true number of regions d ($d=3$ in our case), number of cells c_i ($i \in \{1, 2, 3\}$) per region, estimated rate r_{ai} of copy number variation a per region ($a \in \{0, 1, 2, \ldots, 10\}$, $i \in \{1, 2, 3\}$), and probability p_{mi} that each genomic position that has a nondiploid copy number ($m \in \{1, 2, \ldots, 9934\}$, $i \in \{1, 2, 3\}$).

For each region i, we created a root with diploid copy number in all genomic positions. We then created a binary tree where each node represents one clone. We created copy number vectors for the nodes of the tree by adding mutations according to a Poisson distribution with empirical rate r_{ai} and p_{mi} to mutate the copy number in different genomic positions so that the overall copy number distribution would be similar to that of the real SCS samples. We chose the depth D of the tree to be the minimum value that allows the number of nodes to exceed the number of cells sampled in the real SCS data in each region (in our case, $D=6$).

From the full tree, we constructed subtrees starting from the root with c_i nodes (excluding the root) by random walk with depth first order and with the equal possibility to go right or left (Fig. 11). We collected the chosen nodes to establish an artificial single-cell data set encoded as a matrix whose rows are the

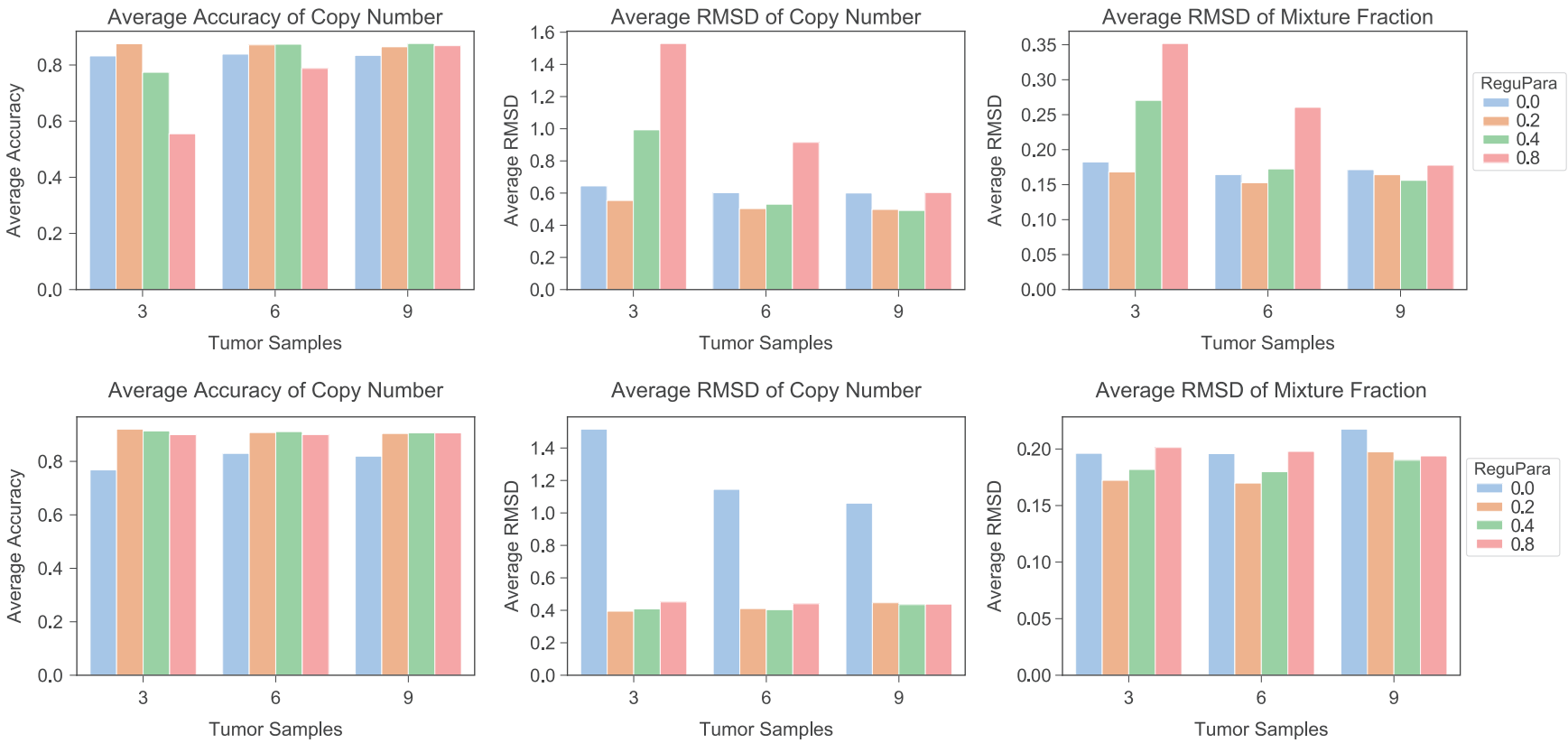

FIG. 11. A subtree chosen from the simulated completed binary tree of an example tumor region in the fully simulated data. The root was included in the tree structure for this visualization, but would be excluded when it was converted to a copy number matrix for use as input to our method.

genomic positions, and columns are the simulated single-cell samples. We applied the methods described in Section 5.4 on these simulated SCS samples to simulate bulk tumor samples, which we call the fully simulated tumor data.

We then further perturbed the artificial single-cell data to evaluate sensitivity of the methods to noise in copy number alteration (CNA) calling. We randomly choose a desired fraction (0, 0.2, or 0.4) of total genomic positions in each cell sample of $\boldsymbol{C}^{(\text{observed})}$ to perturb, then sampled a set of uniformly random genomic loci without replacement, and increased or decreased (if the copy number is greater than 1) the copy number by 1. This modification resulted in noisy versions of $\boldsymbol{C}^{(\text{observed})}$. We then applied our methods and quality assessment as described in Sections 3.1, 3.2, and 5.5 to noiseless and noise-added variants of the fully simulated data.

5.7. Real data analysis

To analyze the effectiveness of the methods on bulk genomic data, we used the three regions of the real bulk data from the patient GBM07 as $\boldsymbol{B}$ in the objective, while we randomly chose 2 cells from single-cell data from each region, used as $\boldsymbol{C}^{(\text{observed})}$ ($k = 6$). We fixed the number of components to be six. Since we tried to take advantage of the resolution of single-cell data, we set the regularization parameter to be slightly larger ($\beta = 1.0$). This setting would not be expected to hurt the inference since the accuracy of $\beta = 1.0$ is only slightly worse than the optimized parameter (Fig. 5). We retrieved a list of known mutated genes that are frequently reported in previous work (Section 3.3) and matched their coordinates to the coordinates of intervals reported in the copy number data by following the matching rules mentioned in Section 2.5. We assert that a gene has a gain if the corresponding interval has copy number at least three, and has a loss if the corresponding interval has a copy number of at most one.

For purposes of visualizing mixture fractions and phylogenies on the fully real data, we randomly selected one result out of the 100 random trials. We present a single trial due to the difficulty of unambiguously matching mixture components and phylogenetic nodes across trials that involved distinct observed single cells and somewhat different inferred clones. We then plotted the fraction of each inferred clone in each tumor region for that one randomly selected trial to obtain Figure 9A and provided the inferred phylogenetic relationship among the inferred clones and observed single cells in that one trial as Figure 9C.

6. APPENDIX 2 (SUPPLEMENTARY RESULTS)

6.1. Inference quality on semisynthetic bulk data via pure NMF

In this subsection, we describe further analysis and additional experimental results on the use of pure NMF or simple heuristic extensions thereof. We assessed the methods with a series of experiments on semisynthetic data designed to assess the phylogeny-free and phylogeny-based methods in comparison with one another and with generic NMF as functions of various parameters of the simulated data and the algorithms. In general, NMF is used to decompose high-dimensional data ($m \times n$) to a low rank basis ($m \times k$) and fractions ($k \times n$), where we usually have $k << \min(m, n)$. Therefore, we tested the pure NMF algorithm on different numbers of tumor samples to assess the data needs with respect to numbers of tumor samples. For these tests, we explored a wider range of sample numbers (3–99) than in the main article to get a sense of whether performance improvement saturates for unrealistically large numbers of samples. As described in the preceding section, we assume that we have $k' = 6$ dominant clones (derived from 2 out of random 25 cells selected in each region) as well as 23 rare clones and assess our ability to infer the dominant clones, treating the remainder as noise.

The copy number is predominantly diploid (2) across genomic positions for these data. We therefore trivially achieve relatively high accuracy either in each clone or in all clones by guessing that all clones are purely diploid. As shown in the left plot in Figure 12, the accuracy in some clones can reach as high as 80%. The overall average accuracy resulting from setting all genetic positions to be diploid, indicated by the dashed line, means that over 70% of the entries are diploid. The right plot in Figure 12 shows that the overall RMSD in copy numbers is also small for an all-diploid baseline solution. The variance means that in each clone, there are some loci of which the copy numbers are far away from diploid, as shown in Figures 4 and 6. However, the all-diploid baseline is otherwise uninformative, since we are specifically interested in

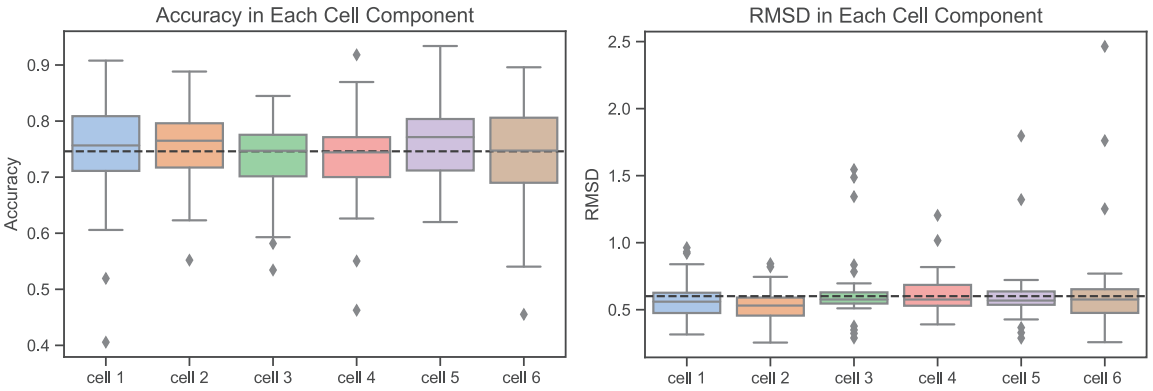

FIG. 12. Results of average and variance of accuracy (left) and RMSD (right) in copy numbers in each cell component (inferred clone). In this analysis, we compared a diploid cell matrix with all entries equal to 2 to the true cells in all the replicates to get the average and the variance of the all-diploid baseline solution. The results show that copy numbers are predominantly diploid in all clones, although with some variance. The black dashed line indicates the overall average value among all clones. Each colored box indicates one cell component. The *Y*-axis shows the accuracy or RMSD, respectively.

studying CNAs that change copy number for a subset of the genome. Nevertheless, we can use the accuracy of the all-diploid solution as a baseline against which to assess our algorithms.

We first applied NMF with random initialization to decompose the bulk tumor to infer the clonal cell components and the fraction matrix. We extended the test from small (3) to large (99) numbers of tumor samples with the regularization parameter α changing from 0 to 1 with the increment of 0.2. The $\alpha = 0$ results in the top three plots in Figure 13 show that the pure NMF method performs poorly in resolving the problem. For some α, we can achieve a good estimation of accuracy and RMSD in copy numbers in smaller numbers of tumor samples (3, 6, 9), but the improvement is generally minimal for larger numbers of tumor samples (33, 99), although some specific α values can return good estimates. These results indicate that adding the penalty to the objective has the potential to lead to a good local optimization. We also note that the performance of mixture fraction inference does not change much either across different numbers of tumor samples or among different α.

The bottom three plots in Figure 13 show results of applying the full phylogeny-free method, with real SCS data for initialization and objective penalty, to a larger range of sample sizes. The results of smaller numbers of tumor samples (3, 6, 9) have been shown in the main article, but here we include the results from larger numbers of tumor samples (33, 99). As mentioned in the main article, using real SCS data in initialization and the objective function together can lead to good estimates of copy numbers and mixture fractions of clones. Increasing the number of tumor samples substantially above the intrinsic dimension of the mixture does not appreciably improve the peak accuracy, but does appear to make the method more robust to variation in α.

6.2. Inference quality on semisynthetic bulk data via phylogeny-free augmentation with SCS data

Although we showed in the main article that the phylogeny-free method can yield reasonably good average results (Fig. 3), we were further interested in how performance of the method might vary across clones inferred. Thus, we calculated the accuracy and RMSD of inferred and true clones pairwise (Fig. 14) to see if the improvement is consistent in each clone. We also assessed this performance with the varying regularization parameter α from 0 to 1 in increments of 0.2. We only show the results for smaller tumor samples (3, 6, 9) since previous results have shown no significant difference for larger numbers of tumor samples. The top plot of Figure 14 shows that fine-tuned α can improve the accuracy of copy numbers in each clone, while continuously increasing α would result in worse performance in each clone. Similar results can also be observed in RMSDs of copy numbers (center, Fig. 14) and RMSDs of mixture fraction (bottom, Fig. 14). These results indicate that the modified NMF method affects the performance for each individual clone rather than only having the effect on a subset of the clones.

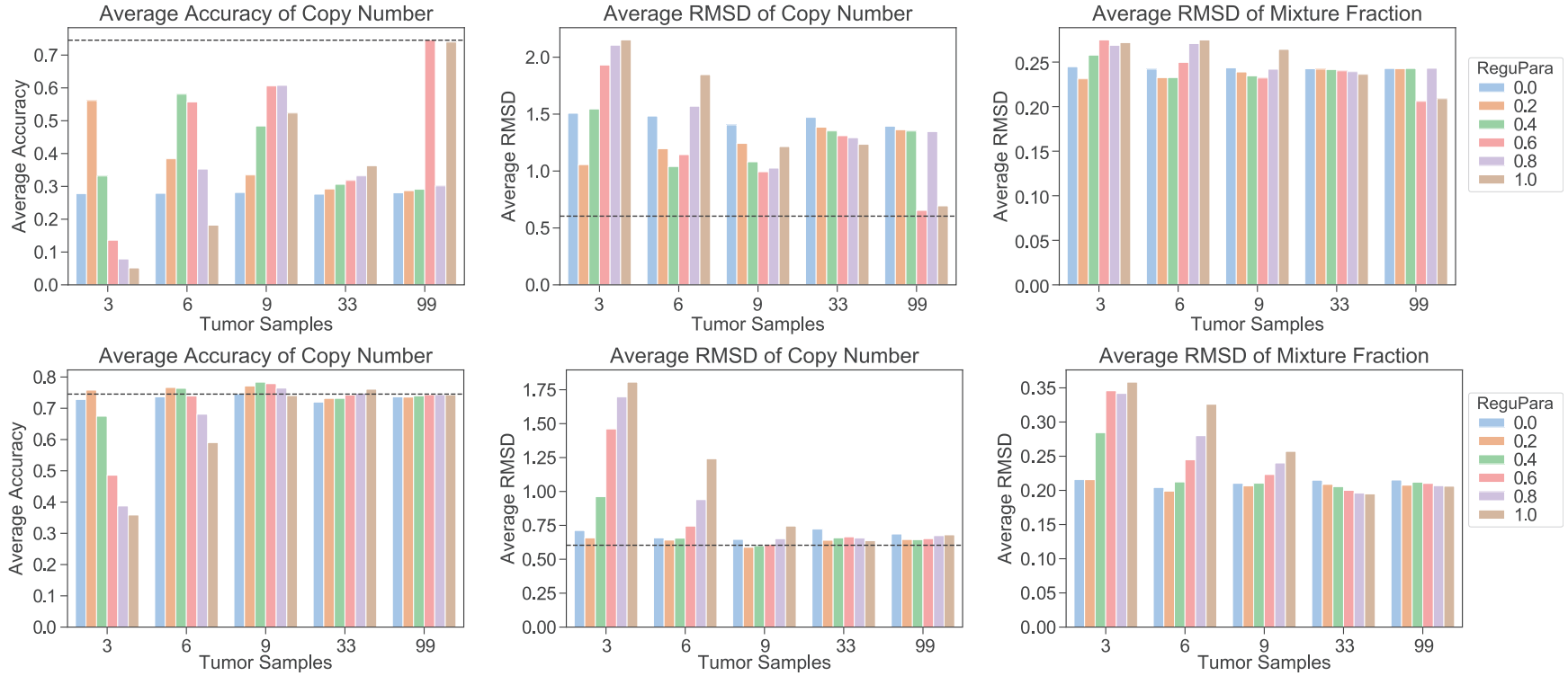

FIG. 13. NMF decomposition with L_2 penalty with random initialization (top) and with SCS initialization (bottom) for varying numbers of tumor samples. The figure shows the overall average accuracy of copy number (left), average RMSD in copy number (center), and average RMSD in mixture fraction (right) for deconvolution of varying numbers of tumor samples ($n = 3, 6, 9, 33, 99$) for random initialization and initialization with true SCS data as a function of varying numbers of tumor samples and varying regularization parameter for L_2 penalty for deviation between true and inferred clones. The black dashed lines in the left and center columns show accuracy and RMSD of copy number assignment for all-diploid inferences.

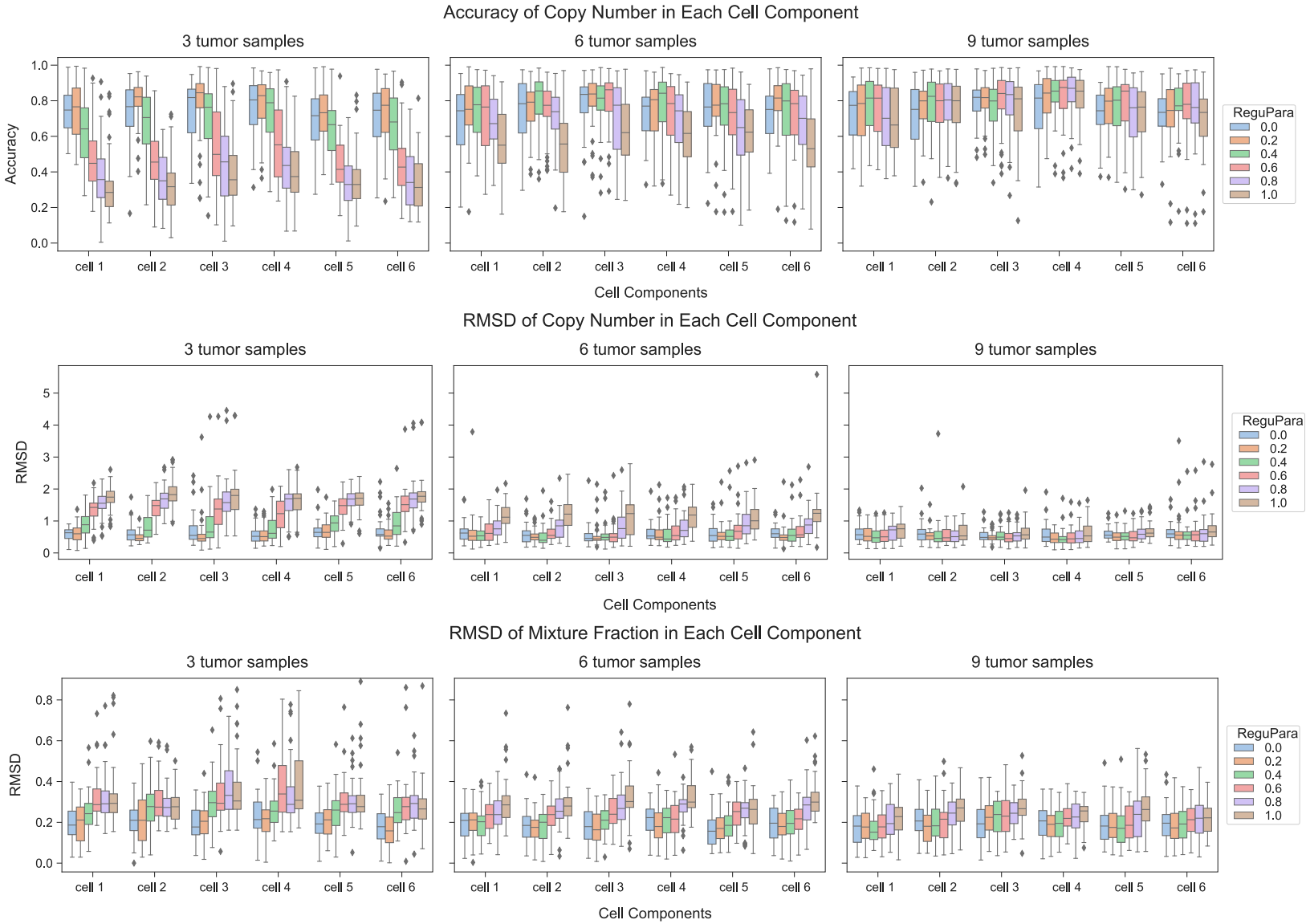

FIG. 14. Performance assessment by cell component (inferred clone) for the phylogeny-free method. The results are from the same experiments as in Figure 3, but analyzed in each clone. The top row shows the accuracy of copy number in each component. The center row shows RMSDs of copy numbers in each clone. The bottom row shows RMSDs of mixture fraction in each clone. Columns (left to right) show results for varying numbers of tumor samples ($n = 3, 6, 9$).

6.3. Inference quality on semisynthetic bulk data via phylogeny-based augmentation with SCS data

This subsection elaborates on the results for the phylogeny-based method of Section 2.3, also specifically in evaluating accuracy and RMSD of clones and mixture fractions in each inferred clone as functions of the regularization parameter β. Figure 15 shows the assessment of performance in each inferred clone with the varying regularization parameter β from 0 to 1 in increments of 0.2. The top and center rows in Figure 15 show that the accuracy and RMSDs in copy numbers were poor in each cell component when $\beta = 0$, but were substantially improved by adding the phylogeny into the objective. Each cell component is fairly robust to β variation and shows substantially better copy number accuracy and RMSD than observed with pure NMF ($\beta = 0$). However, the RMSDs of mixture fractions (bottom row, Fig. 15) do not show appreciable improvements relative to pure NMF. Examination of the variances of the mixture fractions showed that adding the phylogeny increased the variances, which suggests that combining other objective function terms with the phylogeny might lead to better estimates of the fractions for each cell component.

6.4. Comparison of phylogeny-free and phylogeny-based methods

We next compared the performance in each inferred clone of the two methods for three, six, and nine tumor samples. We chose the best-performing regularization parameter, as indicated in Figure 7, for each method in each tumor sample, respectively. When comparing the results in each inferred clone, we found that the phylogeny-based method consistently gave better average accuracy and copy number RMSD for each inferred clone for three tumor samples (top and center row, Fig. 16). Also, the number of outliers from the phylogeny-based method is generally smaller than that from the phylogeny-free method in most of the inferred clones. In contrast, for mixture fraction comparison, the phylogeny-based method yields wider variances. This observation indicates that the performance is less consistent in mixture fraction inference, but, given the higher mean accuracy, can be interpreted as an ability to infer much better mixture fractions in a subset of cases (bottom row, Fig. 16).

6.5. Inference quality on fully synthetic data

In this section, we present the results of both phylogeny-free and phylogeny-based methods on fully synthetic tumor samples (Section 5.6) following procedures similar to those of Section 3.1 and 3.2. We varied numbers of tumor samples (3, 6, 9) for these tests and regularization parameters ($\alpha = 0.0, 0.2, 0.4, 0.8$ for the phylogeny-free method and $\beta = 0.0, 0.2, 0.4, 0.8$ for the phylogeny-based method). For the phylogeny-free method (top row, Fig. 17), the results are qualitatively similar to those on semisimulated data (Fig. 3) across the range of chosen regularization parameters. The phylogeny-free method shows notable improvement over pure NMF for at least some regularization parameters, with the range of effective regularization parameters expanding with increasing numbers of tumor samples. As with the semisimulated data, the results of the phylogeny-based method (bottom row, Fig. 17) for $\beta > 0$ are substantially better than those for the phylogeny-free method. Furthermore, they show much less sensitivity to sample size and to regularization parameter across the range of nonzero values considered. The phylogeny-based results on fully simulated data are somewhat better than those seen on semisimulated data, suggesting that the more complicated error profiles of real SCS data do present some challenge to the methods.

Figure 18 shows sensitivity of the methods to noise in $\boldsymbol{C}^{(\text{reference})}$ at noise levels 0, 0.2, and 0.4. We conducted these tests for a single regularization value of 0.2 for α or β, chosen because it yielded relatively insensitive performance to sample size in the noise-free tests. For the phylogeny-free method (top row, Fig. 18), the plot shows that accuracy by each measure generally degrades with increasing noise, with copy number inference generally more sensitive to noise than mixture fraction inference. Robustness to noise by each measure increases noticeably with larger sample sizes. For the phylogeny-based method (bottom row, Fig. 18), the plot shows that accuracy is stable with increasing noise for all cases. RMSD results show some variability parameter-to-parameter, but no evident trend with increasing noise. These results show that the phylogeny-free method is robust to noise for a sufficiently large sample size, while the phylogeny-based method is robust to noise even for small sample sizes.

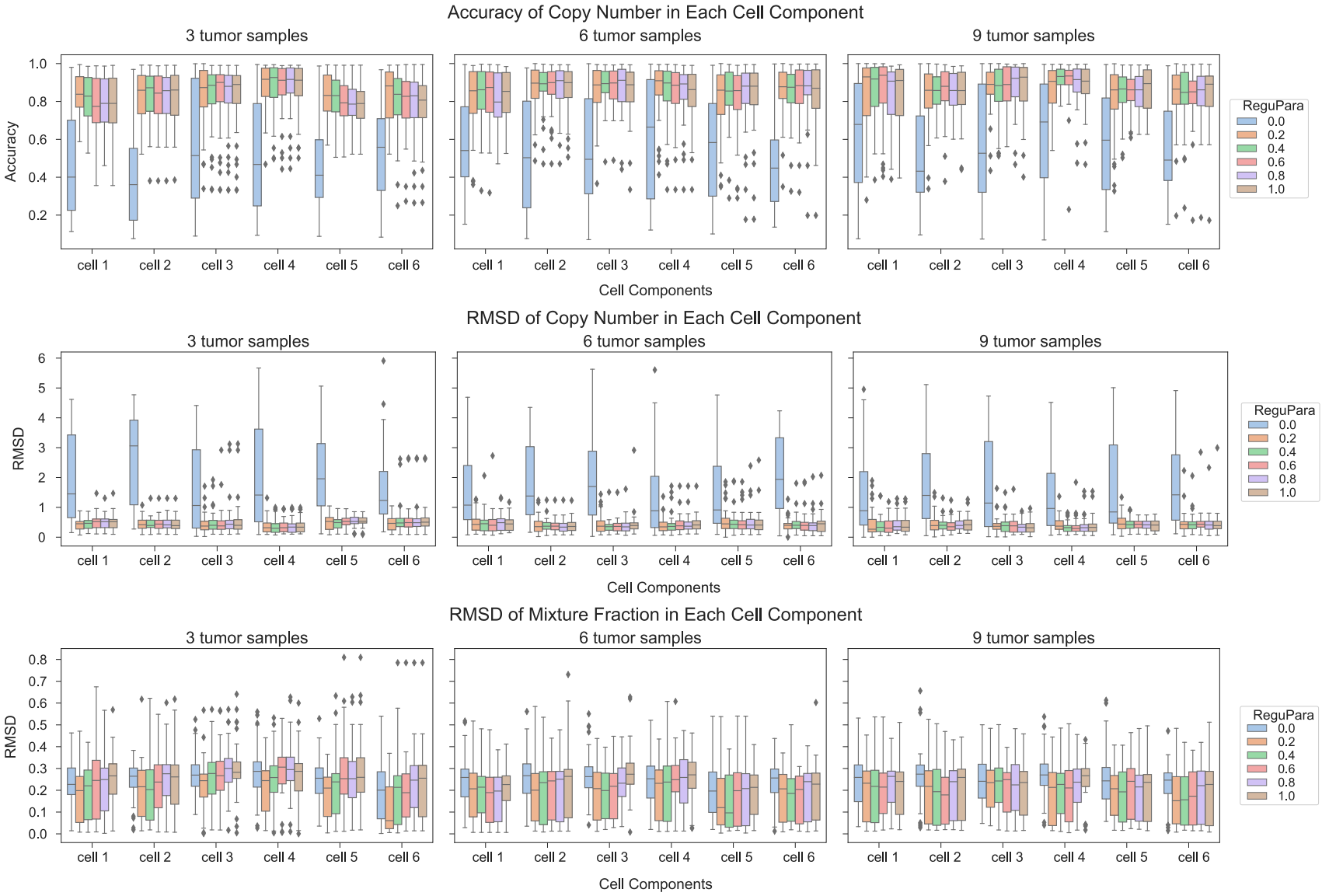

FIG. 15. Performance assessment by cell component (inferred clone) for the phylogeny-based method. The results are from the same experiments as in Figure 5, but analyzed in each clone. The top row shows the accuracy of the copy number predictions in each component. The center row shows RMSDs of copy number predictions in each clone. The bottom row shows RMSDs of mixture fraction predictions in each clone. Columns (left to right) show results for varying numbers of tumor samples ($n = 3, 6, 9$).

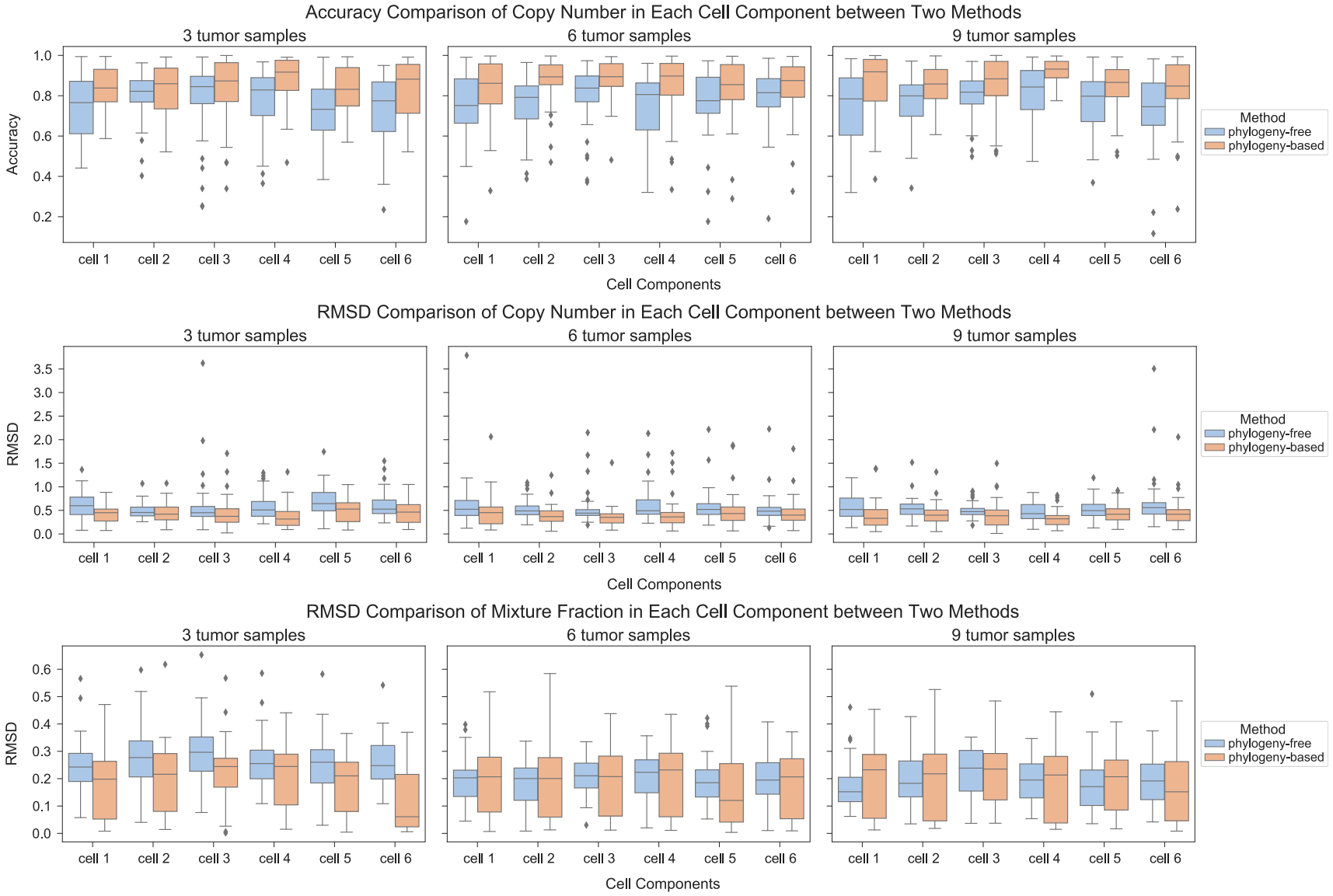

FIG. 16. Comparison in each cell component (inferred clone) between phylogeny-free and phylogeny-based methods. The figure shows the comparison between the two methods for incorporating SCS data, using optimized regularization parameters for both, as indicated in Figure 7. Results are from the same experiment as in Figure 7, but analyzed in each clone. The top row shows the comparison of accuracy of copy number predictions in each clone. The center row shows the comparison of RMSDs of copy number predictions in each clone. The bottom row shows the comparison of RMSDs of inferred mixture fraction in each clone. Columns (left to right) show results for varying numbers of tumor samples ($n = 3, 6, 9$).

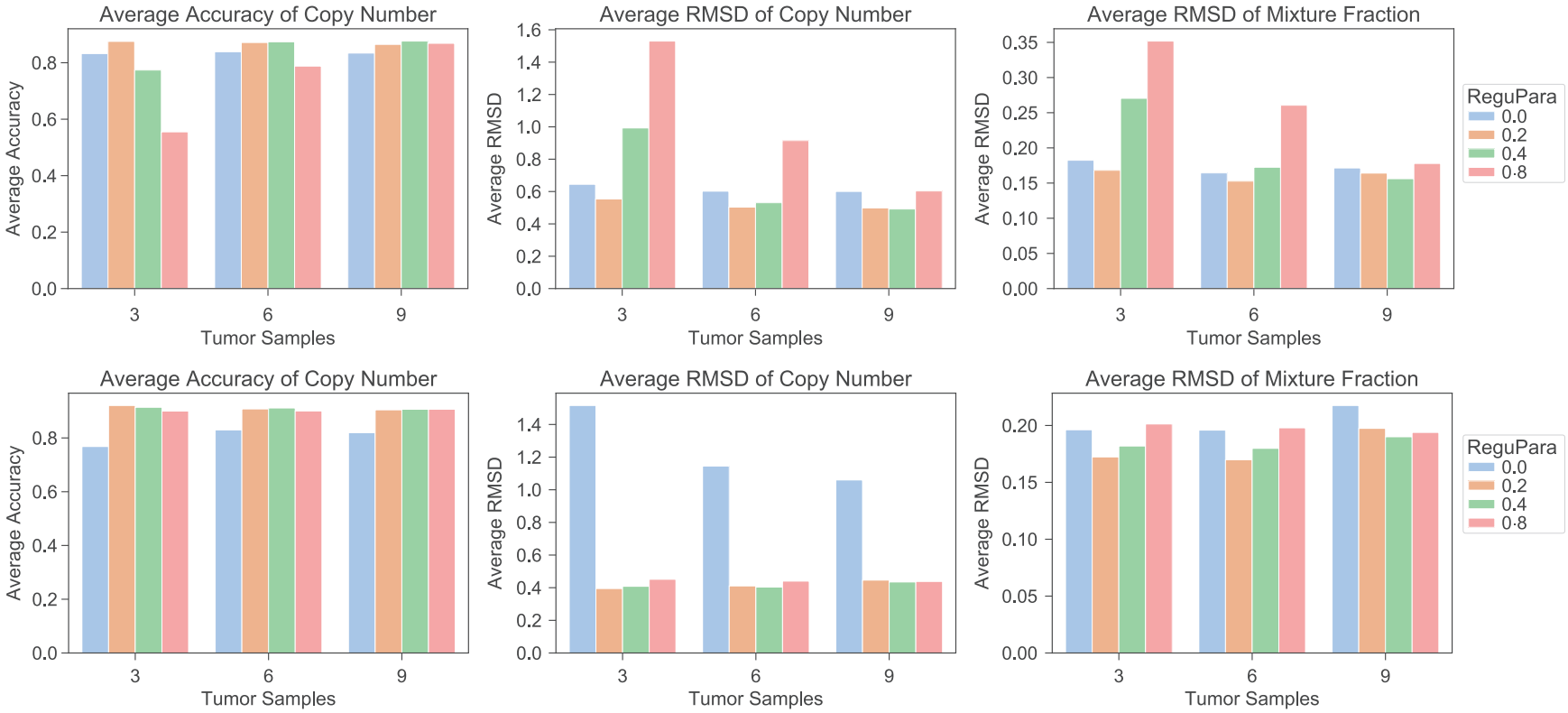

FIG. 17. Average accuracy and RMSD for our two methods as functions of tumor samples and regularization parameter on fully simulated tumor data. The top row shows the results for the phylogeny-free method, while bottom row shows the results for the phylogeny-based method. The left column shows the average accuracy of inferred copy numbers, the center column average RMSD between inferred and true copy numbers, and the right column average RMSD between the inferred and true mixture fractions. Bar plots show performance with regularization parameters 0.0, 0.2, 0.4, and 0.8. The *X*-axis shows the number of tumor samples, and the *Y*-axis the average accuracy or RMSD.

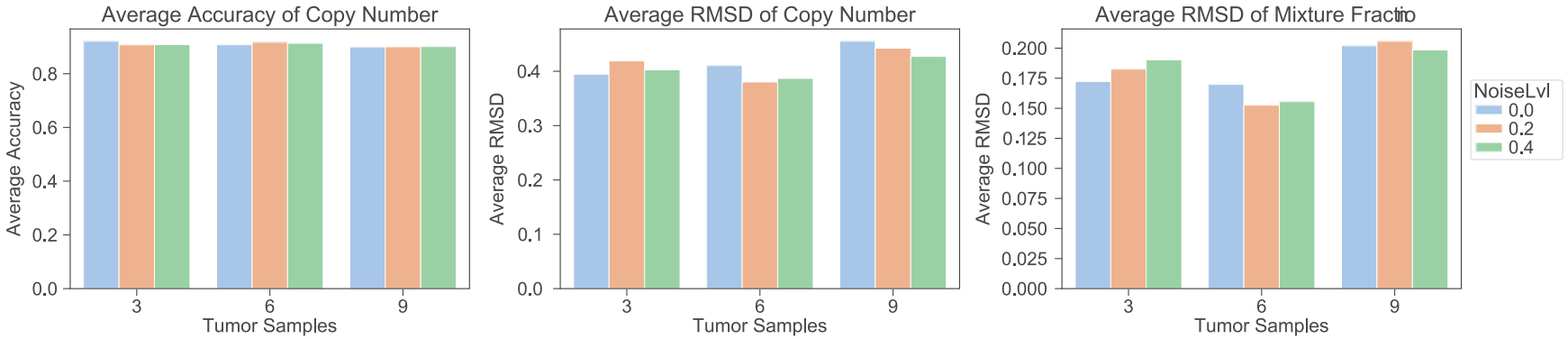

FIG. 18. Average accuracy and RMSD of sensitivity analysis on phylogeny-free (top) and phylogeny-based methods (bottom) as functions of noise level and number of tumor samples. The top row shows the results for phylogeny-free method, and the bottom row shows the results for phylogeny-based method. The left column shows the average accuracy of inferred copy numbers, the center column average RMSD between inferred and true copy numbers, and the right column average RMSD between the inferred and true mixture fractions. Bar plots show performance at different noise levels 0.0, 0.2, and 0.4. The *X*-axis shows the number of tumor samples, and the *Y*-axis the average accuracy or RMSD.

6.6. *Average copy number in real bulk tumor and single-cell data*

In this section, we examine a visualization of average copy numbers along the genomic loci in real bulk tumor samples and corresponding single-cell data. We matched the coordinates of our data to the hg19 human genomic reference and split the loci into different chromosomes (Fig. 19). We observe some variance in the copy number profile of bulk data (top, Fig. 19), in part, because the copy number changes are not identical across the three bulk samples. For example, two samples show the loss of chromosome 10, while the other does not. Also, not every bulk sample shows the gain of the full chromosome 7, with some only showing a gain of subset of the loci in chromosome 7. We further note that the bulk sequencing reflects only the average genomic profiles of the clones. In the single-cell data (center, Fig. 19), by contrast, we can clearly see the gain of chromosome 7 and the loss of chromosome 10. The loss of chromosome 10 is less pronounced, which may due to the coarse resolution of genomic breakpoints and uncertainty of measurement of calling copy number in individual cells. Some single cells exhibit the loss of full chromosome 10, while others do not. Similar reasons may account for why both bulk and single-cell data show the gain and loss of partial chromosome 11.

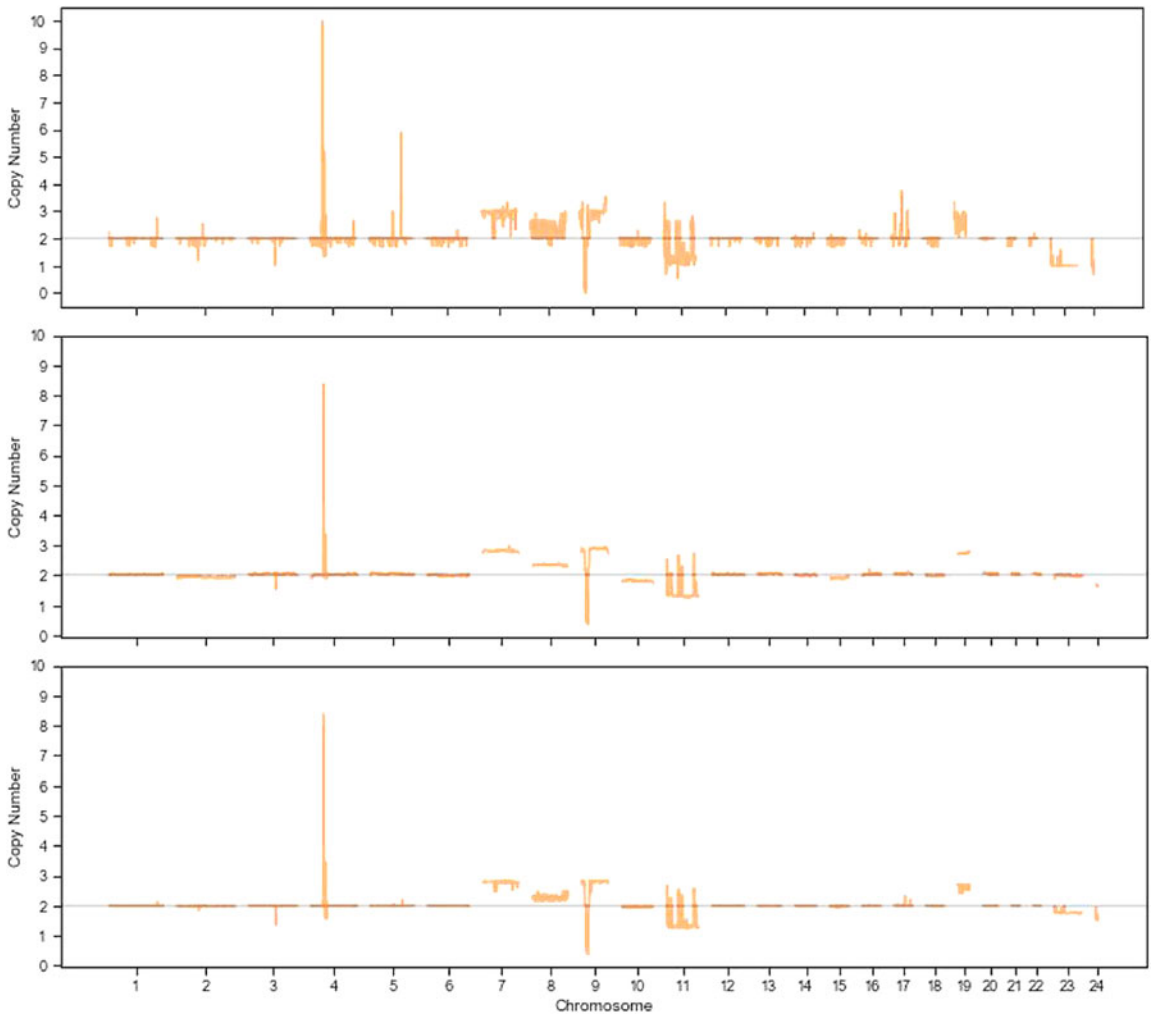

FIG. 19. Average copy number aberration profiles for bulk tumor samples (top), all available single-cell data (center), and the inferred clones (bottom). The bulk data show considerable variability in aberrations, while the single-cell data are more stable. Our method infers clones that reflect events supported by bulk and/or single-cell data, allowing some reduction in noise. It also reduces the variation of averaging bulk samples and allows us to verify events in the single-cell data via support for similar events in the bulk data, addressing some of the noise problems of current SCS technologies. In each plot, the *X*-axis shows different numbered chromosomes, with chromosomes X and Y represented by 23 and 24. The *Y*-axis indicates the copy number.

The copy number profile of inferred cell components (bottom, Fig. 19) not only reflects the common events in both bulk and single-cell data (e.g., the large copy number of *PDGFR*) but also shows the similarity with the available single-cell data. The figure shows the gain of chromosome 7 and the loss of chromosome arm 9p, while the loss of chromosome 10 is less pronounced. Comparing these three average copy number profiles, with the assistance of SCS, the variation of copy number changes in some genomic loci [chromosome 7, 9, 17, 19, X(23), Y(24) in this case] has been minimized, providing more confidence about the copy number aberrations in the genomic loci with which we are interested. We can also derive more confidence about the aberrations in chromosome regions that are less commonly altered (e.g., chromosome 19 and chromosome X, which contains *ATRX*). Most importantly, we show it is possible to accomplish this using only a very small subset of the single-cell data (six cells in this case) to take the advantage of the resolution of SCS technology.

DISCLAIMER

The Pennsylvania Department of Health specifically disclaims responsibility for any analyses, interpretations, or conclusions.

AUTHOR DISCLOSURE STATEMENT

R.S. is funded, in part, by the University of Pittsburgh Medical Center.

FUNDING INFORMATION

This research was supported, in part, by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, and both Center for Cancer Research and Division of Cancer Epidemiology and Genetics within the National Cancer Institute. This research was supported, in part, by the Exploration Program of the Shenzhen Science and Technology Innovation Committee [JCYJ20170303151334808]. Portions of this work have been funded by the U.S. N.I.H. award R21CA216452 and the Pennsylvania Department of Health award 4100070287.

REFERENCES

Abou-El-Ardat, K., Seifert, M., Becker, K., et al. 2017. Comprehensive molecular characterization of multifocal glioblastoma proves its monoclonal origin and reveals novel insights into clonal evolution and heterogeneity of glioblastomas. *Neuro Oncol.* 19, 546–557.

Barber, L.J., Davies, M.N., and Gerlinger, M. 2015. Dissecting cancer evolution at the macro-heterogeneity and micro-heterogeneity scale. *Curr. Opin. Genet. Dev.* 30, 1–6.

Baslan, T., Kendall, J., Rodgers, L., et al. 2012. Genome-wide copy number analysis of single cells. *Nat. Proto.* 7, 1024.

Berry, M.W., Browne, M., Langville, A.N., et al. 2007. Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Stat. Data Anal.* 52, 155–173.

Brennan, C.W., Verhaak, R.G.W., McKenna, A., et al. 2013. The somatic genomic landscape of glioblastoma. *Cell* 155, 462–477.

Cancer Genome Atlas Research Network et al. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068.

Chowdhury, S.A., Gertz, E.M., Wangsa, D., et al. 2015. Inferring models of multiscale copy number evolution for single-tumor phylogenetics. *Bioinformatics* 31, i258–i267.

Chowdhury, S.A., Shackney, S.E., Heselmeyer-Haddad, K., et al. 2014. Algorithms to model single gene, single chromosome, and whole genome copy number changes jointly in tumor phylogenetics. *PLoS Comput. Biol.* 10, e1003740.

Coyne, G.O., Takebe, N., and Chen, A.P. 2017. Defining precision: The precision medicine initiative trials NCI-IMPACT and NCI-MATCH. *Curr. Probl. Cancer* 41, 182–193.

Crespo, I., Vital, A.L., Nieto, A.B., et al. 2011. Detailed characterization of alterations of chromosomes 7, 9, and 10 in glioblastomas as assessed by single-nucleotide polymorphism arrays. *J. Mol. Diagn.* 13, 634–647.

Davis, B., Shen, Y., Poon, C.C., et al. 2015. Comparative genomic and genetic analysis of glioblastoma-derived brain tumor-initiating cells and their parent tumors. *Neuro Oncol.* 18, 350–360.

Deshwar, A.G., Vembu, S., Yung, C.K., et al. 2015. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* 16, 35.

Dexter, D.L., and Leith, J.T. 1986. Tumor heterogeneity and drug resistance. *J. Clin. Oncol.* 4, 244–257.

Eaton, J., Wang, J., and Schwartz, R. 2018. Deconvolution and phylogeny inference of structural variations in tumor genomic samples. *Bioinformatics* 34, i357–i365.

El-Kebir, M., Raphael, B. J., Shamir, R., et al. 2017. Complexity and algorithms for copy-number evolution problems. *Algorithms Mol. Biol.* 12, 13.

El-Kebir, M., Satas, G., Oesper, L., et al. 2016. Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. *Cell Syst.* 3, 43–53.

Fisher, R., Pusztai, L., and Swanton, C. 2013. Cancer heterogeneity: Implications for targeted therapeutics. *Br. J. Cancer* 108, 479–485.

Gomes, A.L., Reis-Filho, J.S., Lopes, J.M., et al. 2007. Molecular alterations of *KIT* oncogene in gliomas. *Anal. Cell. Pathol.* 29, 399–408.

Heselmeyer-Haddad, K., Berroa Garcia, L.Y., Bradley, A., et al. 2012. Single-cell genetic analysis of ductal carcinoma in situ and invasive breast cancer reveals enormous tumor heterogeneity yet conserved genomic imbalances and gain of *MYC* during progression. *Am. J. Pathol.* 181, 1807–1822.

Hou, Y., Song, L., Zhu, P., et al. 2012. Single-cell exome sequencing and monoclonal evolution of a JAK-2 negative myeloproliferative neoplasm. *Cell* 148, 873–885.

Huhn, S.L., Mohapatra, G., Bollen, A., et al. 1999. Chromosomal abnormalities in glioblastoma multiforme by comparative genomic hybridization: Correlation with radiation treatment outcome. *Clin. Cancer Res.* 5, 1435–1443.

Jahn, K., Kuipers, J., and Beerenwinkel, N. 2016. Tree inference for single-cell data. *Genome Biol.* 17, 86.

Jiang, Y., Qiu, Y., Minn, A.J., et al. 2016. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc. Natl Acad. Sci. U. S. A.* 113, E5528–E5537.

Kiwamoto, T., Kawasaki, N., Paulson, J.C., et al. 2012. Siglec-8 as a drugable target to treat eosinophil and mast cell-associated conditions. *Pharmacol. Ther.* 135, 327–336.

Körber, V., Yang, J., Barah, P., et al. 2019. Evolutionary trajectories of idhwt glioblastomas reveal a common path of early tumorigenesis instigated years ahead of initial diagnosis. *Cancer Cell* 35, 692–704.

Koschmann, C., Lowenstein, P.R., and Castro, M.G. 2016. ATRX mutations and glioblastoma: Impaired dna damage repair, alternative lengthening of telomeres, and genetic instability. *Mol. Cell. Oncol.* 3, e1167158.

Kuipers, J., Jahn, K., and Beerenwinkel, N. 2017. Advances in understanding tumour evolution through single-cell sequencing. *Biochim. Biophys. Acta Rev. Cancer* 1867, 127–138.

Lee, D.D., and Seung, H.S. 2001. Algorithms for non-negative matrix factorization, 556–562. *In* Jordan M.I., LeCun, Y., and Solla, S.A., eds: A*dvances in Neural Information Processing Systems.* The MIT Press, Cambridge, MA.

Lei, H., Ma, F., Chapman, A., et al. 2015. Single-cell whole-genome amplification and sequencing: Methodology and applications. *Annu Rev Genomics Hum Genet* 16, 79–102.

Loeb, L.A. 2001. A mutator phenotype in cancer. *Cancer Res* 61, 3230–3239.

Macintyre, G., Goranova, T.E., De Silva, D., et al. 2018. Copy number signatures and mutational processes in ovarian carcinoma. *Nat Genet* 50, 1262–1270.

Magali, O., Hollstein, M., and Hainaut, P. 2010. TP53 mutations in human cancers: Origins, consequences, and clinical use. *Cold Spring Harb. Perspect. Biol.* 2, a001008.

Malikic, S., Ciccolella, S., Mehrabadi, F.R., et al. 2018. PhISCS-a combinatorial approach for sub-perfect tumor phylogeny reconstruction via integrative use of single cell and bulk sequencing data. *bioRxiv* 376996.

Malikic, S., Jahn, K., Kuipers, J., et al. 2019. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nat. Commun.* 10, 2750.

Marusyk, A., and Polyak, K. 2010. Tumor heterogeneity: Causes and consequences. *Biochim. Biophys. Acta.* 1805, 105–117.

McGranahan, N., Rosenthal, R., Hilley, C.T., et al. 2017. Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell* 171, 1259–1271.

Nakamura, M., Yang, F., Fujisawa, H., et al. 2000. Loss of heterozygosity on chromosome 19 in secondary glioblastomas. *J. Neuropathol. Exp. Neurol.* 59, 539–543.

Nandakumar, P., Mansouri, A., and Das, S. 2017. The role of ATRX in glioma biology. *Front. Oncol.* 7, 236.

Navin, N., Kendall, J., Troge, J., et al. 2011. Tumour evolution inferred by single-cell sequencing. *Nature* 472, 90–94.

Nobusawa, S., Stawski, R., Kim, Y.-H., et al. 2011. Amplification of the PDGFRA, KIT and KDR genes in glioblastoma: A population-based study. *Neuropathology* 31, 583–588.

Nowell, P.C. 1976. The clonal evolution of tumor cell populations. *Science* 194, 23–28.

Ohgaki, H., and Kleihues, P. 2007. Genetic pathways to primary and secondary glioblastoma. *Am. J. Pathol.* 170, 1445–1453.

Ortega, M.A., Poirion, O., Zhu, X., et al. 2017. Using single-cell multiple omics approaches to resolve tumor heterogeneity. *Clin. Transl. Med.* 6, 46.

Pearson, J.R., and Regad, T. 2017. Targeting cellular pathways in glioblastoma multiforme. *Signal Transduct. Target. Ther.* 2, 17040.

Ross, E.M., and Markowetz, F. 2016. OncoNEM: Inferring tumor evolution from single-cell sequencing data. *Genome Biol.* 17, 69.

Schwartz, R., and Schäffer, A.A. 2017. The evolution of tumour phylogenetics: Principles and practice. *Nat. Rev. Genet.* 18, 213–229.

Schwartz, R., and Shackney, S.E. 2010. Applying unmixing to gene expression data for tumor phylogeny inference. *BMC Bioinformatics* 11, 42.

Schwarz, R.F., Ng, C.K.Y., Cooke, S.L., et al. 2015. Spatial and temporal heterogeneity in high-grade serous ovarian cancer: A phylogenetic analysis. *PLoS Med.* 12, e1001789.

Shackleton, M., Quintana, E., Fearon, E.R., et al. 2009. Heterogeneity in cancer: Cancer stem cells versus clonal evolution. *Cell* 138, 822–829.

Siegel, R.L., Miller, K.D., Fedewa, S.A., et al. 2017. Colorectal cancer statistics, 2017. *CA Cancer J. Clin.* 67, 177–193.

Sridhar, S., Lam, F., Blelloch, G.E., et al. 2007. Efficiently finding the most parsimonious phylogenetic tree via linear programming, 37–48. *In* Mandoiu, I., and Zelikovsky, A., eds: *International Symposium on Bioinformatics Research and Applications*, Lecture Notes in Bioinformatics, Vol. 4463. Springer, Berlin, Heidelbery and New York.

Subramanian, A., and Schwartz, R. 2015. Reference-free inference of tumor phylogenies from single-cell sequencing data. *BMC Genomics* 16, S7.

Thurau, C., Kersting, K., and Bauckhage, C. 2009. Convex non-negative matrix factorization in the wild, 523–532. *In* Wang, W., Kargupta, H., RanKa, S., Yu, P.S., and Wu, X., eds: *2009 Ninth IEEE International Conference on Data Mining*. The IEEE Computer Society, Washington, DC.

Tolliver, D., Tsourakakis, C., Subramanian, A., et al. 2010. Robust unmixing of tumor states in array comparative genomic hybridization data. *Bioinformatics* 26, i106–i114.

Wang, Y., Waters, J., Leung, M. L., et al. 2014. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* 512, 155–160.

Wang, Y., and Zhang, Y. 2013. Nonnegative matrix factorization: A comprehensive review. *IEEE Trans. Knowl. Data Eng.* 25, 1336–1353.

Williams, M.J., Werner, B., Barnes, C.P., et al. 2016. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* 48, 238–244.

Wu, K., Wan, W., Hou, Y., et al. 2016. Diverse evolutionary dynamics in glioblastoma inference by multi-region and single-cell sequencing. *J. Clin. Oncol.* 34, 11580–11580.

Zaccaria, S., El-Kebir, M., Klau, G.W., et al. 2018. Phylogenetic copy-number factorization of multiple tumor samples. *J. Comput. Biol.* 25, 689–708.

Zack, T.I., Schumacher, S.E., Carter, S.L., et al. 2013. Pan-cancer patterns of somatic copy number alteration. *Nat Genet* 45, 1134–1140.

Zafar, H., Tzen, A., Navin, N., et al. 2017. SiFit: Inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biol.* 18, 178.

Zahn, H., Steif, A., Laks, E., et al. 2017. Scalable whole-genome single-cell library preparation without pre-amplification. *Nat. Methods* 14, 167.

Address correspondence to:
Dr. Russell Schwartz
Department of Biological Sciences
Carnegie Mellon University
4400 Fifth Avenue
Pittsburgh, PA 15213

E-mail: russells@andrew.cmu.edu