

Joint Clustering of Single-Cell Sequencing and Fluorescence In Situ Hybridization Data for Reconstructing Clonal Heterogeneity in Cancers

XUECONG FU,¹ HAOYUN LEI,² YIFENG TAO,^{2,i} KERSTIN HESELMAYER-HADDAD,³
IRIANNA TORRES,^{3,*} MICHAEL DEAN,⁴ THOMAS RIED,³ and RUSSELL SCHWARTZ^{1,2,ii}

ABSTRACT

Aneuploidy and whole genome duplication (WGD) events are common features of cancers associated with poor outcomes, but the ways they influence trajectories of clonal evolution are poorly understood. Phylogenetic methods for reconstructing clonal evolution from genomic data have proven a powerful tool for understanding how clonal evolution occurs in the process of cancer progression, but extant methods so far have limited the ability to resolve tumor evolution via ploidy changes. This limitation exists in part because single-cell DNA-sequencing (scSeq), which has been crucial to developing detailed profiles of clonal evolution, has difficulty in resolving ploidy changes and WGD. Multiplex interphase fluorescence in situ hybridization (miFISH) provides a more unambiguous signal of single-cell ploidy changes but it is limited to profiling small numbers of single markers. Here, we develop a joint clustering method to combine these two data sources with the goal of better resolving ploidy changes in tumor evolution. We develop a probabilistic framework to maximize the probability of latent variables given the pre-clustered datasets, which we optimize via Markov chain Monte Carlo sampling combined with linear regression. We validate the method by using simulated data derived from a glioblastoma (GBM) case profiled by both scSeq and miFISH. We further apply the method to two GBM cases with scSeq and miFISH data by reconstructing a phylogenetic tree from the joint clustering results, demonstrating their synergistic value in understanding how focal copy number changes and WGD events can collectively contribute to tumor progression.

Keywords: cancer, clonal evolution, sequencing, single-cell genomics.

¹Department of Biological Sciences, and ²Computational Biology Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

³Genetics Branch, Cancer Genomics Section, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA.

⁴Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Rockville, Maryland, USA.

^{*}Current Affiliation: School of Medicine and Dentistry, University of Rochester, Rochester, New York, USA.

ⁱORCID ID (<https://orcid.org/0000-0001-5076-0081>).

ⁱⁱORCID ID (<https://orcid.org/0000-0002-4970-2252>).

1. INTRODUCTION

TUMOR PROGRESSION OCCURS through a process of clonal evolution (Nowell, 1976), during which tumor cells typically accumulate a variety of different types of genetic aberrations. Copy number alterations (CNAs), which play a particularly vital role in how tumors adapt functionally over progression, are a common outcome of chromosome instability (CIN) (Heng et al., 2013; Bakhoum et al., 2018). The CNAs can occur at multiple scales, from small focal gains or losses, to gain or loss of whole chromosomes or chromosome arms, to whole genome duplication (WGD). The WGD, in particular, is a primary driving force of aneuploidy in tumors that is associated with poor outcomes (Van de Peer et al., 2017; Bielski et al., 2018). Characterizing how CNAs at various scales act in tumor evolution is, thus, crucial to accurately reconstructing how tumor genomes functionally adapt during evolution.

The reconstruction of clonal evolution in cancers from genomic data has become the subject of a field known as tumor phylogenetics (Schwartz and Sch  ffer, 2017). Historically, tumor phylogeny methods have focused primarily on evolution by single-nucleotide variants (SNVs), with mechanisms of evolution by copy number variations seeing comparatively little attention in the tumor phylogeny literature. Letouz   et al. (2010) used breakpoints to construct relationships between different tumors. Multiplex interphase DNA fluorescence in situ hybridization (miFISH), a technique for measuring copy numbers of small numbers of probes in many single cells, has been previously used to study copy number evolution in cancers (Pennington et al., 2007; Chowdhury et al., 2013) and eventually extended to explicitly model WGD as distinct from focal and chromosome-scale gain-and-loss mechanisms (Chowdhury et al., 2014, 2015). A particularly influential variant of copy number evolution was the MEDICC model (Schwarz et al., 2014a), which allowed for CNAs at arbitrary scales. Later work showed the MEDICC model to be solvable in linear time (Zeira et al., 2017) and extended to a weighted version (Zeira and Raphael, 2020).

Single-cell DNA-sequencing (scSeq) was revolutionary for tumor phylogenetics (Navin et al., 2011) and in early versions used primarily to study evolution by CNAs, but it has not generally been used to study aneuploidy or WGD specifically. There is so far little research on algorithms for copy number evolution from large-scale single-cell sequencing data. The CNAs have been used as an evidence of single-nucleotide variant (SNV) changes caused by copy number changes or simply treated as a confounding factor in studies of SNV evolution (Satas et al., 2020). Methods have begun to appear explicitly modeling WGD events and other large chromosomal aberrations from sequence data (Zaccaria and Raphael, 2020), but with a restrictive model for WGD derived from the common assumption that WGD is a one-time early event in tumor evolution. The MEDICC (Schwarz et al., 2014b) method for CNA phylogenies has also recently been extended to allow for explicit WGD modeling (Petkovic et al., 2021). Prior studies have suggested that WGD may occur repeatedly in clonal evolution of single tumors (Shackney et al., 1995; Oltmann et al., 2018). Fitting more complex models from scSeq data is problematic because of the difficulty of reliably identifying WGD events from scSeq. Another obstacle is that scSeq data are noisy and prone to challenging technical artifacts, which makes it hard to infer precise phylogenetic information.

The miFISH provides an alternative way to measure copy numbers of target DNA sequences at the single-cell level for use in reconstructing clonal evolution (Pennington et al., 2007; Chowdhury et al., 2013). The miFISH has the advantage of providing absolute copy numbers, in contrast to the relative amounts offered by scSeq that are typically normalized with respect to ploidy. The miFISH is also scalable to large numbers of cells at a comparatively low cost. These advantages make miFISH particularly useful for studying ploidy changes and WGD as well as for more accurately estimating clonal frequencies. However, miFISH tracks copy numbers only at a limited number of probes per cell, in contrast to the whole-genome information offered by scSeq. Effective use of miFISH, thus, requires some knowledge of which genes might be important and has limited ability to allow one to discover novel variants of importance or to develop a comprehensive portrait of all variations in a tumor.

The present work is intended to make use of the complementary advantages of scSeq and miFISH for profiling clonal evolution by CNAs at the single-cell level, with the particular goal of better capturing WGD events and other sources of ploidy changes for which miFISH is ideal while also studying tumor evolution at the whole-genome as scSeq allows. For this purpose, we developed a joint clustering method, which pre-clusters each type of data to denoise single-cell copy number profiles and obtain clonal fractions, then synthesizes them to find the best joint matching based on a biclustering likelihood function. We applied our joint clustering method to simulated data and to two glioblastoma (GBM) cases for which both

miFISH and scSeq datasets are available. Results from simulated data confirm that our method can effectively jointly cluster cell populations profiled by both data types, allowing one to simultaneously reconstruct aspects of the clonal evolution process poorly captured by each data type alone. Results from the real GBM data demonstrate the ability of the method to successfully infer simultaneous evolution by focal and large-scale copy number aberrations and by WGD, showing results that are consistent with previous research on GBM (Beroukhi et al., 2007; Crespo et al., 2011) and other efforts to use miFISH to assist the deconvolution of bulk sequence data (Lei et al., 2020a,b) while providing new insights into how ploidy changes can drive the global evolution process in CIN tumors.

2. METHODS

The core of our contribution is a method to jointly cluster both scSeq and miFISH data gathered from distinct single cells of a common tumor cell population. This co-clustering is posed as an optimization problem through which we seek to identify a set of clones making up the tumor and assign each a whole-genome copy number profile, a ploidy, and an estimated clonal frequency. We then use these inferred clones as the input for phylogenetic inference of putative clonal lineage trees describing evolution of the clonal cell populations in the tumor. The whole process is illustrated in Figure 1. Table 1 shows a summary of definitions of key variables and model parameters used in defining the method.

For both miFISH (D_{scFISH}) and scSeq (D_{scSeq}) data, we first initialize the method by clustering each dataset separately with a conventional clustering algorithm. Here, we explore three options: k -means clustering, Gaussian mixture models (GMMs), and GMMs combined with uniform manifold approximation and projection (UMAP) (McInnes et al., 2018) for dimensionality reduction. After that, we obtain the separate cluster centers \mathbf{F} and \mathbf{S} and cluster frequencies \vec{f} and \vec{g} for miFISH and scSeq data.

We then define a joint likelihood function [Eqs. (1–3)] for the co-clustering problem. The goal of this problem is to infer a matrix A (Fig. 2) that identifies pairs of miFISH and scSeq cluster centers that are inferred to correspond to the same clone and assigns a clonal frequency to each such pair. An optimal solution of matrix A then yields an estimated population frequency for the clone corresponding to each miFISH-scSeq matched subgroup. For convenience in specifying the optimization, we also define an indicator matrix Z , which is a 0–1 matrix where a given entry z_{ij} is 1 when the corresponding entry of a_{ij} of A is nonzero and 0 when the corresponding entry a_{ij} of A is zero. Our goal is to find the configuration that maximizes the likelihood function:

$$\operatorname{argmax}_{A, Z} L(A, Z) = P(\mathbf{F}, \mathbf{S}, \vec{f}, \vec{g}; A, Z) \quad (1)$$

$$= P(\mathbf{F}, \mathbf{S}; A, Z) P(\vec{f}, \vec{g}; A, Z) \quad (2)$$

$$= P(\mathbf{F}, \mathbf{S}; Z) P(\vec{f}, \vec{g}; A), \quad (3)$$

where

$$P(\vec{f}, \vec{g}; A) = \left[\prod_{i=1}^m \left(\sum_{j=1}^n a_{ij} \right)^{f_i} \right] \left[\prod_{j=1}^n \left(\sum_{i=1}^m a_{ij} \right)^{g_j} \right]. \quad (4)$$

$P(\vec{f}, \vec{g}; A)$ shows the likelihood of the observed counts of cells in each subclone given the estimated actual frequencies. The overall proportions of different matched subgroups should sum up to one. Therefore, we apply the following constraint to the optimization problem:

$$\sum_{i=1}^m \sum_{j=1}^n a_{ij} = 1. \quad (5)$$

We further used a Gaussian noise model to account for errors in estimated copy number profiles, based on which we estimated a probability p_{ij} of observing the discrepancy between i th miFISH and j th scSeq profiles given that they are from the same tumor subclone. The statistical meaning of a Gaussian error distribution $\operatorname{erf}(x)$ is the probability of a Gaussian variable $X \sim \mathcal{N}(0, \frac{1}{2})$ falling in the range of $[-x, x]$. Its

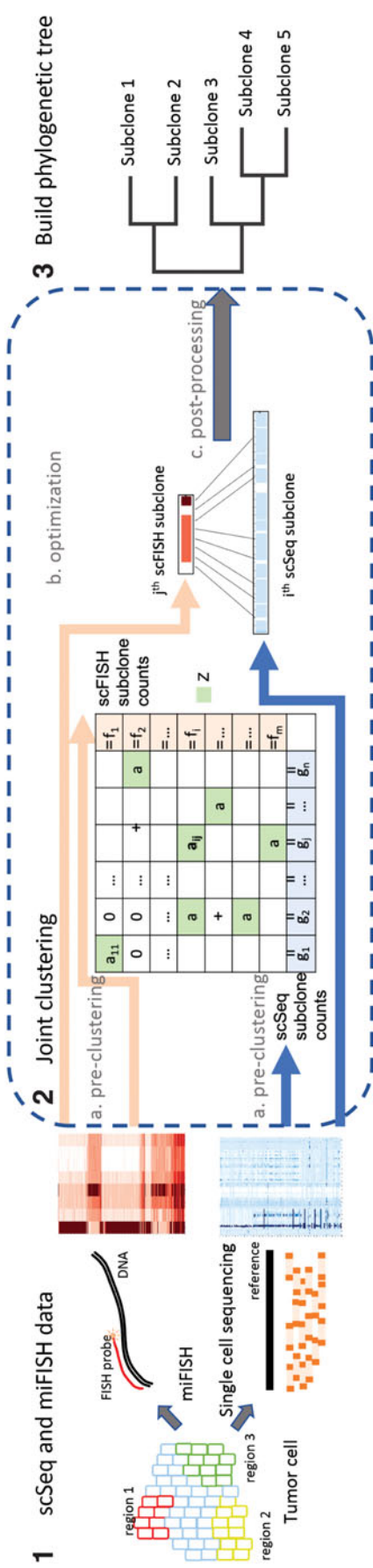


FIG. 1. Pipeline of the method. (1) We assume that single cells have been isolated from a tumor and samples of these cells have been profiled by scSeq and miFISH. (2) We then perform a joint clustering analysis in which we seek to infer a common set of clones consistent with both data sets. (3) After that, we conduct phylogenetic inference by MEDICC2 (Petkovic et al., 2021) to yield phylogeny inference that is able to take advantage of multiple scales of CNA events while working with whole-genome scale data. CNAs, copy number alterations; miFISH, multiplex interphase fluorescence in situ hybridization; scSeq, single-cell DNA-sequencing.

TABLE 1. PARAMETERS AND CONSTANTS FOR JOINT CLUSTERING

Notation	Meaning
$A^{m \times n}$	a_{ij} is the estimated frequency of the joint cluster of i th cluster of miFISH and j th cluster of scSeq
$Z^{m \times n}$	z_{ij} is a 0–1 indicator for a_{ij}
$\mathbf{F}^{m \times l}$	$\mathbf{F}_{i,p}$ is the miFISH profile of i th pre-cluster at FISH marker p
$\mathbf{S}^{n \times k}$	$\mathbf{S}_{j,q}$ is the scSeq profile j th pre-cluster at genome position q
$\vec{f}^{m \times 1}$	f_i is the miFISH frequency of i th pre-clusters
$\vec{g}^{n \times 1}$	g_j is the scSeq frequency of j th pre-clusters
$\mathbf{P}^{m \times n}$	p_{ij} is the probability matrix for matched clusters of i th miFISH and j th scSeq pre-cluster
$Probe\text{-}Indices^{l \times 1}$	The list of whole genome profile indices at FISH marker positions
m	The number of miFISH clusters
n	The number of scSeq clusters
l	The number of FISH markers
k	The number of genome features

FISH, fluorescence in situ hybridization; miFISH, multiplex interphase fluorescence in situ hybridization; scSeq, single cell DNA-sequencing.

complementary function $\text{erfc}(x)$ is used to estimate the probability p_{ij} of the observation of the difference between the profiles given the matching of i th scFISH and j th scSeq cluster ($z_{ij} = 1$):

$$\text{erfc}(x) = 1 - \text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^{+\infty} \exp(-t^2) dt. \quad (6)$$

We first normalize the miFISH cluster by half of its ploidy number, because it should always be diploid after normalization (denoted by \mathbf{F}^{ploidy}). We define the extent of discrepancy between the matched clusters given ploidy number $ploidy$ and probe index p [denoted as $(\mathbf{F}^{ploidy} \ominus \mathbf{S})_{ijp}$ in eq. (8)], specifically between the normalized profile of i th miFISH cluster at a specific probe p (denoted as $2\mathbf{F}_{i,p}/ploidy$) and j th scSeq cluster at the probe position [denoted as $\mathbf{S}_{j, ProbeIndices(p)}$] as the absolute distance. And then we apply the complementary Gaussian error function on it as an assessment of probabilities at every probe over all probable $ploidy \in [1, 2, \dots, 8]$ for miFISH cluster profile [eq. (9)] and choose the maximum probability as the final probability of matching clusters.

$$\mathbf{F}_{i,p}^{ploidy} = \frac{2\mathbf{F}_{i,p}}{ploidy} \quad (7)$$

$$(\mathbf{F}^{ploidy} \ominus \mathbf{S})_{ijp} = \left| \frac{2\mathbf{F}_{i,p}}{ploidy} - \mathbf{S}_{j, ProbeIndices(p)} \right|, \quad (8)$$

$$p_{ij} = \max_{ploidy} \left[\prod_p \text{erfc} \left(\frac{1}{\sigma} (\mathbf{F}^{ploidy} \ominus \mathbf{S})_{ijp} \right) \right]. \quad (9)$$

a_{11}	0						f_1
0	0					a_{2n}	f_2
		...					f_3
			0	0			...
			0	a_{ij}			f_i
				
		a_{m3}					f_m
g_1	g_2	g_3	...	g_j	...	g_n	

FIG. 2. Matrix A , which is the primary output of the co-clustering algorithm and identifies a set of inferred clones, each with corresponding miFISH and scSeq data, and their inferred population frequencies. Each nonzero entry a_{ij} indicates an inferred matching between the i th miFISH cluster and the j th scSeq cluster, proposing that they each at least in part derive from a common clone with clonal frequency a_{ij} . Zero entries of A indicate no such matching. Here, f_i and g_j are the frequencies of the pre-clustered miFISH cluster center i and scSeq cluster center j separately.

σ can be used to model the expected noise level of the sequence data, effectively regularizing the weight placed on the noise model relative to the likelihood of the clonal frequencies given the scSeq and fluorescence in situ hybridization (FISH) data. Larger σ means a more flatten curve of error function, where the matching can allow more randomness, whereas smaller σ sharpens the curve, which confines the matching more strictly depending on the difference.

The overall probability $P(\mathbf{F}, \mathbf{S}; Z)$ is estimated by the product of the probability p_{ij} where $z_{ij} = 1$.

$$P(\mathbf{F}, \mathbf{S}; Z) = \prod_{i=1}^n \prod_{j=1}^n p_{ij}^{z_{ij}}. \quad (10)$$

Therefore, the log likelihood function used for actual optimization is

$$\begin{aligned} l(A, Z) &= \log L(A, Z) \\ &= \sum_{i=1}^m f_i \log \left(\sum_{j=1}^n a_{ij} \right) + \sum_{j=1}^n g_j \log \left(\sum_{i=1}^m a_{ij} \right) + \sum_{i=1}^m \sum_{j=1}^n z_{ij} \log p_{ij}. \end{aligned} \quad (11)$$

We used Metropolis-Hasting sampling to sample the likelihood function (Algorithm 1). Because of the large search space for A , the sampling method is broken into two steps. First, it randomly samples a configuration given the previous configuration, based on a heuristic distribution. Then, it seeks to estimate the real-valued frequencies of the non-zero elements of probability matrix A so that the given observation of frequencies for separately clustered scFISH or scSeq subclones is most likely, that is, to maximize the likelihood function $\bar{L}(A) = P(\vec{f}, \vec{g}; A)$. Applying the Metropolis criterion to the likelihood ratio of $L(A, Z)$ and the proposal probability of the current sample and the previous sample, we are able to obtain a partial sample of the likelihood function from which we identify the configuration of A yielding the maximum likelihood.

In the Metropolis-Hasting sampling, we use the log-likelihood function $\bar{l}(A)$ in our implementation when solving for the most probable matrix A given indicator Z in the second step (eq. 12), where we adopt a regression model to estimate the parameters:

$$\begin{aligned} \max \bar{l}(A) &= \log \bar{L}(A) \\ &= \sum_{i=1}^m f_i \log \left(\sum_{j=1}^n a_{ij} \right) + \sum_{j=1}^n g_j \log \left(\sum_{i=1}^m a_{ij} \right). \end{aligned} \quad (12)$$

$$\text{s.t. } \sum_{i=1}^m \sum_{j=1}^n a_{ij} = 1. \quad (13)$$

Algorithm 1: Algorithm to match single-cell sequencing and single-cell FISH subgroups and identify frequencies of each matched subclones using Metropolis-Hasting sampling

Input: matching probabilities P , miFISH pre-clustered profiles \mathbf{F} and frequencies \vec{f} , scSeq pre-clustered profiles \mathbf{S} and frequencies \vec{g} , probe indices list *ProbeIndices*

Output: indicator matrix $Z^{(maxiter)}$ and corresponding matching matrix $A^{(maxiter)}$

```

Randomly initialize  $Z^{(0)}$ 
for  $t$  in  $0 : maxiter$  do
     $Z = Z^{(t)}$ ,  $A = A^{(t)}$ 
    Sample  $Z^* \sim Prop(Z'|Z)$ 
    Sample  $u \sim U[0, 1]$ 
    Solve  $A^*$  for optimization problem for  $\max \bar{l}(A)$  based on non-zero entries indicator matrix  $Z^*$ 
    if  $u < \min \left( 1, \frac{L(A^*, Z^*) Prop(Z|Z^*)}{L(A, Z) Prop(Z^*|Z)} \right)$  then
         $Z^{(t+1)} = Z^*$ ,  $A^{(t+1)} = A^*$ 
    else
         $Z^{(t+1)} = Z$ ,  $A^{(t+1)} = A$ 
    end if
end for

```

The estimated p_{ij} is also used for derivation of a transformation distribution as a proposal distribution $Prop(Z'|Z)$ during the sampling [Eqs. (14–17)]. The proposal distribution is used to sampling a position in the matrix to either increase or decrease a non-zero entry depending on the Z matrix in the previous step. We utilized a multinomial distribution with the parameters of normalized Gaussian error probabilities in each entry. Let \hat{p}_{ij} be the probability of Z of the previous step, where $\hat{p}_{ij} = \bar{p}_{ij}$ if the corresponding entry of Z is 0 and $\hat{p}_{ij} = 1 - \bar{p}_{ij}$ if the corresponding entry of Z is 1.

$$\bar{p}_{ij} = \sqrt{\frac{p_{ij}^2}{\sum_{i'=1}^m p_{i'j} \sum_{j'=1}^n p_{ij'}}}, \quad (14)$$

$$\bar{p}_{ij} \leftarrow \frac{\bar{p}_{ij}}{\sum_{i'=1}^m \sum_{j'=1}^n \bar{p}_{i'j'}}, \quad (15)$$

$$\hat{p}_{ij} = \bar{p}_{ij} \mathbf{1}_{z_{ij}=0} + (1 - \bar{p}_{ij}) \mathbf{1}_{z_{ij}=1}, \quad (16)$$

transition probability of $Z'|Z \sim$

$$\text{Mult}(mn, \vec{\pi} = (\hat{p}_{11}, \hat{p}_{12}, \dots, \hat{p}_{21}, \dots, \hat{p}_{ij}, \dots, \hat{p}_{mn})). \quad (17)$$

3. RESULTS

3.1. Validation on simulated data

We began our validation with simulated data to provide a comprehensive test of accuracy and model robustness on data of known ground truth. We first sought to evaluate the impact of different choices for existing pre-clustering methods for scSeq copy number data. We tested two different aspects of this issue: how well a method can identify the cluster number, and how well it can recover the true subclonal information. We test the first aspect by using different criteria for determining cluster number for different clustering methods. We generated miFISH and scSeq data by creating a series of model trees following (Zafar et al., 2019) and then separately simulating both data types from the common tree. This was repeated for 20 simulations of both scSeq and scFISH simultaneously with different tree structures and also different copy number mutations with mutation rate γ . We tested the methods by using $\gamma = 10, 20, 30$ to check the sensitivity for mutation rate (See Supplementary Methods S1 for more details on the simulation protocol.).

We used the measure $k_{\text{estimate}} - k_{\text{gt}}$ to evaluate the relative performance of different pre-clustering methods. We considered k -means, GMMs, and GMMs with UMAP feature reduction. The combination of UMAP and GMM proved superior at determining cluster numbers for both miFISH and scSeq across three mutation rates considered, with k -means generally somewhat better than GMM without UMAP (see Supplementary Fig. S1 for details).

We also tested the ability of each method to recover the true profiles of the cluster centers. Root mean square deviation of the copy number profiles and frequencies compared with the ground truth profiles and fractions were used for identifying the accuracy of clustering the single-cell copy numbers with different dimensions. By this measure, k -means proved the most successful method for both reconstructing the subclone profiles and estimating frequencies across mutation rates and data types. The GMM alone was generally superior to GMM+UMAP for miFISH data but inferior to scSeq (see Supplementary Fig. S2 for details).

We next applied the method to a test of the ability to draw correct inferences in the face of data noise. We examined two simulation cases: one in which miFISH data are error-free, which means that the ground truth miFISH profiles are the same as the scSeq profiles on their shared markers except in that they are not normalized by the ploidy. The other assumes that errors in miFISH probe counts are modeled by normal random perturbations in marker counts. We used precisions and recalls of the identified matchings as the evaluation matrix. The separate subclone profiles and fractions are obtained with three different pre-clustering methods applied to both single-cell data types. We also tested the method with the input of ground truth fractions and profiles for comparison. We further varied the mutation rates γ and a scaling factors σ used by the Gaussian error function to set sensitivity of the method to presumed noise in copy number estimates. The simulation is repeated 40 times for each set of parameters.

The overall results show that this method can identify matchings between the miFISH and scSeq measures of common subclones with high precision provided the pre-clustering is accurate (Figs. 3 and 4). Different pre-clustering methods affect the recalls and precisions to different extents, which are largely related to the way how the data are actually distributed. Joint clustering with the data pre-clustered by k -means yields similar results to joint clustering from the ground truth subclonal information, which is consistent with the prior observation that k -means is the most effective of the clustering methods considered

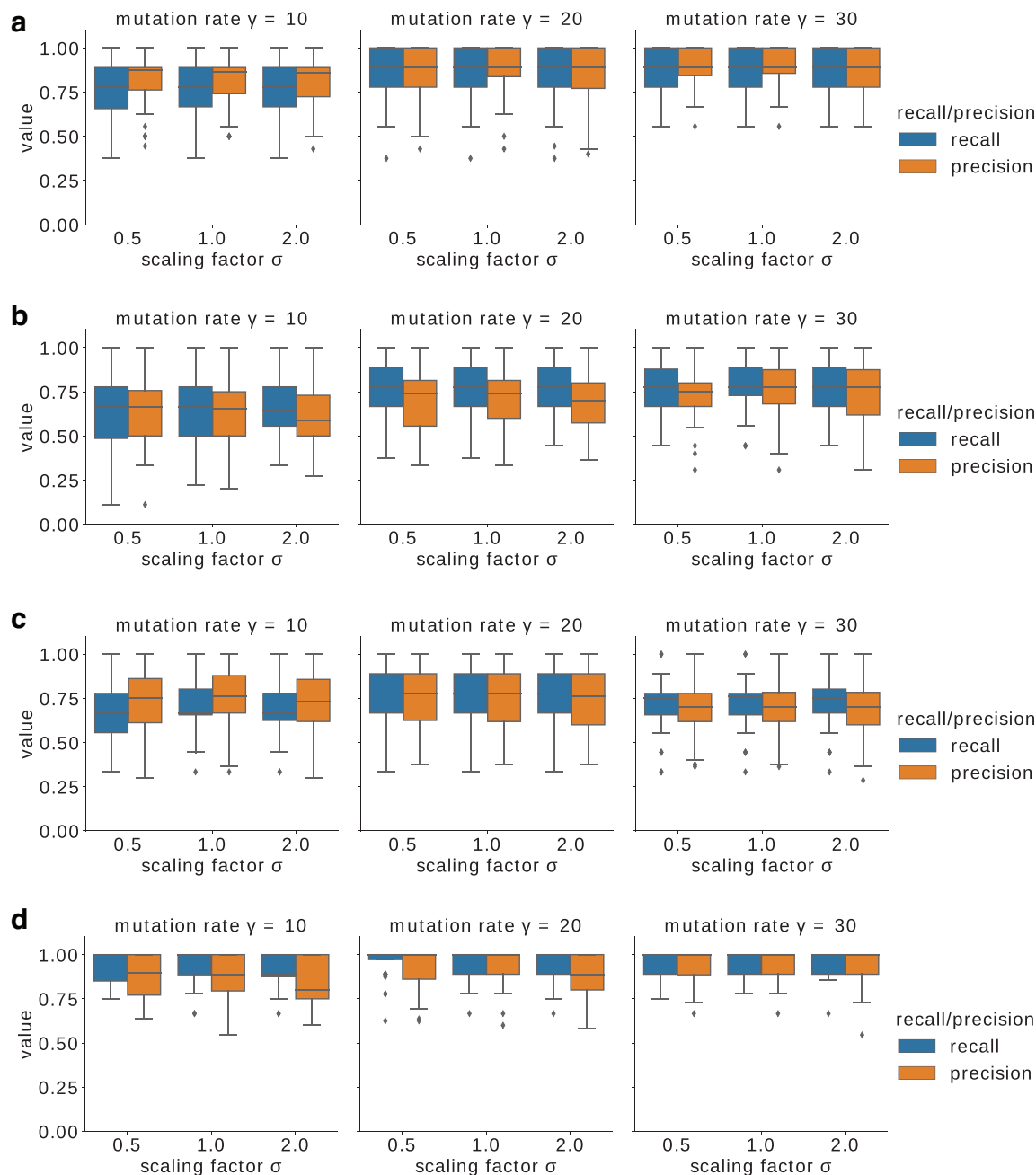


FIG. 3. Results from noise-free simulation data. Recalls and precisions are calculated from 40 sets of simulation data, each with different mutation rates $\gamma = 10, 20, 30$. The method was applied with different parameters $\sigma = 0.5, 1, 2$ with iterations=50,000. The simulated cluster number is 9. The structure is randomly generated. The inputs of the method are scFISH and scSeq profiles clustered by (a) k -means, (b) GMM, (c) UMAP+GMM (spherical for scSeq and diag for scFISH), and (d) ground truth profiles. In simulation, the copy number changes are not normalized. The perturbation rate is 0.05 for scSeq and 0.1 for miFISH. GMMs, Gaussian mixture models; UMAP, uniform manifold approximation and projection.

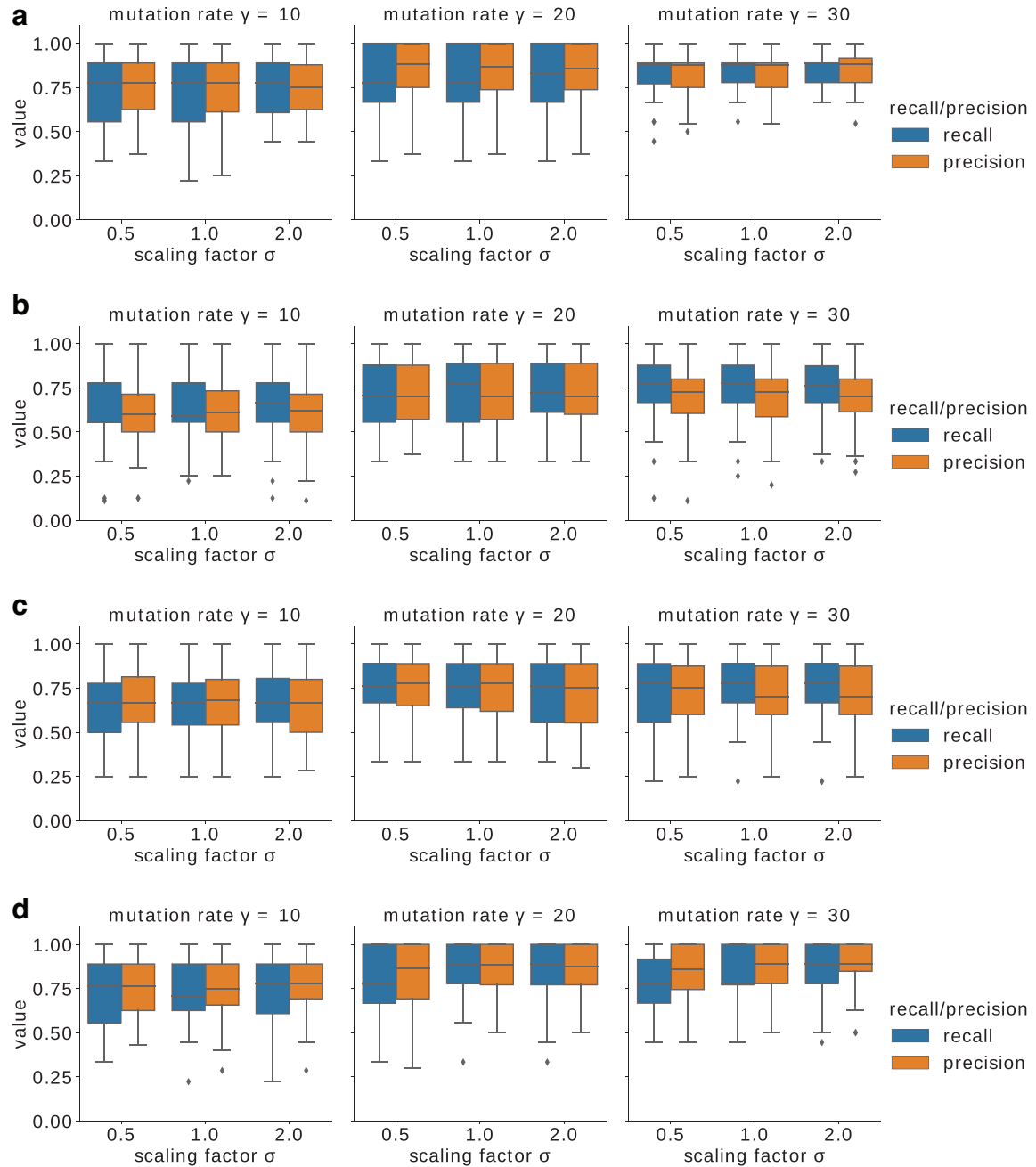


FIG. 4. Results from noisy simulation data. We use the same simulation protocol as in Figure 3, except that there is noise added to the ground truth scFISH data at markers sites with $\epsilon \sim \mathcal{N}(0, 0.3)$. Subfigure labels are defined identically to those in Figure 3.

at recapitulating the ground truth cluster assignments. We observe a decrease in performance when adding a noise model to miFISH counts relative to noise-free data, but the method is still able to jointly cluster both data types with a fairly good accuracy. Little sensitivity is observed to varying scaling factor in both cases, suggesting that Gaussian error function can give a good estimate of the joint generative probability even without a very precise estimate of the true data noise level.

Variations in mutation rate do produce larger changes in outcomes. Simulation data with low mutation rates tend to have lower recalls and precisions, mainly because low mutation rates make it more difficult to get a good pre-clustering where all the generated clusters are well distinguished. Also, lower mutation rate means a lower probability for the copy numbers at the FISH marker positions to be altered, resulting in less

distinct FISH subclones, which, in turn, introduces more imprecision during the joint clustering since no significant features can be identified. In the real data, FISH markers are usually selected based on prior knowledge of the particular tumor type. Therefore, even if the overall mutation rate is low, the FISH gene markers are still generally distinguishable in real single-cell tumor data. Also, although we set the number of markers to be 8, as in the available real data, miFISH can accommodate larger numbers of markers, with potential for improved ability to distinguish miFISH clusters and match them accurately to scSeq clusters.

3.2. Application to two GBM cases

We apply our method to two real GBM samples each taken from three tumor regions of a patient for whom scSeq and miFISH are available. One data set, GBM07, includes 202 single cells for sequencing and 450 single cells for FISH. The other, GBM33, includes 208 single cells for sequencing and 450 single cells for FISH in total (Wu et al., 2016). Each scSeq cell profile contains 9934 copy number features, representing copy numbers of consecutive regions spanning the whole genome. The FISH data contain copy numbers of the following gene markers: PDGFRA (4q), APC (5q), EGFR (7p), MET (7q), MYC (8q), CCND1 (11q), CHEK1 (11q), and ERG (21q). Both the miFISH and scSeq data can be inaccurate when the copy numbers are extremely high. Therefore, copy numbers over 20 for scFISH data and 10 for scSeq were manually capped to those upper bounds.

We first post-process the joint matching results by matching the clusters and calculating estimated pseudoploidies. Possible pseudoploidies that indicate the potential ratios between normalized scSeq at markers and miFISH profiles are calculated based on the matched pairs of clusters.

$$pseudoploidy_{ij} = \text{median}_p \frac{2\mathbf{F}_{i,p}}{\mathbf{S}_{j, ProbesIndices(p)}}. \quad (18)$$

where i th miFISH cluster and j th scSeq cluster are matched and p is the index of the FISH marker.

The actual whole genome copy number profile for each joint cluster is calculated by unnormalizing the scSeq profiles with the pseudoploidy.

The estimated real ploidy for each joint cluster and internal nodes is then computed by taking the mean of the copy numbers and rounding to the nearest integer. We then infer a branch to be WGD with a heuristic criterion based on the model that ploidy changes tends to follow a pattern of increase through doubling followed by reduction through cascading focal deletions. Under this model, we interpret ploidy change $2 \rightarrow 4$ as WGD, $2 \rightarrow 3$ as WGD followed by ploidy loss, $3 \rightarrow 4$ as ploidy loss followed by WGD, $3 \rightarrow 5$ as WGD followed by ploidy loss, $4 \rightarrow 6$ as ploidy loss followed by WGD, and $4 \rightarrow 8$ as WGD.

$$\mathbf{S}_{ij}^{unnormalized} = \mathbf{S}_j \times pseudoploidy_{ij} \quad (19)$$

$$ploidy_{ij} = \text{round}(\text{mean}(\mathbf{S}_{ij}^{unnormalized})). \quad (20)$$

Because of the high dimensionality of the scSeq data and the high noise of both types of single cell data, we use UMAP for feature reduction and then apply GMM and use Bayesian information criterion (BIC) to identify the optimal numbers of clusters for both data types, as shown in the simulation results. k -Means was then utilized to cluster both single-cell data given the optimal cluster numbers inferred via UMAP and GMM. We then applied our method to the pre-clustering results.

Figure 5 shows the inferred frequency matrices A for the two samples. At a high level, we observe clonal heterogeneity in both data sets with a few dominant clones together with a larger number of minor clones. We observe substantially more heterogeneity in GBM07 versus GBM33, reflected in a larger number of cluster centers by both data types, more diverse copy number aberrations in general, fewer widely shared mutations across clones, and a larger number of relatively rare clones inferred, consistent with our prior study of these data (Lei et al., 2020a). GBM33 also still keeps a relatively large diploid normal cell cluster, whereas GBM07 has a very small diploid normal cell cluster. Both, however, experience diverse chromosome-scale copy number aberrations.

To evaluate the utility of the clustering for tumor phylogenetics and assess its implications on clonal evolution, we apply MEDICC2 (Petkovic et al., 2021), which derives a clonal phylogeny of supplied clones and infers CNA profiles of ancestral internal nodes.

Figure 6a shows trees for the inferred centers of the joint clustering for GBM07. (The cluster centers themselves are visualized in Supplementary Fig. S3a.) For GBM07, the joint clustering and subsequent tree

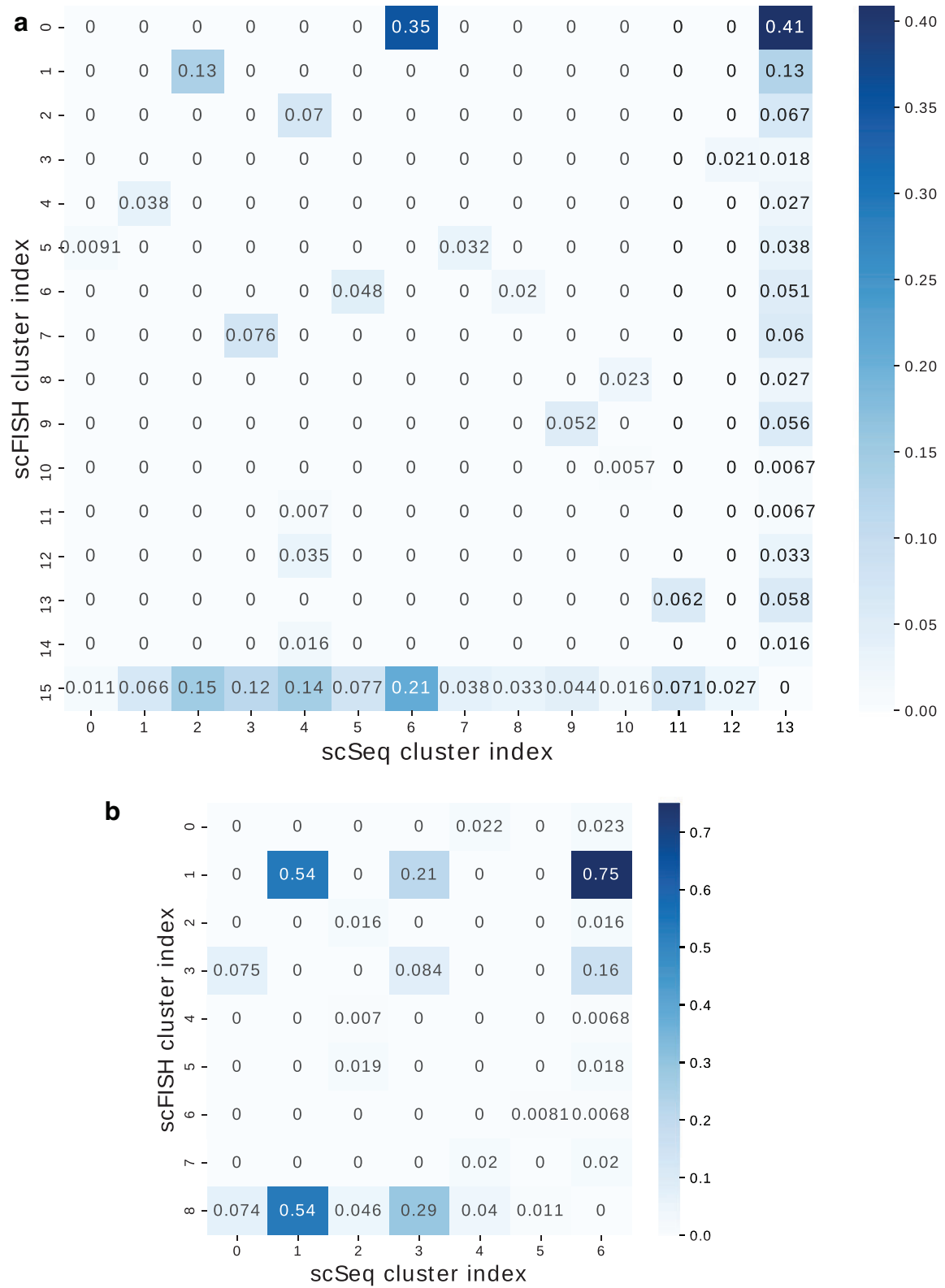


FIG. 5. The estimated frequency matrix A for GBM data sets **(a)** GBM07 and **(b)** GBM33. The last row and last column represent the fractions from separate clustering of scSeq and miFISH, respectively. GBM, glioblastoma.

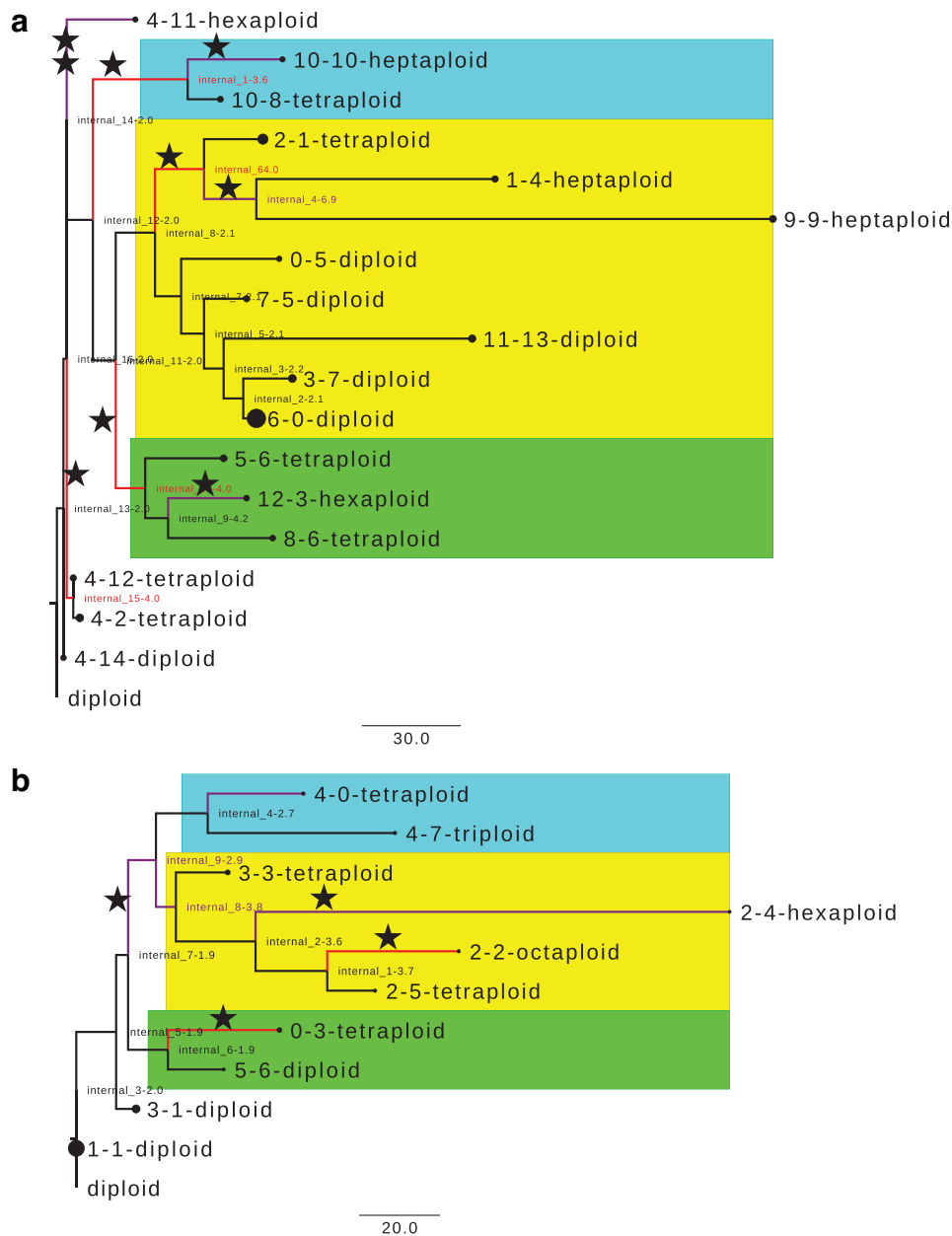


FIG. 6. GBM07 (a) and GBM33 (b) clonal lineage trees inferred from the joint cluster centers (normalized clusters profiles of scSeq multiplied by estimated pseudoploidies). The names of each leaf are in the form of normalized scSeq cluster index—scFISH cluster index—estimated ploidy. The diameter of the tip circle represents the fraction of the cell type among all the cells. The red branches show possible WGD events. The purple branches show possible WGD combined with ploidy loss. Stars show inferred WGD events. The inference was performed with scaling factor $\sigma = 1$ and k -means preclustering. Diploid node is the diploid root node. WGD, whole genome duplication.

inference suggests nine possible WGD events. Some of these may reflect errors in phylogenetic inference, although we would expect MEDICC2 to find a reasonably parsimonious explanation of the data. The most dominant individual clones are a diploid lineage inferred to have no ancestral WGD (6-0-diploid) and a tetraploid lineage inferred to have one ancestral WGD (2-1-tetraploid). Several more minor lineages show evidence of independent WGD events, and several individual clones show evidence of multiple WGD events along their evolution.

It has been shown in different cancer types that WGD can touch off a cascade of other copy number gains and especially losses (Zack et al., 2013). This pattern of WGD followed by cascading focal CNAs was previously inferred to be the source of patterns of triploidy and subsequent hexaploidy or pentaploidy

observed in prior studies of aneuploid tumors (Oltmann et al., 2018). These mechanisms explain the observation of triploid or heptaploid subclones, which we interpret to have likely resulted from massive copy number losses after a WGD event. For comparison, Figure 7a and b shows trees for the scFISH-only and scSeq-only clusters for GBM07, the former identifying some of the same major focal events but without capturing the global patterns of whole-genome variation and the latter yielding a quite different tree topology uninformed by WGD.

Examining the clusters in more depth shows copy gain of PDGFRA in the early branching 4-14-diploid cluster, before an early WGD in almost-euploid lineage, resulting in a tetraploid lineage (4-12-tetraploid and 4-2-tetraploid), suggesting that amplification of PDGFRA was an early event in the GBM. This conclusion is also suggested by the fact that most of the clusters have large copy numbers of PDGFRA. Whole chromosome 9 gain and chromosome 9p loss is a one common feature of blue-yellow-green lineage, also suggesting that these are early events. High amplification of PDGFRA is the common characteristics in the yellow-green lineage. Whole chromosome 11 loss and intermittent copy number gains in chromosome 11 are found in almost all of the subclones in yellow lineage. The diploid subtree of this lineage experiences a whole chromosome 8 gain during development, and 0-5-diploid has a unique whole chromosome 2 gain. It is inferred that one WGD event happens in a relatively early stage and subsequent WGDs later in the blue, yellow, and green lineages. Subclone 5-6-tetraploid has a unique whole-chromosome 21 and 22 gain. Other large-scale genome copy number aberrations are observed, such as loss of whole chromosome 10 and either whole chromosome loss of 15 or gain of whole chromosome 17 observed in subclones from the blue and green lineages, demonstrating substantial intra-tumor heterogeneity in copy number variations (Fujisawa et al., 2000; Beroukhi et al., 2007; Crespo et al., 2011).

Figure 6b shows the trees for the inferred centers of the joint clustering for GBM33. (The inferred cluster centers themselves are visualized in Supplementary Fig. S3b.) Although there is substantially less clonal heterogeneity than for GBM07, there are still four WGD events inferred in subclonal populations for GBM33. There is a major apparently normal clone (1-1-diploid) identified. Losses of whole chromosome 21; loss of chromosome 4p, 5q, 9p, 14p, and 18p are among the large-scale copy number mutations in the early stages of development (Supplementary Fig. S3b and Fig. 6b) before the first WGD. As with GBM07, it is possible the data could be explained by other trees but there is no plausible explanation that would not require multiple WGD events, including multiple events along single lineages to account for octaploid and hexaploid clones. 5-6-Diploid has a unique chromosome 3p loss, and a common chromosome 12p loss is identified in the blue-yellow lineage. Whole chromosome 13 loss was identified in blue and green lineages. These chromosomal aberrations are consistent with previous studies on GBM (Beroukhi et al., 2007).

For comparison, we provide trees for the scFISH-only and scSeq-only clusters for GBM33 as Figure 7c and d, with each lacking some features of the inferred pattern of clonal differentiation only observable by the combination of the two data types. Despite the differences between them, both tumors show some common features. Both show substantial clonal heterogeneity. They show a common amplification of PDGFRA as an early and important event in GBM genesis. Both GBM07 and GBM33 show different copy number gains in the early stages, exhibiting behavior consistent with the proneural subclass of GBM (Verhaak et al., 2010). Also, there are observations of common chromosome-scale copy number aberrations among both cases, such as the loss of chromosomes 4p and 9p.

One distinct feature for both cases compared with previous related studies (Lei et al., 2020b) is that the present study inferred a larger number of WGD events than that prior work. This might be the result of very low fractions of the tetraploid or octaploid subclones compared with major diploid subclones, which the present work can detect, whereas the prior works depends on deconvolution with bulk data as well, an approach that is only sensitive to relatively frequent subclones.

4. DISCUSSION

Single-cell technologies, including scSeq and miFISH, are becoming more widely used in tumor studies because they can capture intra-tumor heterogeneity with much finer grain than older bulk genomic methods, yet these technologies still bring technical challenges. We proposed a method combining both scSeq and miFISH for improved clonal inference from copy number mutations by leveraging the relative advantages of each technology. Results on simulated data show that the combination can achieve high accuracy in capturing both the whole-genome variant data of scSeq and better resolution for ploidy variations offered by miFISH.

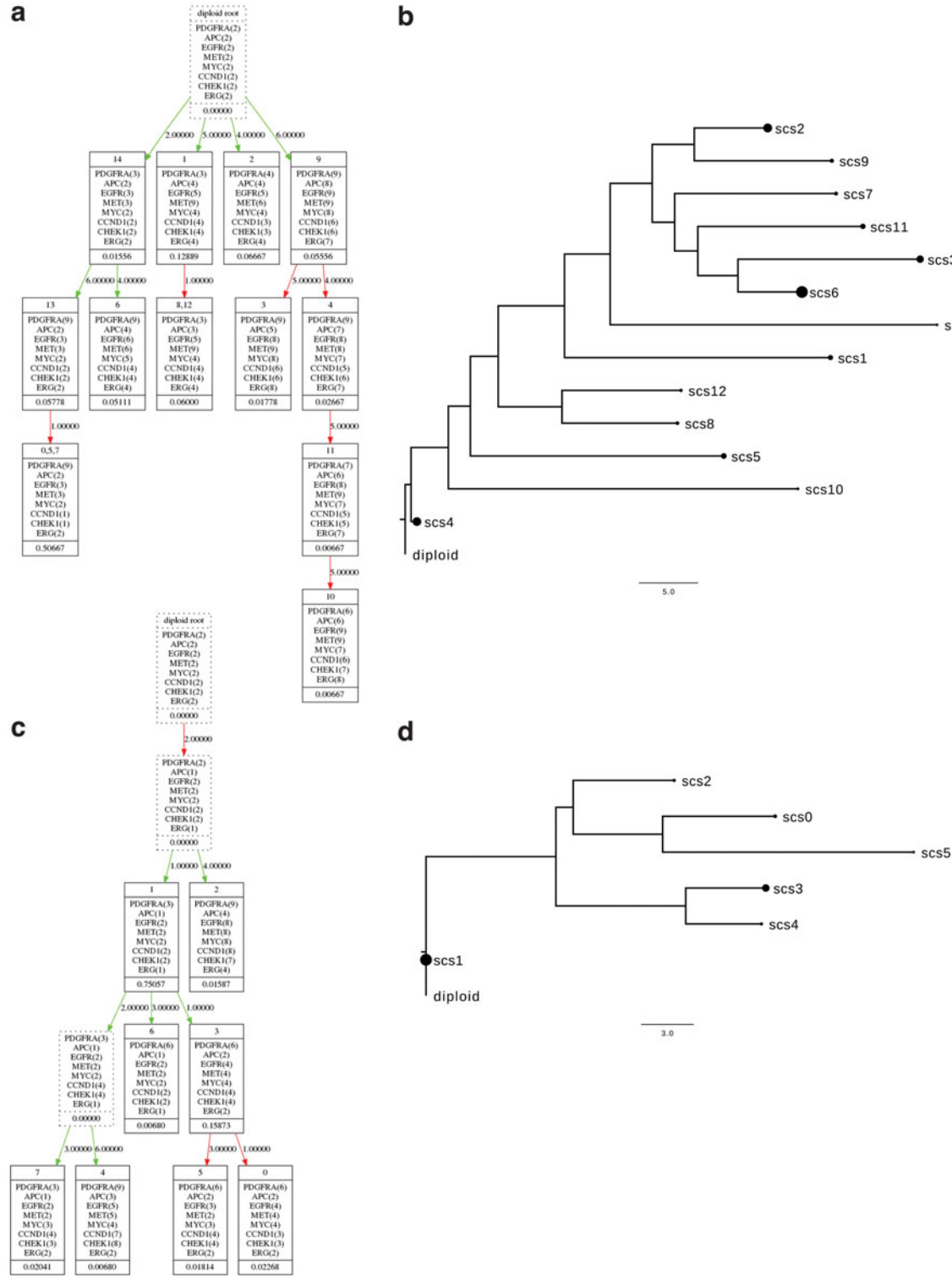


FIG. 7. GBM clonal lineage trees from miFISH or scSeq data alone. GBM07 clonal lineage trees inferred from the (a) miFISH and (b) scSeq cluster centers. GBM33 clonal lineage trees inferred from the (c) miFISH and (d) scSeq cluster centers. miFISH trees are produced by FISHtrees (Gertz et al., 2016) and scSeq trees via MEDICC2 (Petkovic et al., 2021). For miFISH trees, the red branches show possible WGD events and green branches focal changes. For scSeq trees, the diameter of each tip circle represents the fraction of the cell type among all the cells. The inference was performed with scaling factor $\sigma = 1$ and k -means preclustering. The node labeled “diploid” is a presumed diploid root node.

We also apply our method to two GBM cases and demonstrate the ability of the joint data to reconstruct a more comprehensive model of clonal heterogeneity in these cancers. Application of the cluster results in clonal tree inference reveals a model of focal copy number aberrations occurring alongside repeated chromosome-scale and WGD mutation events. Consistent with our prior analysis of these data sets (Lei et al., 2020a) and earlier models of persistent CIN (Storchova and Pellman, 2004), we find that WGD is not a one-time event but rather an ongoing process in these tumors that may act on distinct lineages or multiple times in single-cell lineages within a tumor. In fact, the present method, which offers a higher sensitivity for the detection of relatively rare clones than prior approaches to WGD-sensitive clonal inference, suggests an even greater activity of these mechanisms. Given the apparent importance of these events in touching off cascades of rapid clonal evolution, the results confirm the crucial importance of incorporating accurate models of CNAs at all scales into tumor evolution studies and the value of integrating multiple data sources, notably miFISH or similar methods, in that process.

This work can be further extended in some aspects. First, the model can be modified to be adaptable for noisy single-cell whole-genome sequencing and miFISH data with different types. Currently, there is a lack of computational tools that are able to cluster high-dimensional whole-genome copy numbers of single tumor cells with high accuracy. Better pre-clustering for both scFISH and scSeq for copy numbers could improve the accuracy for joint clustering. Second, the model is computationally expensive and prone to converge to local optimum for highly heterogeneous tumors. Markov Chain Monte Carlo (MCMC) sampling combined with a prior distribution on the indicator matrix Z or other techniques might be applied to speed up the algorithm and get global optimum. Moreover, the incorporation of other types of data from different biotechnologies can provide evidence for single cells from different sources to be grouped together. Alternative ways of measuring absolute ploidy may have advantages over miFISH, such as facilitating clinical application, despite tradeoffs. Integration of other types of genome aberrations such as SNVs and structural variations (SVs), as well as other genomic data, such as RNA-seq or epigenetic data, could help us develop more accurate and comprehensive models of clonal evolution and its connections to cellular phenotypes.

DATA AVAILABILITY

<https://github.com/CMUSchwartzLab/JointClustering>

ACKNOWLEDGMENTS

The authors thank Guibo Li, Xulian Shi, Liqin Xu, Yong Hou, and Kui Wu for the use of sequence data analyzed in this work.

DISCLAIMER

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The Pennsylvania Department of Health specifically disclaims the responsibility for any analyses, interpretations, or conclusions.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no competing financial interests.

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 US NIH award R21CA216452 and Pennsylvania Dept. of Health awards 4100070287 and FP00003273. Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award number R01HG010589.

SUPPLEMENTARY MATERIAL

Supplementary Data
 Supplementary Figure S1
 Supplementary Figure S2
 Supplementary Figure S3

REFERENCES

- Bakhom, S.F., Ngo, B., Laughney, A.M., et al. 2018. Chromosomal instability drives metastasis through a cytosolic DNA response. *Nature* 553, 467–472.
- Beroukhi, R., Getz, G., Nghiemphu, L., et al. 2007. Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc. Natl Acad. Sci. U. S. A.* 104, 20007–20012.
- Bielski, C.M., Zehir, A., Penson, A.V., et al. 2018. Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat. Genet.* 50, 1189–1195.
- Chowdhury, S.A., Gertz, E.M., Wangsa, D., et al. 2015. Inferring models of multiscale copy number evolution for single-tumor phylogenetics. *Bioinformatics* 31, i258–267.
- Chowdhury, S.A., Shackney, S.E., Heselmeyer-Haddad, K., et al. 2013. Phylogenetic analysis of multiprobe fluorescence in situ hybridization data from tumor cell populations. *Bioinformatics* 29, i189–i198.
- 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.
- 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.
- Fujisawa, H., Reis, R.M., Nakamura, M., et al. 2000. Loss of heterozygosity on chromosome 10 is more extensive in primary (de novo) than in secondary glioblastomas. *Lab. Invest.* 80, 65–72.
- Gertz, E.M., Chowdhury, S.A., Lee, W.-J., et al. 2016. FISHtrees 3.0: Tumor phylogenetics using a ploidy probe. *PLoS One* 11, e0158569.
- Heng, H.H., Bremer, S.W., Stevens, J.B., et al. 2013. Chromosomal instability (CIN): What it is and why it is crucial to cancer evolution. *Cancer Metastasis Rev.* 32, 325–340.
- Lei, H., Lyu, B., Gertz, E.M., et al. 2020a. Tumor copy number deconvolution integrating bulk and single-cell sequencing data. *J. Comput. Biol.* 27, 565–598.
- Lei, H., Gertz, E.M., Schäffer, A., et al. 2020b. Tumor heterogeneity assessed by sequencing and fluorescence in situ hybridization (FISH) data. *bioRxiv* 2020.02.29.970392.
- Letouze, E., Allory, Y., Bollet, M.A., et al. 2010. Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. *Genome Biol.* 11(Suppl 1), P25.
- McInnes, L., Healy, J., and Melville, J. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint; arXiv.1802.03426*.
- Navin, N., Kendall, J., Troge, J., et al. 2011. Tumour evolution inferred by single-cell sequencing. *Nature* 472, 90–94.
- Nowell, P. 1976. The clonal evolution of tumor cell populations. *Science* 194, 23–28.
- Oltmann, J., Heselmeyer-Haddad, K., Hernandez, L.S., et al. 2018. Aneuploidy, tp53 mutation, and amplification of myc correlate with increased intratumor heterogeneity and poor prognosis of breast cancer patients. *Genes Chromosomes Cancer*. 57, 165–175.
- Pennington, G., Smith, C.A., Shackney, S., et al. 2007. Reconstructing tumor phylogenies from heterogeneous single-cell data. *J. Bioinform. Comput. Biol.* 5, 407–427.
- Petkovic, M., Watkins, T.B., Colliver, E.C., et al. 2021. Whole-genome doubling-aware copy number phylogenies for cancer evolution with MEDICC2. *bioRxiv* 2021.02.28.433227.
- Sainudiin, R., and Veber, A. 2015. A beta-splitting model for evolutionary trees. *R. Soc. Open Sci.* 3, 160016.
- Satas, G., Zaccaria, S., Mon, G., et al. 2020. SCARLET: Single-cell tumor phylogeny inference with copy-number constrained mutation losses. *Cell Syst.* 10, 323–332.
- Schwartz, R., and Schäffer, A.A. 2017. The evolution of tumour phylogenetics: Principles and practice. *Nat. Rev. Genet.* 18, 213–229.

- Schwarz, R.F., Trinh, A., Sipos, B., et al. 2014a. Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Comput. Biol.* 10, 1003535.
- Schwarz, R.F., Trinh, A., Sipos, B., et al. 2014b. Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Comput. Biol.* 10, e1003535.
- Shackney, S.E., Singh, S.G., Yakulis, R., et al. 1995. Aneuploidy in breast cancer: A fluorescence in situ hybridization study. *Cytometry* 22, 282–291.
- Storchova, Z., and Pellman, D. 2004. From polyploidy to aneuploidy, genome instability and cancer. *Nat. Rev. Mol. Cell Biol.* 5, 45–54.
- Van de Peer, Y., Mizrachi, E., and Marchal, K. 2017. The evolutionary significance of polyploidy. *Nat Rev Genet.* 18, 411–424.
- Verhaak, R.G., Hoadley, K.A., Purdom, E., et al. 2010. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110.
- 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.
- Zaccaria, S., and Raphael, B.J. 2020. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. *Nat. Biotechnol.* 39, 1–8.
- 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., Navin, N., Chen, K., et al. 2019. SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome Res.* 29, 1847–1859.
- Zeira, R., and Raphael, B.J. 2020. Copy number evolution with weighted aberrations in cancer. *Bioinformatics* 36, i344–i352.
- Zeira, R., Zehavi, M., and Shamir, R. 2017. A linear-time algorithm for the copy number transformation problem. *J. Comput. Biol.* 24, 1179–1194.

Address correspondence to:

Dr. Russell Schwartz
Department of Biological Sciences
Carnegie Mellon University
4400 Fifth Avenue
Pittsburgh, PA 15213
USA

E-mail: russells@andrew.cmu.edu