JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 25, Number 7, 2018
© Mary Ann Liebert, Inc.
Pp. 689–708
DOI: 10.1089/cmb.2017.0253

Phylogenetic Copy-Number Factorization of Multiple Tumor Samples

SIMONE ZACCARIA,[1,2,*] MOHAMMED EL-KEBIR,[1,*]
GUNNAR W. KLAU,[3] and BENJAMIN J. RAPHAEL[1,**]

ABSTRACT

Cancer is an evolutionary process driven by somatic mutations. This process can be represented as a phylogenetic tree. Constructing such a phylogenetic tree from genome sequencing data is a challenging task due to the many types of mutations in cancer and the fact that nearly all cancer sequencing is of a bulk tumor, measuring a superposition of somatic mutations present in different cells. We study the problem of reconstructing tumor phylogenies from copy-number aberrations (CNAs) measured in bulk-sequencing data. We introduce the Copy-Number Tree Mixture Deconvolution (CNTMD) problem, which aims to find the phylogenetic tree with the fewest number of CNAs that explain the copy-number data from multiple samples of a tumor. We design an algorithm for solving the CNTMD problem and apply the algorithm to both simulated and real data. On simulated data, we find that our algorithm outperforms existing approaches that either perform deconvolution/factorization of mixed tumor samples or build phylogenetic trees assuming homogeneous tumor samples. On real data, we analyze multiple samples from a prostate cancer patient, identifying clones within these samples and a phylogenetic tree that relates these clones and their differing proportions across samples. This phylogenetic tree provides a higher resolution view of copy-number evolution of this cancer than published analyses.

Keywords: copy-number aberrations, factorization, integer linear programming, intratumor heterogeneity, multiple tumor samples, tumor phylogeny.

1. INTRODUCTION

Cancer results from an evolutionary process where somatic mutations accumulate in a population of cells during the lifetime of an individual (Nowell, 1976). Thus, a tumor consists of heterogeneous subpopulations of cells, or *clones*. Each clone comprises cells that share a unique complement of somatic mutations. Quantifying this intratumor heterogeneity has been shown to be important in cancer treatment (Venkatesan and Swanton, 2015). While intratumor heterogeneity complicates the identification of mutations

[1] Department of Computer Science, Princeton University, Princeton, New Jersey.
[2] Dipartimento di Informatica Sistemistica e Comunicazione (DISCo), Università degli Studi di Milano-Bicocca, Milan, Italy.
[3] Algorithmic Bioinformatics, Heinrich Heine University, Düsseldorf, Germany.

*Joint first authorship.

**Corresponding author.

in bulk-sequencing data from a tumor sample containing millions of cells, it also provides a signal for inferring the tumor composition—the number and proportion of clones within a sample—as well as the ancestral history of somatic mutations during cancer development (Gerlinger et al., 2012). Thus, a number of methods have been developed to infer phylogenetic trees from DNA sequencing data from one or more samples of a tumor (Gerlinger et al., 2012; Nik-Zainal et al., 2012; Deshwar et al., 2015; Gundem et al., 2015; Malikic et al., 2015; Sottoriva et al., 2015; El-Kebir et al., 2016b; Jiang et al., 2016; McPherson et al., 2016).

One class of mutations that are particularly useful for inferring tumor composition and tumor evolution are copy-number aberrations (CNAs), which include duplications and deletions of large genomic regions. CNAs are ubiquitous in solid tumors and can be readily detected from DNA sequencing data, making them good candidates for phylogenetic analysis. However, there are two major challenges in using CNAs to quantify intratumor heterogeneity and evolution.

The first challenge is that nearly all cancer sequencing studies perform bulk sequencing, where mutations are measured in tumor samples composed of mixtures of millions of different cells. While single-cell sequencing provides a higher resolution measurement of tumor heterogeneity, it remains a specialized technique that is cost prohibitive and error prone for whole-genome analysis of thousands of cells (Gawad et al., 2016). Thus, we require techniques to deconvolve CNA measurements from mixed tumor samples. Typically, CNAs are detected in sequencing data by examining the depth of aligned sequencing reads to genomic regions. More specifically, segmentation algorithms use this signal to partition the genome into *segments* with the same *integer* copy number (Baumbusch et al., 2008; Van Loo et al., 2010). When a sample is heterogeneous, that is, composed of a mixture of distinct clones, a *fractional* copy number may be obtained for each segment instead of an integer copy number.

A number of methods have been developed to infer tumor composition from fractional copy numbers, taking advantage of the fact that larger CNAs perturb thousands-millions of sequencing reads, providing a signal to infer their proportions, even with modest coverage sequencing (Van Loo et al., 2010; Carter et al., 2012; Nik-Zainal et al., 2012; Oesper et al., 2013; Fischer et al., 2014; Ha et al., 2014). However, these methods have certain limitations that limit their applicability and performance. For example, ASCAT (Van Loo et al., 2010) and ABSOLUTE (Carter et al., 2012) use the data from heterogeneous samples for inferring the tumor purity (the proportion of normal clone in a sample), but they do not distinguish the copy numbers of different tumor clones. Other methods, such as THetA (Oesper et al., 2013), Battenberg (Nik-Zainal et al., 2012), cloneHD (Fischer et al., 2014), and TITAN (Ha et al., 2014), infer the clonal composition independently for each sample by deconvolving the fractional copy numbers into the integer copy numbers of the extant clones and their proportions. However, one can obtain more information by jointly considering more samples from the same tumor (Gerlinger et al., 2012), as successfully done for single-nucleotide mutations (Deshwar et al., 2015; Malikic et al., 2015; El-Kebir et al., 2016b; Jiang et al., 2016) or noninteger copy numbers (Roman et al., 2016). Moreover, there may be multiple ways to deconvolve fractional copy numbers, especially without imposing a structure on the inferred CNAs. Therefore, the inference of distinct clones may benefit from jointly inferring their evolution.

The second challenge in using CNAs to reconstruct tumor evolution is that one requires a model of the evolution of CNAs. Defining such a model is not straightforward because CNAs can overlap, and thus, positions in the genome cannot be treated independently. Standard phylogenetic models represent a genome as a sequence of ‘‘characters’’ with mutations acting independently on individual characters. A number of models have been introduced to study CNA evolution, and these models can be classified into two categories. The first considers *single events* such that each of those independently affects the copy number of a single segment (Chowdhury et al., 2014; McPherson et al., 2016). However, these models do not account for dependency between adjacent segments in the genome. The second category considers the effects of CNAs on multiple segments as *interval events* that amplify or delete copies of contiguous segments; the most prominent such approach is MEDICC (Schwarz et al., 2014). Recently, Shamir et al. (2016), El-Kebir et al. (2016a), and El-Kebir et al. (2017) improved the model in MEDICC. Specifically, Shamir et al. (2016) formally investigated the effects of interval events on segments of a single clone. In El-Kebir et al. (2016a) and El-Kebir et al. (2017), the authors formalized the *Copy-Number Tree (CNT)* problem that aims to find the most parsimonious evolution of clones explained by the minimum number of interval events, and derived an integer linear programming (ILP) that solves this problem. However, all of the studies applying these methods either assume that each sample is homogeneous and consisting of a single clone (Schwarz et al., 2015; Sottoriva et al., 2015; Mangiola et al., 2016) or first attempt to infer the clones independently on each sample before performing a phylogenetic analysis of CNAs (McPherson et al., 2016).

In this article, we propose an approach combining the deconvolution of fractional copy numbers from multiple samples with the inference of CNAs that describes the evolution of the clones. We introduce the *Copy-Number Tree Mixture Deconvolution (CNTMD)* problem that aims to deconvolve the fractional copy numbers into the integer copy numbers of the extant clones and their proportions such that the evolution of the clones is explained by a minimum number of CNAs modeled as interval events (Fig. 1). CNTMD generalizes two problems that have been researched intensively in recent years: deconvolution/factorization problems that aim to infer the composition of a tumor sample, including the proportions of tumor subclones; and phylogenetic models of copy number evolution that model the dependencies between copy number events that affect the same genomic loci. We design a coordinate-descent algorithm for solving this problem and we compare our method with alternative approaches on real-size simulations. We find that combining the deconvolution of fractional copy numbers with a phylogenetic tree outperforms other methods. We apply our method on multisample sequencing data of a prostate cancer patient (Gundem et al., 2015). Our inference shows well-supported patterns that reveal the clonal composition in terms of CNAs. The software is available at https://github.com/raphael-group/CNT-MD.

2. COPY-NUMBER TREE MIXTURE DECONVOLUTION PROBLEM

We start by reviewing the CNT problem, where given integer copy-number profiles one is asked to infer a *CNT*, whose leaves correspond to the profiles with the minimum of events. Specifically, we define the interval events that label the edges of this tree. We conclude this section by introducing the problem of deconvolving fractional copy numbers from multiple heterogeneous samples into integer copy-number profiles of distinct clones and their proportions such that the resulting profiles form the leaves of a parsimonious CNT.

2.1. *Profiles and events*

Following the model in Schwarz et al. (2014), Shamir et al. (2016), El-Kebir et al. (2016a), and El-Kebir et al. (2017), we represent a chromosome as a sequence of *m segments*. A *copy-number profile*, or *profile* for short, specifies the number of copies of each segment in a clone. Formally, a profile $\mathbf{c}_i = [c_{s,i}]$ is a (column) vector of *m* integers whose entries $c_{s,i} \in \mathbb{N}$ indicate the number of copies of segment *s* in a clone *i*. For brevity, we consider a single chromosome in the definitions.

We consider mutations that amplify or delete contiguous segments. An *interval event*, or *event* for short, increases or decreases the copy numbers of contiguous segments of a profile $\mathbf{c}_i$. Formally, an event is a triple (s, t, b) with segments $s \leq t$ and integer $b \in \mathbb{Z}$. If *b* is positive, then the event is an *amplification* and

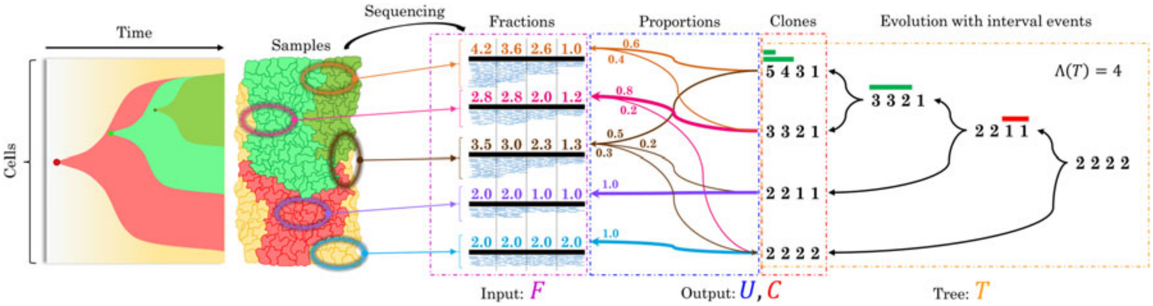

FIG. 1. CNTMD problem. A tumor acquires somatic mutations over time and thus consists of heterogeneous subpopulations of cells, or clones. The tumor clones are colored in red, dark and light green, whereas the normal clone is colored in yellow. Five samples are sequenced with bulk-sequencing technology yielding fractional copy-number matrix *F*. We model the evolution of the CNAs by a CNT *T* (right), whose vertices are labeled by integer copy numbers, whose leaves correspond to clones, and whose edges are labeled by interval events that correspond to amplifications (green) or deletions (red) of an interval in the copy-number profile. We combine the deconvolution of these fractional copy numbers (*F*) with the inference of the evolution of clones (*T*). More specifically, CNTMD deconvolves/factors *F* into the integer copy numbers *C* of the extant clones and their proportions *U* such that $F = CU$ and *C* generate a CNT *T* with the minimum number $\Lambda(T)$ of interval events. CNA, copy-number aberration; CNT, copy-number tree; CNTMD, Copy-Number Tree Mixture Deconvolution.

the nonzero segments between s and t are incremented by b, whereas for negative b, the events are a *deletion* and the same segments are decremented by at most $|b|$ (segments must be non-negative). Thus, the event (s, t, b) applied on $\mathbf{c}_i = [c_{\ell,i}]$ results in $\mathbf{c}'_i = [c'_{\ell,i}]$ such that for each segment ℓ,

$$c'_{\ell,i} = \begin{cases} \max\{c_{\ell,i} + b, 0\}, & \text{if } s \leq \ell \leq t \text{ and } c_{\ell,i} \neq 0, \\ c_{\ell,i} & \text{otherwise.} \end{cases} \tag{1}$$

Thus, once a segment ℓ has been lost, that is, $c_{\ell,i} = 0$, it can never be regained (or deleted).

2.2. The copy-number tree problem

We model the evolutionary process that led to n extant tumor clones by a *CNT T* defined as follows.

Definition 1 *Given a number n of clones, a* CNT *is a rooted full binary tree on n leaves, such that each vertex $v_i \in V(T)$ is labeled by a profile $\mathbf{c}_i$ and each edge (v_i, v_j) is labeled by a set $\mathcal{E}_{i,j}$ of events. The root vertex $r(T)$, corresponding to the* normal clone, *is diploid, that is, $c_{s,r(T)} = 2$ for each segment s.*

The requirement that T is a full binary tree is without loss of generality, as each vertex with out-degree greater than 2 of a general tree can be split into vertices of out-degree 2, and each vertex with out-degree 1 can be removed and the associated events assigned to the outgoing edge. Thus, each vertex $v_i \in V(T)$ has either zero or two children and is labeled by a profile $\mathbf{c}_i$. To avoid degenerate solutions, we impose a maximum copy number $c_{\max} \in \mathbb{N}$ for each segment s of any vertex v_i of T such that $c_{s,i} \leq c_{\max}$. Moreover, each leaf $v_i \in L(T)$ corresponds to the clone i. As such, we order the vertices $V(T) = \{v_1, \ldots, v_{2n-1}\}$ such that $L(T) = \{v_1, \ldots, v_n\}$ and $r(T) = v_{2n-1}$. An edge $(v_i, v_j) \in E(T)$ relates a parent vertex v_i to its child v_j such that the label $\mathcal{E}(i,j)$ is a set of events that transform $\mathbf{c}_i$ to $\mathbf{c}_j$. In general, the order in which events $\mathcal{E}(i,j)$ are applied matters. Following a result by Shamir et al. (2016), it suffices to consider an unordered set of events instead of an ordered sequence. In fact, any sequence of events, where amplifications and deletions occur in an arbitrary order, can be transformed into a *sorted* sequence, where deletions are followed by amplifications, without changing the cost of events, as defined in the following. The cost of an event (s, t, b) is the number of changes in the segment and is thus equal to $|b|$. Therefore, the cost $\Lambda(i,j)$ of an edge (v_i, v_j) is the total cost of the events in $\mathcal{E}(i,j)$, that is, $\Lambda(i,j) = \sum_{(s,t,b) \in \mathcal{E}(i,j)} |b|$. The cost $\Lambda(T)$ of the tree T is the sum of the costs of all edges, that is, $\Lambda(T) = \sum_{(v_i, v_j) \in E(T)} \Lambda(i,j)$.

In the ideal case of single-cell sequencing data with no errors, each clone is a single cell and we observe the copy-number profiles $\mathbf{c}_1, \ldots, \mathbf{c}_n$ of n tumor clones. As such, we wish to find the most parsimonious explanation, that is, a minimum-cost CNT T^* whose n leaves are labeled by $\mathbf{c}_1, \ldots, \mathbf{c}_n$. Previously, we have shown that this problem, the CNT problem, is NP-hard and we introduced an ILP formulation for solving it (El-Kebir et al., 2016a, 2017). However, with bulk-sequencing data, the observations correspond to k *samples* obtained from a single tumor in different regions or at different time points. Each sample corresponds to a *mixture* of n extant clones (leaves) of an unknown CNT in unknown proportions. Recall that m is the number of segments. Our observations are thus described by the $m \times k$ *fractional copy-number matrix* $F = [f_{s,p}]$ where the *fraction* $f_{s,p} \in \mathbb{R}_{\geq 0}$ is the average copy number of segment s in sample p.

Let T be a CNT with n leaves. We represent the profiles of the clones of T by the $m \times n$ *copy-number matrix* $C = [c_{s,i}]$ such that the i-th column of C corresponds to the profile $\mathbf{c}_i$ of clone i, that is, $C = (\mathbf{c}_1, \ldots, \mathbf{c}_n)$. We say that C *generates* C if the leaves of T are labeled by the profiles in C and such that each internal vertex v_i is labeled by a profile $\mathbf{c}_i = [c_{s,i}]$ with $c_{s,i} \leq c_{\max}$ for each segment s. The $n \times k$ *usage matrix* $U = [u_{i,p}]$ describes the *mixing proportion* $u_{i,p} \in \mathbb{R}_{\geq 0}$ of clone i in sample p such that the sum $\sum_{1 \leq i \leq n} u_{i,p}$ of the mixing proportions for each sample p is 1. The observed fractional copy numbers F are thus modeled by $F = CU$. We have the following problem (Fig. 1).

Problem 1 (Copy-Number Tree Mixture Deconvolution (CNTMD)) *Given an $m \times k$ fractional copy-number matrix F, a number n of clones, and a maximum copy number $c_{\max}$, find an $m \times n$ copy-number matrix C generating T^* and an $n \times k$ usage matrix U such that $F = CU$ and $\Lambda(T^*)$ are minimum.*

The goal of the problem is thus to deconvolve the observed fractional copy numbers F as mixtures of the copy-number profiles C of the leaves of the most parsimonious tree T^*, according to mixing proportions U.

3. METHODS

The hardness of CNTMD is an open question. However, we suspect the problem to be NP hard, as the related unmixed version, the CNT problem, is NP hard (El-Kebir et al., 2016a, 2017). Moreover, other similar deconvolution problems under a tree constraint are NP hard as well (El-Kebir et al., 2015, 2016b). As such, we design a heuristic algorithm based on the coordinate-descent paradigm for solving a distance-based version of CNTMD where we aim to infer copy numbers C with n clones (columns) and mixing proportions U that minimize the distance between the *observed* fractional copy numbers F and the *inferred* fractional copy numbers CU:

$$\|F-CU\| = \sum_{1\le s\le m}\sum_{1\le p\le k}\left|f_{s,p}-\sum_{1\le i\le n} c_{s,i}u_{i,p}\right|. \tag{2}$$

Under a parsimony constraint, we impose a maximum cost $\Lambda_{\max}$ on the CNT T generated by C. That is, we require that C generates T such that $\Lambda(T)\le\Lambda_{\max}$ and we consider the following problem.

Problem 2 (Distance-based Copy-Number Tree Mixture Deconvolution (d-CNTMD)) *Given an $m\times k$ fractional copy-number matrix F, a number n of clones, a maximum copy number $c_{\max}$, and a maximum cost $\Lambda_{\max}$, find an $m\times n$ copy-number matrix $C=[c_{s,i}]$ generating T and an $n\times k$ usage matrix U such that $c_{s,i}\le c_{\max}$, $\Lambda(T)\le\Lambda_{\max}$, and $\|F-CU\|$ are minimum.*

Following the coordinate-descent paradigm, we split the variables of d-CNTMD and obtain two subproblems, where either matrix C or matrix U is fixed, with the same objective of minimizing the distance $\|F-CU\|$. An iteration t consists of two steps. In the *C-step*, we are given a usage matrix U_{t-1} and we search for a copy-number matrix $C_t=[c_{s,i}]$ minimizing $\|F-C_tU_{t-1}\|$ such that $c_{s,i}\le c_{\max}$ and C generate T with cost $\Lambda(T)\le\Lambda_{\max}$. Conversely, in the *U-step* we take the matrix C_t as input and seek a usage matrix U_t such that $\|F-C_tU_t\|$ is minimized.

To account for local optima, we use Q restarts with different initial usage matrices $U_{0,0},\ldots,U_{Q,0}$. We generate these usage matrices in a sparse way using a method based on random-number partitions (Nijenhuis and Wilf, 2014). This procedure yields a sequence of pairs of matrices, where for consecutive pairs $(C_{q,t},U_{q,t}),(C_{q,t+1},U_{q,t+1})$ it holds that $\|F-C_{q,t}U_{q,t}\|\ge\|F-C_{q,t+1}U_{q,t+1}\|$. This is because both $C_{q,t+1}$ and $U_{q,t+1}$ can be chosen equal to the previous matrices $C_{q,t}$ and $U_{q,t}$, respectively, resulting in the same distance. We iterate until $\|F-C_{q,t}U_{q,t}\|$ drops below a convergence threshold or the number of iterations reaches a specified number K.

Our algorithm thus computes Q pairs $(C_{q,K},U_{q,K})$ of matrices for each restart $U_{q,0}$ and returns a pair (C',U^*) of matrices that minimize the distance $\|F-C_{q,K}U_{q,K}\|$. In the distance-based formulation, we do not directly optimize for the cost $\Lambda(T')$ of a tree T' generated by C'. Instead, we only require that each identified matrix $C_{q,K}$ generates a CNT $T_{q,K}$ with cost $\Lambda(T_{q,K})\le\Lambda_{\max}$ and, consequently, we have that the final matrix C' generates a CNT T' with cost $\Lambda(T')\le\Lambda_{\max}$. Thus, it may be the case that for the same usage matrix U^*, there exists another copy-number matrix C'' different from C' that generates a CNT T'' whose cost is $\Lambda(T'') < \Lambda(T')$ while having the same distance $\|F-C'U^*\| = \|F-C''U^*\|$. To find the best such matrix C^* that generates a tree T^* with the smallest cost $\Lambda(T^*)$, we introduce a *refinement step* with a slightly adjusted ILP formulation of the C-step. Figure 2 depicts the entire procedure of the coordinate-descent algorithm.

We present a linear programming (LP) formulation for the U-step in Section 3.1 followed by an ILP formulation for the C-step in Section 3.2. Since the distance-based variant of the problem does not directly minimize the cost of the tree, we present in Section 3.3 an algorithm for finding the smallest maximum cost Λ^* that achieves the largest decrease in the distance $\|F-CU\|$.

3.1. U-step

In the U-step, we are given a fractional matrix F and a copy-number matrix C, and seek a usage matrix $U=[u_{i,p}]$ with real-valued entries $u_{i,p}$ minimizing the distance $\|F-CU\|$. We linearize the distance function $\|F-CU\|$ and formulate the resulting optimization problem as an LP with $O(km)$ variables and $O(km)$ constraints. To model the absolute difference in Equation (2), we introduce variables $\bar{f}_{s,p}$ for each segment s and sample p, and model $\bar{f}_{s,p}=|f_{s,p}-\sum_{1\le i\le n}c_{s,i}u_{i,p}|$ using the following linear constraints.

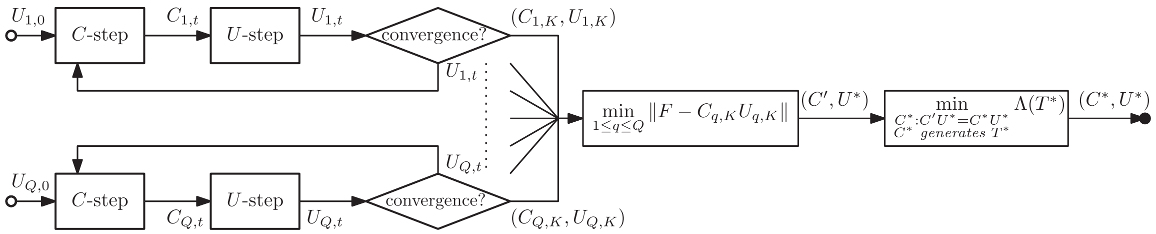

FIG. 2. Scheme of our coordinate-descent algorithm for solving the CNTMD problem. Given an initial usage matrix $U_{q,0}$, the algorithm performs K alternating iterations of two distinct steps: the C-step and the U-step. For each iteration t, the C-step computes a copy-number matrix $C_{q,t}$ given the previous usage matrix $U_{q,t-1}$. Then, the U-step computes a usage matrix $U_{q,t}$ given $C_{q,t}$. We repeat the entire procedure using Q restarts with random usage matrices $U_{1,0}, \ldots, U_{Q,0}$. We show the first repetition starting with $U_{1,0}$ and the last repetition starting with $U_{Q,0}$, while $(C_{0,K}, U_{0,K})$ and $(C_{Q,K}, U_{Q,K})$ correspond to the pairs of matrices returned by these steps, respectively. The pair (C', U^*) corresponds to the matrices $(C_{q,K}, U_{q,K})$ from the repetition q that achieve the minimum distance $\|F - C_{q,K}U_{q,K}\|$. Finally, the refinement step searches for a copy-number matrix C^* that generates a CNT T^* with minimum cost $\Lambda(T^*)$ such that $\|F - C'U^*\| = \|F - C^*U^*\|$.

$$\bar{f}_{s,p} \geq f_{s,p} - \sum_{1\leq i\leq n} c_{s,i}u_{i,p} \qquad 1 \leq s \leq m, 1 \leq p \leq k \tag{3}$$

$$\bar{f}_{s,p} \geq \sum_{1\leq i\leq n} c_{s,i}u_{i,p} - f_{s,p} \qquad 1 \leq s \leq m, 1 \leq p \leq k \tag{4}$$

Moreover, we introduce variables $0 \leq u_{i,p} \leq 1$ that represent the usage of a clone i in sample p. We constrain the usage of each sample to sum to 1 using the following constraint.

$$\sum_{1\leq i\leq n} u_{i,p} = 1 \qquad 1 \leq p \leq k \tag{5}$$

In sum, the following LP solves the U-step:

$$\min_{\mathrm{u}, \bar{\mathrm{f}}} \sum_{\substack{1\leq s\leq m\\ 1\leq p\leq k}} \bar{f}_{s,p} \tag{6}$$

$$\text{s.t. (3), (4) and (5)}$$

3.1.1. Complete LP formulation. Below is the complete LP formulation used in the U-step.

$$\begin{aligned} &\min \sum_{1\leq p\leq k} \sum_{1\leq s\leq m} \bar{f}_{s,p} \\ &\bar{f}_{s,p} \geq f_{s,p} - \sum_{1\leq i\leq n} c_{s,i}u_{i,p} && 1 \leq s \leq m, 1 \leq p \leq k \\ &\bar{f}_{s,p} \geq \sum_{1\leq i\leq n} c_{s,i}u_{i,p} - f_{s,p} && 1 \leq s \leq m, 1 \leq p \leq k \\ &\sum_{1\leq i\leq n} u_{i,p} = 1 && 1 \leq p \leq k \end{aligned}$$

3.2. C-*step*

In the C-step, we are given a fractional matrix F and a usage matrix U, and seek a copy-number matrix $C = [c_{s,i}]$ with integer entries $c_{s,i}$ minimizing the distance $\|F - CU\|$ such that $c_{s,i} \leq c_{\max}$ and C generate a tree T with $\Lambda(T) \leq \Lambda_{\max}$. Similarly, to the U-step, we model the distance function $\|F - CU\|$ with variables $\bar{f}_{s,p}$ and their corresponding constraints (3) and (4). We formulate the optimization problem of the C-step as an ILP with $O(n^2m + nm \log \Lambda_{\max} + km)$ variables and constraints. Our formulation introduces new constraints that improve on the model introduced in El-Kebir et al. (2016a, 2017).

We introduce binary variables $X = [x_{i,j}]$ to model the topology of T and integer variables $\tilde{C}$ to label the vertices and edges of T. Note that C, which is the collection of profiles for the leaves $L(T)$, is a submatrix of $\tilde{C}$.

3.2.1. Topology of T. Recall that T is a full binary tree (Definition 1). We construct a directed acyclic graph $G=(V,E)$ that contains all CNTs T with n leaves as spanning trees. More specifically, we order the vertices $V=\{v_1, \ldots, v_{2n-1}\}$ such that $L(T)=\{v_1, \ldots, v_n\}$ and $r(T)=v_{2n-1}$. The edge set E contains edges $\{(v_i, v_j)|n+1 \leq i < 2n-1, 1 \leq j < i \leq 2n-1\}$. We introduce a variable $x_{i,j}$ for each edge $(v_i, v_j) \in E$, which indicates whether (v_i, v_j) is an edge of T. To encode that T is a full binary spanning tree of G, we require that each non-root vertex has exactly one incoming edge and that each internal vertex has two outgoing edges with the following constraints.

$$\sum_{\substack{i \geq j \\ i \geq n+1}} x_{i,j} = 1 \qquad 1 \leq j < 2n-1 \tag{7}$$

$$\sum_{1 \leq j < i} x_{i,j} = 2 \qquad n < i \leq 2n-1 \tag{8}$$

3.2.2. Vertex and edge labeling. Integer variables $\tilde{C}=[c_{s,i}]$ where $c_{s,i} \in \{0, \ldots, c_{\max}\}$ encode the profiles of each vertex v_i. More precisely, $c_{s,i}$ encodes the copy-number state of segment s of vertex v_i. Since the profile of the root vertex is diploid, we add the following constraints.

$$c_{s,2n-1} = 2 \qquad 1 \leq s \leq m \tag{9}$$

From these profiles and the topology of T (as captured by variables $x_{i,j}$), we obtain the events $\mathcal{E}(i,j)$ that transform the profile $\mathbf{c}_i$ into the profile $\mathbf{c}_j$ and thereby the cost for the edge (v_i, v_j). Recall that an event is a triple (s,t,b) and corresponds to an amplification if $b > 0$ and a deletion otherwise. We model the amplifications and deletions covering any segment s in $\mathcal{E}(i,j)$ with two separate variables $a_{s,i,j} \in \{0, \ldots, c_{\max}\}$ and $d_{s,i,j} \in \{0, \ldots, c_{\max}\}$, respectively. Note that we require $\mathcal{E}(i,j)$ to be empty when the corresponding edge (v_i, v_j) is not in T. As such, we introduce the following constraints that force variables $a_{s,i,j}$ and $d_{s,i,j}$ to be 0 when (v_i, v_j) is not in T.

$$a_{s,i,j} \leq c_{\max} x_{i,j} \qquad 1 \leq s \leq m, (v_i, v_j) \in E(G) \tag{10}$$

$$d_{s,i,j} \leq c_{\max} x_{i,j} \qquad 1 \leq s \leq m, (v_i, v_j) \in E(G) \tag{11}$$

Due to these constraints, the cost of every pair (v_i, v_j) of vertices that do not form an edge of T, that is, $x_{i,j}=0$, is fixed to 0. Therefore, only the cost of the edges of T is computed, which significantly constraints the model and thereby improves the performance over the formulation presented for the unmixed CNT problem (El-Kebir et al., 2016a, 2017).

Now, we consider the effect of amplifications and deletions on a segment s. As described above, we assume that deletions are applied before amplifications. Moreover, if a subset of deletions results in segment s reaching value 0, the remaining amplifications and deletions will not change the value of that segment. We distinguish the following four different cases.

(a) $c_{s,i}=0$ and $c_{s,j}=0$: Since both segments have value 0, we have that, following a result in Shamir et al. (2016), the number of amplifications $a_{s,i,j}$ and deletions $d_{s,i,j}$ must be between 0 and $c_{\max}$.
(b) $c_{s,i} \neq 0$ and $c_{s,j} \neq 0$: Since $c_{s,j} > 0$, the number of deletions $d_{s,i,j}$ must be strictly smaller than $c_{s,i}$. Moreover, it must hold that $c_{s,j}+d_{s,i,j}=c_{s,i}+a_{s,i,j}$.
(c) $c_{s,i} \neq 0$ and $c_{s,j}=0$: Since deletions precede amplifications, the number of deletions $d_{s,i,j}$ must be at least $c_{s,i}$.
(d) $c_{s,i}=0$ and $c_{s,j} \neq 0$: Once a segment s has been lost it cannot be regained. As such, this case is infeasible.

These cases are also summarized in Table 1.

To capture the conditions of the four cases, we introduce binary variables $z_{i,s,q}$ that provide a binary representation of the integer variable $c_{s,i}$. In addition, we introduce binary variables $\bar{c}_{s,i} \in \{0, 1\}$ and the following constraints such that $\bar{c}_{s,i}=1$ iff $c_{s,i} \neq 0$.

$$c_{s,i} = \sum_{q=0}^{\lfloor \log_2(c_{\max}) \rfloor + 1} 2^q \cdot z_{i,s,q} \qquad 1 \leq i \leq 2n-1, 1 \leq s \leq m \tag{12}$$

Table 1. Case Analysis on the Values of Variables $c_{s,i}$ and $c_{s,j}$ for a Segment s When $a_{s,i,j}$ and $d_{s,i,j}$ Are the Amplifications and the Deletions, Respectively, Covering s and Labeling the Edge (v_i, v_j) of a Corresponding Copy-Number Tree T

	$a_{s,i,j}$	$d_{s,i,j}$	*Additional*
(a) $c_{s,i}=0 \wedge c_{s,j}=0$	$\leq c_{\max}$	$\leq c_{\max}$	
(b) $c_{s,i} \neq 0 \wedge c_{s,j} \neq 0$	$\leq c_{\max}$	$< c_{s,i}$	$c_{s,j}+d_{i,j,s}=c_{s,i}+a_{s,i,j}$
(c) $c_{s,i} \neq 0 \wedge c_{s,j}=0$	$\leq c_{\max}$	$\leq c_{\max},\ \geq c_{s,i}$	
(d) $c_{s,i}=0 \wedge c_{s,j} \neq 0$	*infeasible*	*infeasible*	*infeasible*

Note that in our model, deletions are applied before amplifications and when segment s reaches value 0, the remaining amplifications and deletions will not change the copy number of s.

$$\bar{c}_{s,i} \leq \sum_{q=0}^{\lfloor \log_2(c_{\max})\rfloor+1} z_{i,s,q} \qquad 1 \leq i \leq 2n-1, 1 \leq s \leq m \tag{13}$$

$$\bar{c}_{s,i} \geq z_{i,s,q} \qquad 1 \leq i \leq 2n-1, 1 \leq s \leq m, 0 \leq q \leq \lfloor \log_2(c_{\max})\rfloor+1 \tag{14}$$

$$z_{i,s,q} \in \{0, 1\} \qquad 1 \leq i \leq 2n-1, 1 \leq s \leq m, 0 \leq q \leq \lfloor \log_2(c_{\max})\rfloor+1 \tag{15}$$

Since $a_{s,i,j}, d_{s,i,j} \in \{0, \ldots, c_{\max}\}$, the upper bound constraints involving $c_{\max}$ are covered. In particular, case (a) is captured in its entirety. We capture case (b) with the following constraints.

$$c_{s,j} \leq c_{s,i} - d_{s,i,j} + a_{s,i,j} + 2c_{\max}(3 - \bar{c}_{i,s} - \bar{c}_{j,s} - x_{i,j}) \qquad 1 \leq s \leq m, (v_i, v_j) \in E(G) \tag{16}$$

$$c_{s,j} + 2c_{\max}(3 - \bar{c}_{s,i} - \bar{c}_{s,j} - x_{i,j}) \geq c_{s,i} - d_{s,i,j} + a_{s,i,j} \qquad 1 \leq s \leq m, (v_i, v_j) \in E(G) \tag{17}$$

$$d_{i,j,s} \leq c_{s,i} - 1 + (c_{\max}+1)(2 - \bar{c}_{s,i} - \bar{c}_{s,j}) \qquad 1 \leq s \leq m, (v_i, v_j) \in E(G) \tag{18}$$

In fact, in the case of $x_{i,j}=1$ (i.e., (v_i, v_j) is in T), $\bar{c}_{s,i}=1$, and $\bar{c}_{s,j}=1$, constraints (16) and (17) model the equation $c_{s,j}+d_{s,i,j}=c_{s,i}+a_{s,i,j}$, whereas constraint (18) ensures that $d_{s,i,j} < c_{s,i}$. Otherwise, in the case of $x_{i,j}=0$, the constraints are always satisfied and the corresponding variables $a_{s,i,j}, d_{s,i,j}$ for every segment s are forced to 0 (which is different from the ILP formulation in El-Kebir et al. (2016a, 2017). Note that $d_{s,i,j}$ can be always equal to zero by constraint (18), and hence, we do not need to distinguish whether $x_{i,j}=0$ or $x_{i,j}=1$. Next, we model case (c), when $x_{i,j}=1$, using the following constraints.

$$c_{s,i} \leq d_{s,i,j} + c_{\max}(2 - \bar{c}_{s,i} + \bar{c}_{s,j} - x_{i,j}) \qquad 1 \leq s \leq m, (v_i, v_j) \in E(G) \tag{19}$$

Finally, the following constraints, which encode that if $x_{i,j}=1$ and then $\bar{c}_{s,i}=0$ implies $\bar{c}_{s,j}=0$, prevent case (d) from happening.

$$1 - x_{i,j} + \bar{c}_{s,i} \geq \bar{c}_{s,j} \qquad 1 \leq s \leq m, (v_i, v_j) \in E(G) \tag{20}$$

We model the cost of an edge (v_i, v_j) as the sum of the amplifications and deletions starting at each segment s by introducing variables $\bar{a}_{s,i,j} \in \{0, \ldots, c_{\max}\}$ and $\bar{d}_{s,i,j} \in \{0, \ldots, c_{\max}\}$. Variables $\bar{a}_{s,i,j}$ correspond to the amplifications starting at segment s and is equal to $\max\{a_{s,i,j} - a_{s-1,i,j}, 0\}$. Symmetrically, variables $\bar{d}_{s,i,j}$ correspond to the deletions starting at segment s and are equal to $\max\{d_{s,i,j} - d_{s-1,i,j}, 0\}$. We model this using the following constraints.

$$\bar{a}_{s,i,j} \geq a_{s,i,j} - a_{s-1,i,j} \qquad 1 \leq s \leq m, (v_i, v_j) \in E(G) \tag{21}$$

$$\bar{d}_{s,i,j} \geq d_{s,i,j} - d_{s-1,i,j} \qquad 1 \leq s \leq m, (v_i, v_j) \in E(G) \tag{22}$$

$$a_{0,i,j} = 0 \qquad (v_i, v_j) \in E(G) \tag{23}$$

$$d_{0,i,j} = 0 \qquad (v_i, v_j) \in E(G) \tag{24}$$

As before, we force $\bar{a}_{s,i,j}$ and $\bar{d}_{s,i,j}$ to 0 when the corresponding pair (v_i, v_j) of vertices is not an edge of T using the following constraints.

$$\bar{a}_{s,i,j} \leq c_{\max} x_{i,j} \qquad 1 \leq s \leq m, (v_i, v_j) \in E(G) \tag{25}$$

$$\bar{d}_{s,i,j} \leq c_{\max} x_{i,j} \qquad 1 \leq s \leq m, (v_i, v_j) \in E(G) \tag{26}$$

Now, the cost of an edge (v_i, v_j) can indeed be expressed as $\sum_{1\leq s\leq m}(\bar{a}_{s,i,j}+\bar{d}_{s,i,j})$. Hence, the cost $\Lambda(T)$ is simply the sum of the costs of all the edges, and we require that this cost is at most $\Lambda_{\max}$ with the following constraint.

$$\sum_{(v_i,v_j)\in E(G)}\sum_{1\leq s\leq m}(\bar{a}_{s,i,j}+\bar{d}_{s,i,j})\leq\Lambda_{\max} \tag{27}$$

Thus, the following ILP solves the C-step:

$$\min_{c,\bar{f}}\sum_{\substack{1\leq s\leq m\\1\leq p\leq k}}\bar{f}_{s,p} \tag{28}$$

$$\text{s.t.}\quad \begin{array}{l}(3),(4),(7),(8),(9),(10),(11),\ (12),(13),(14),(15),(16),\\(17),(18),(19),(20),(21),(22),(23),(24),(25),(26)\text{ and }(27).\end{array}$$

3.2.3. *Complete ILP formulation.*

Below is the complete ILP formulation of the C-step.

$$\min\sum_{1\leq s\leq m,\,1\leq p\leq k}\bar{f}_{s,p}$$

$$\bar{f}_{s,p}\geq f_{s,p}-\sum_{1\leq i\leq n}c_{s,i}u_{i,p}\qquad 1\leq s\leq m,1\leq p\leq k$$

$$\bar{f}_{s,p}\geq\sum_{1\leq i\leq n}c_{s,i}u_{i,p}-f_{s,p}\qquad 1\leq s\leq m,1\leq p\leq k$$

$$\sum_{(v_i,v_j)\in E(G)}\sum_{1\leq s\leq m}\bar{a}_{s,i,j}+\bar{d}_{s,i,j}\leq\Lambda_{\max}$$

$$\sum_{j\leq i\wedge i\geq n+1}x_{i,j}=1\qquad 1\leq j<2n-1$$

$$\sum_{1\leq j<i}x_{i,j}=2\qquad n<i\leq 2n-1$$

$$c_{s,2n-1}=2\qquad 1\leq s\leq m$$

$$a_{s,i,j}\leq c_{\max}x_{i,j}\qquad 1\leq s\leq m,(v_i,v_j)\in E(G)$$

$$d_{s,i,j}\leq c_{\max}x_{i,j}\qquad 1\leq s\leq m,(v_i,v_j)\in E(G)$$

$$c_{s,i}=\sum_{q=0}^{\lfloor\log_2(c_{\max})\rfloor+1}2^q\cdot z_{i,s,q}\qquad 1\leq i\leq 2n-1,1\leq s\leq m$$

$$\bar{c}_{s,i}\leq\sum_{q=0}^{\lfloor\log_2(c_{\max})\rfloor+1}z_{i,s,q}\qquad 1\leq i\leq 2n-1,1\leq s\leq m$$

$$\bar{c}_{s,i}\geq z_{i,s,q}\qquad 1\leq i\leq 2n-1,1\leq s\leq m,0\leq q\leq\lfloor\log_2(c_{\max})\rfloor+1$$

$$c_{s,j}\leq c_{s,i}-d_{s,i,j}+a_{s,i,j}+2c_{\max}(3-\bar{c}_{i,s}-\bar{c}_{j,s}-x_{i,j})\qquad 1\leq s\leq m,(v_i,v_j)\in E(G)$$

$$c_{s,j}+2c_{\max}(3-\bar{c}_{s,i}-\bar{c}_{s,j}-x_{i,j})\geq c_{s,i}-d_{s,i,j}+a_{s,i,j}\qquad 1\leq s\leq m,(v_i,v_j)\in E(G)$$

$$d_{i,j,s}\leq c_{s,i}-1+(c_{\max}+1)(2-\bar{c}_{s,i}-\bar{c}_{s,j})\qquad 1\leq s\leq m,(v_i,v_j)\in E(G)$$

$$c_{s,i}\leq d_{s,i,j}+c_{\max}(2-\bar{c}_{s,i}+\bar{c}_{s,j}-x_{i,j})\qquad 1\leq s\leq m,(v_i,v_j)\in E(G)$$

$$1-x_{i,j}+\bar{c}_{s,i}\geq\bar{c}_{s,j}\qquad 1\leq s\leq m,(v_i,v_j)\in E(G)$$

$$\bar{a}_{s,i,j}\geq a_{s,i,j}-a_{s-1,i,j}\qquad 1\leq s\leq m,(v_i,v_j)\in E(G)$$

$$\bar{d}_{s,i,j}\geq d_{s,i,j}-d_{s-1,i,j}\qquad 1\leq s\leq m,(v_i,v_j)\in E(G)$$

$$a_{0,i,j}=0\qquad (v_i,v_j)\in E(G)$$

$$
\begin{array}{lr}
d_{0,i,j}=0 & (v_i,v_j)\in E(G)\\
\bar{a}_{s,i,j}\le c_{\max}x_{i,j} & 1\le s\le m,(v_i,v_j)\in E(G)\\
\bar{d}_{s,i,j}\le c_{\max}x_{i,j} & 1\le s\le m,(v_i,v_j)\in E(G)\\
x_{i,j}\in\{0,1\} & (v_i,v_j)\in E(G)\\
c_{s,i}\in\{0,\ldots,c_{\max}\} & 1\le i\le 2n-1,1\le s\le m\\
\bar{c}_{s,i}\in\{0,1\} & 1\le i\le 2n-1,1\le s\le m\\
z_{i,s,q}\in\{0,1\} & 1\le i\le 2n-1,1\le s\le m,0\le q\le\lfloor\log_2(c_{\max})\rfloor+1\\
a_{s,i,j},d_{s,i,j}\in\{0,\ldots,c_{\max}\} & 1\le s\le m,(v_i,v_j)\in E(G)\\
\bar{a}_{s,i,j},\bar{d}_{s,i,j}\in\{0,\ldots,c_{\max}\} & 1\le s\le m,(v_i,v_j)\in E(G)\\
\bar{f}_{s,p}\in[0,\ldots,c_{\max}] & 1\le s\le m,1\le p\le k
\end{array}
$$

3.3. Choosing $\Lambda_{\max}$ to balance cost $\Lambda(T)$ and distance $\|F-CU\|$

We indicate by $(C^{\Lambda},U^{\Lambda})$ the matrices found by our approach with maximum cost $\Lambda_{\max}=\Lambda$ and we define $d(\Lambda)=\|F-C^{\Lambda}U^{\Lambda}\|$. First, observe that the objective function $d(\Lambda_t)$ is nonincreasing with larger values of Λ_t. That is, if $\Lambda_t\ge\Lambda$ then $d(\Lambda_t)\le d(\Lambda)$, as C^{Λ} generates T with cost $\Lambda(T)<\Lambda_t$. The parameter $\Lambda_{\max}$ controls the trade-off between the cost $\Lambda(T)$ of the tree T and the distance $\|F-CU\|$. In the following, we describe an algorithm for finding the smallest maximum cost Λ^* such that $d(\Lambda^*)=0$.

However, requiring that $d(\Lambda^*)=0$ is too stringent as the value $d(\Lambda_t)$ depends on the number of restarts and is further confounded by the presence of noise that may result from mapping errors or amplification biases (such as GC-content bias). It is thus reasonable to expect that $d(\Lambda^*)>0$ and that small decreases in the value of $d(\Lambda_t)$ for any $\Lambda_t>\Lambda^*$ may be not significant due to these confounding factors. We therefore introduce the parameter ε and say that $\Lambda_2>\Lambda_1$ provides a better solution than Λ_1 if and only if $d(\Lambda_1)-d(\Lambda_2)>\varepsilon$. Intuitively, the user-specified threshold ε controls the trade-off between greater robustness to noise (larger ε) or more precision (smaller ε). We redefine Λ^* as the smallest integer whose solution cannot be improved by increasing the maximum cost, that is, $d(\Lambda^*)-d(\Lambda_t)\le\varepsilon$ for any $\Lambda_t\ge\Lambda^*$. Note that in a similar manner, ε plays a role in the refinement step described previously.

We use the monotonicity of the function $d(\Lambda_t)$ and use binary search for finding the value Λ^*. Initially, we consider a large interval $[\Lambda_{\mathrm{L}_0},\Lambda_{\mathrm{R}_0}]$, which we assume to contain Λ^*. We then recursively consider intervals $[\Lambda_{\mathrm{L}_t},\Lambda_{\mathrm{R}_t}]$ to search for Λ^*. At each recursive step, we evaluate $d(\Lambda_{\mathrm{M}_t})$, where Λ_{M_t} is the integer at the center of the interval $[\Lambda_{\mathrm{L}_t},\Lambda_{\mathrm{R}_t}]$. If Λ_{R_0} provides a better solution than Λ_{M_t}, that is, $d(\Lambda_{\mathrm{M}_t})-d(\Lambda_{\mathrm{R}_0})>\varepsilon$, then we know that Λ^* is contained in $[\Lambda_{\mathrm{M}_t},\Lambda_{\mathrm{R}_t}]$ and recurse on this interval. Otherwise, we know that Λ^* is contained in $[\Lambda_{\mathrm{L}_t},\Lambda_{\mathrm{M}_t}]$ due to the monotonicity of d. We terminate when $\Lambda_{\mathrm{L}_t}=\Lambda_{\mathrm{R}_t}$. The algorithm thus performs a logarithmic number of iterations.

4. RESULTS

We applied our algorithm for CNTMD to simulated data and to data from two patients from a prostate cancer data set (Gundem et al., 2015). We ran every experiment in this section on a compute cluster, and every execution lasted up to 2 days, with 160 restarts for the simulated data and 300 restarts for the real data. The implementation of our method and related details, as well as the simulated and processed data, and the implementation of alternative methods are available at https://github.com/raphael-group/CNT-MD.

4.1. Benchmark on simulated data

4.1.1. Alternative methods. We benchmarked CNTMD on simulated data, comparing its performance to several other approaches, which we now describe and whose features we summarize in Table 2. The first alternative approach is a ''factorization-only'' approach that aims to factorize a fractional copy-number matrix F into a copy-number matrix C and a usage matrix U such that $F=CU$ without imposing a tree constraint. Published approaches to this problem perform this factorization (sometimes called deconvolution) independently on each sample (Nik-Zainal et al., 2012; Oesper et al., 2013; Fischer et al.,

Table 2. Comparison of Features for Alternative Methods Inferring the Integer Copy Numbers of Distinct Clones and Their Evolution from Multiple Heterogeneous Samples

Method	*Tree*	*Interval events*	*Factorization*	*Ref.*
Integer matrix factorization	No	No	Yes	This article
Copy-number tree	Yes	Yes	No	El-Kebir et al. (2017, 2016a), Schwarz et al. (2014)
Soft CNT	Yes	Yes	No	El-Kebir et al. (2017, 2016a)
Copy-Number Tree Mixture Deconvolution	Yes	Yes	Yes	This article
Single CNTMD	Yes	No	Yes	This article

For each method, the supported features are listed, including the construction of a phylogenetic tree, the presence of interval events to model the effects of copy-number aberrations on contiguous segments, and the factorization of fractional copy numbers into integer copy numbers and mixing proportions.

CNT, copy-number tree; CNTMD, Copy-Number Tree Mixture Deconvolution.

2014; McPherson et al., 2016)—one exception is Roman et al. (2016), but this infers noninteger copy numbers and it has not been applied to multiple samples from the same tumor. These methods do not take into account any information from the context and may provide unlikely profiles characterized by many interval events without a reasonable structure (Fig. 3B). To the best of our knowledge, there is no published method that solves the matrix factorization problem for the case where F comprises multiple vectors and C is composed of integers. Thus, we implemented *Integer Matrix Factorization (IMF)*, which performs the factorization by splitting the variables, C and U, and applying a coordinate-descent algorithm in a similar manner as the procedure described in Section 3.

Another class of approaches uses the same copy number model as CNTMD, but assumes that each sample is homogeneous and unmixed. One strategy is to first round the entries of F before inferring a CNT. We will do this by solving the CNT problem with an ILP model (El-Kebir et al., 2016a, 2017) mimicking the strategy that has been used by MEDICC (Schwarz et al., 2015; Sottoriva et al., 2015; Mangiola et al., 2016). We also consider a second rounding approach, which we call *soft CNT*, where we round the fractions in F either up or down such that we obtain a copy-number matrix C that generates T with minimum cost. We do this by extending the ILP formulation of the CNT problem described in El-Kebir et al. (2016a, 2017).

Finally, we also consider a variant of CNTMD, which we call *single CNTMD*. Here, we replace the interval events by single events; this is equivalent to a model where the cost of an interval event depends on the number of segments in the interval. However, the single-event model is not a good representation of true CNAs in cancer, as the length distribution of somatic CNAs is not simply a function of length (Zack et al., 2013). Such a copy number model was used by McPherson et al. (2016) and Chowdhury et al. (2014) for inferring the evolution comprising the minimum number of single events from the profile of clones inferred independently from each sample. Figure 3 shows an example highlighting the weaknesses of all the alternative methods presented above.

4.1.2. Simulated instances. We compare CNTMD with the methods described above on simulated instances composed of 22 chromosomes with a total of 350 segments. These instances have the same size as real data. The number of segments per chromosome ranges from 5 to 50 and follows the distribution of the number of segments in the prostate cancer data sets available in Gundem et al. (2015). Using a procedure similar to the one described in El-Kebir et al. (2016a) and El-Kebir et al. (2017), we randomly generate three CNTs, denoted by $\hat{T}$, which in turn were generated by copy number matrices $\hat{C}$ composed of four tumor clones plus the normal diploid clone. We mix the leaves of each tree according to a usage matrix $\hat{U}$ and obtain fractional copy-number matrices with $k \in \{2, 5, 10\}$ samples. For each tree and value for k, we generate three instances with different usage matrices. Thus, we consider 27 simulated instances in total.

We use three quality measures to compare the inferred tree T, inferred copy-number matrix C, and inferred usage matrix U to the simulated $\hat{T}$, $\hat{C}$, and $\hat{U}$. We compare T to $\hat{T}$ by considering the relative difference of events $|\Lambda(T) - \Lambda(\hat{T})|/\Lambda(\hat{T})$. To compare U to $\hat{U}$, we need to associate each inferred clone i to a corresponding true clone $\hat{i}$. Similarly to Malikic et al. (2015) and El-Kebir et al. (2015), we search for a maximum-weight bipartite matching that minimizes the value of the *usage difference* $\|U - \hat{U}\|$ in a bipartite graph where there is a an edge $(v_i, v_{\hat{i}})$ with weight $|\mathbf{c}_i - \mathbf{c}_{\hat{i}}|$ for all pairs $(i, \hat{i})$. To compare C to $\hat{C}$,

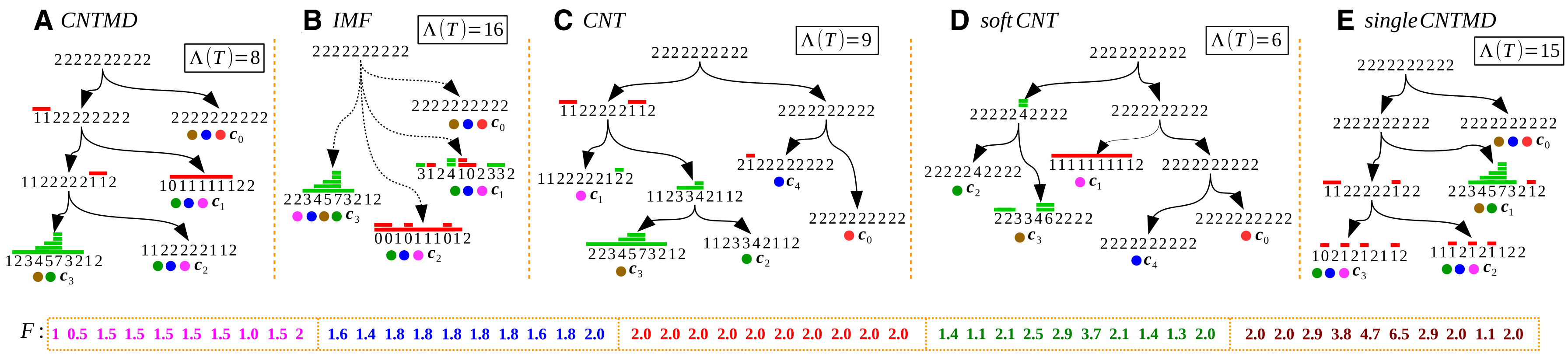

FIG. 3. Alternative methods infer trees that differ significantly from the true tree, which is inferred by our approach CNTMD. CNTs inferred by the methods in Table 2 where deletions $(s, t, -1)$ are red and amplifications $(s, t, 1)$ are green. **(A)** Shows the true tree composed of four clones $\mathbf{c}_0$ (normal), $\mathbf{c}_1$, $\mathbf{c}_2$, $\mathbf{c}_3$ with a cost of 8. This tree is correctly retrieved by CNTMD. All the alternative methods fail to infer the clonal mutation $(1, 2, -1)$. **(B)** The tree inferred by IMF contains too many events and differs significantly from the true tree. **(C, D)** CNT and soft CNT infer clones that are very different from the true clones that comprise the corresponding mixed samples such as $\mathbf{c}_1$ in **(C)** and $\mathbf{c}_2$ in **(D)**. **(E)** Single CNTMD splits the effect of the deletion $(1, 8, -1)$ across two distinct clones $\mathbf{c}_2$ and $\mathbf{c}_3$ resulting in a cost of 15. IMF, integer matrix factorization.

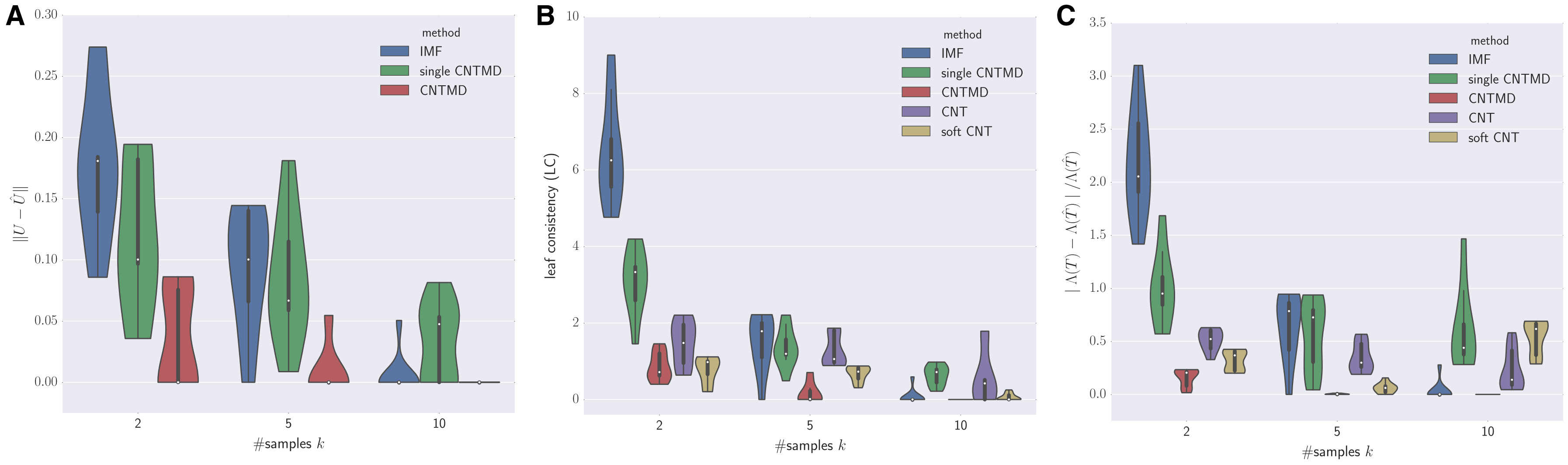

FIG. 4. CNTMD outperforms alternative methods on simulated data. Comparison of five methods (IMF, single CNTMD, CNTMD, CNT, and soft CNT) across 27 simulated data sets with $k \in \{2, 5, 10\}$ samples, consisting of four tumor clones and a normal diploid clone, each with a total of 350 segments across 22 chromosomes. **(A)** Normalized usage difference $\| U - \hat{U} \|$ between true and inferred mixing proportion. Methods CNT and soft CNT are not shown, as these methods do not compute U. **(B)** LC measure. **(C)** Difference $|\Lambda(T) - \Lambda(\hat{T})| / \Lambda(\hat{T})$ between the cost of the inferred tree T and the cost of the true tree $\hat{T}$. Each simulated instance was solved with the parameter n set to the true number of clones. LC, leaf consistency.

we compute a maximum-weight bipartite matching on the same complete bipartite graph where the edges are weighted by a similarity metric, called *leaf consistency (LC)*. This value is computed by solving an instance of CNT (El-Kebir et al., 2016a, 2017) for every pair $(\mathbf{c}_i, \mathbf{c}_j)$ of profiles where $\mathbf{c}_i$ is a column of C and $\mathbf{c}_j$ is a column of $\hat{C}$. More specifically, the LC value of $(\mathbf{c}_i, \mathbf{c}_j)$ is the minimum cost of a CNT with two leaves labeled by $\mathbf{c}_i, \mathbf{c}_j$ and with an unfixed root. Note that LC is 0 if and only if $\mathbf{c}_i, \mathbf{c}_j$ are equal. Similarly to the other metrics, we compute a maximum weight bipartite matching where the edges are weighted by the LC values for every pair $\mathbf{c}_i, \mathbf{c}_j$ of columns from C and $\hat{C}$, respectively. We normalize the weight of the obtained matching by the number of clones and chromosomes.

Figure 4 shows the results on the simulations. First, we observe that CNTMD, which combines both factorization and a proper interval tree-based model, outperforms all other methods across all number of samples. Second, we see that the quality metrics improve with increasing number k of samples for all the methods. This is especially the case for the factorization-based methods (IMF, single CNTMD, CNTMD), where differences in the clonal composition across samples provide a strong signal for deconvolution (Fig. 4A–C). In contrast, the rounding methods (CNT and soft CNT) show only a modest improvement with increasing number of samples (Fig. 4B, C), which is not surprising since rounding does not directly exploit differences in clonal composition across samples. Finally, observe that with a small number of samples ($k=2$), CNTMD dramatically outperforms IMF (Fig. 4A–C), demonstrating how CNTMD leverages the extra information given by the tree constraint. Moreover, by not accounting for interval events, single CNTMD results in CNTs that are inconsistent with the simulated trees and have many more events (Fig. 4C).

4.2. *Application to prostate cancer data set*

We apply our approach on prostate cancer patient A22 from the data set of Gundem et al. (2015). Patient A22 comprises 10 samples. We use the published fractional copy numbers that were obtained by the Battenberg algorithm (Nik-Zainal et al., 2012), which relies on the sample purity and tumor ploidy estimated by the ASCAT package (Van Loo et al., 2010).

Since the true clonal structure of these samples is unknown, we examine the consistency of different measures on the results obtained by running CNTMD with varying number of clones $n \in \{2, \ldots, 8\}$. We observe a number of patterns that suggest that there are six clones in the tumor that are distinguishable by CNAs; in comparison Gundem et al. (2015) estimate 16 clones using single-nucleotide variants (SNVs).

First, we observe that the value of $\| F - CU \|$ decreases significantly with increasing values of n (Fig. 5). However, the rate of decrease declines for $n > 6$, suggesting that additional clones are not providing

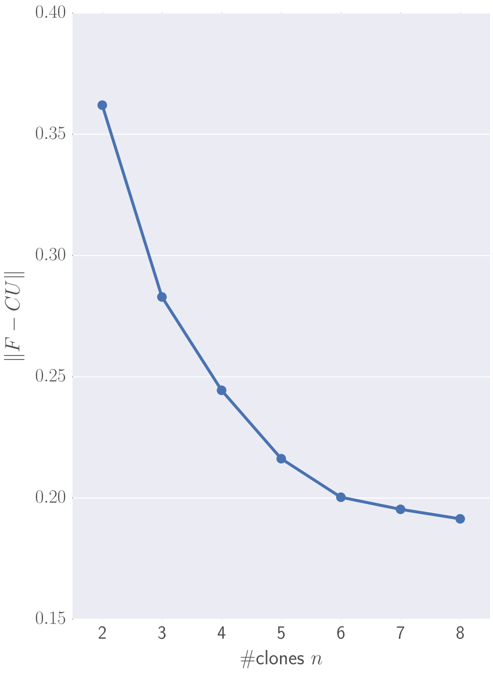

FIG. 5. The trade-off between the number n of clones and the distance $\| F - CU \|$ between the observed fractional copy numbers F and the inferred fractional copy numbers CU for patient A22 from Gundem et al. (2015). Note that the distance stabilizes with $n > 6$ clones and so we choose $n=6$ as the best fit in the data.

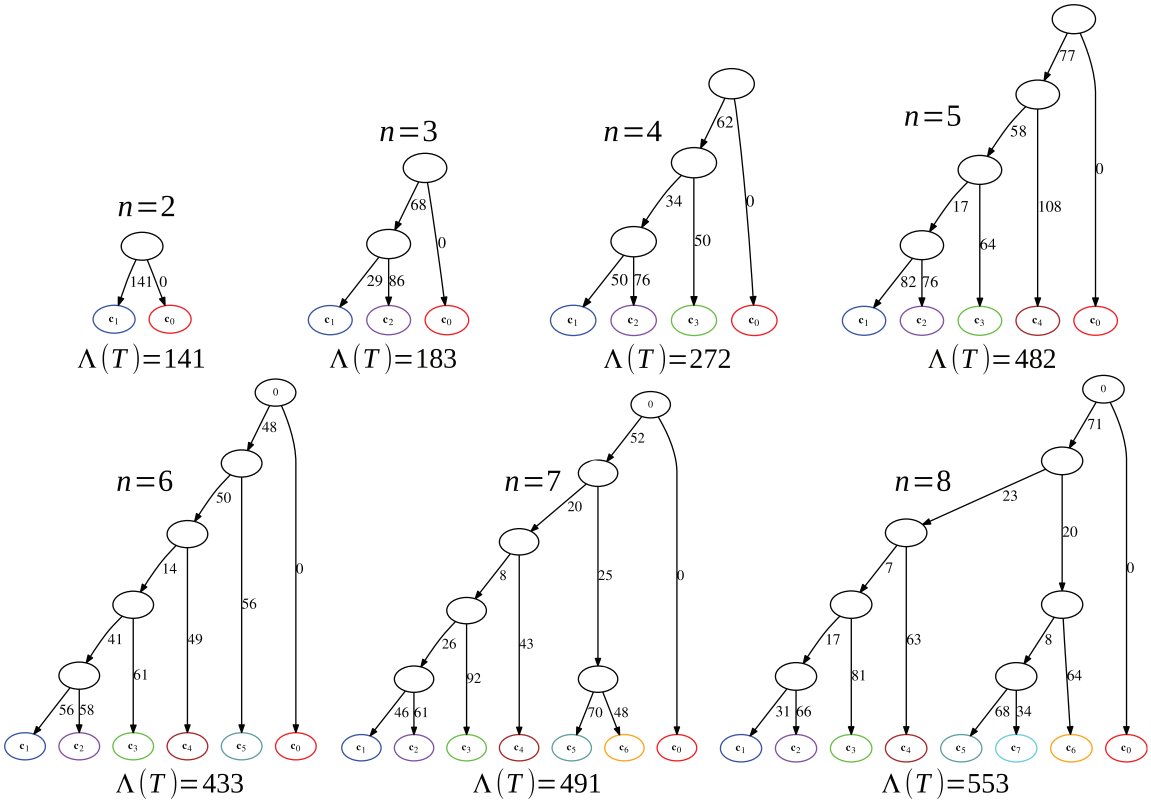

FIG. 6. Trees with $n \leq 6$ clones for patient A22 from Gundem et al. (2015) have a cascading topology and well-supported edges, whereas trees with $n > 6$ clones have the same cascading topology but have less-supported edges. For each CNT T inferred with $n \in \{2, \ldots, 8\}$ clones, we show the cost $\Lambda(T)$ and label the edges by their corresponding costs. The colors of leaves map the corresponding clones in the topologies. The red clone corresponds to the normal diploid clone.

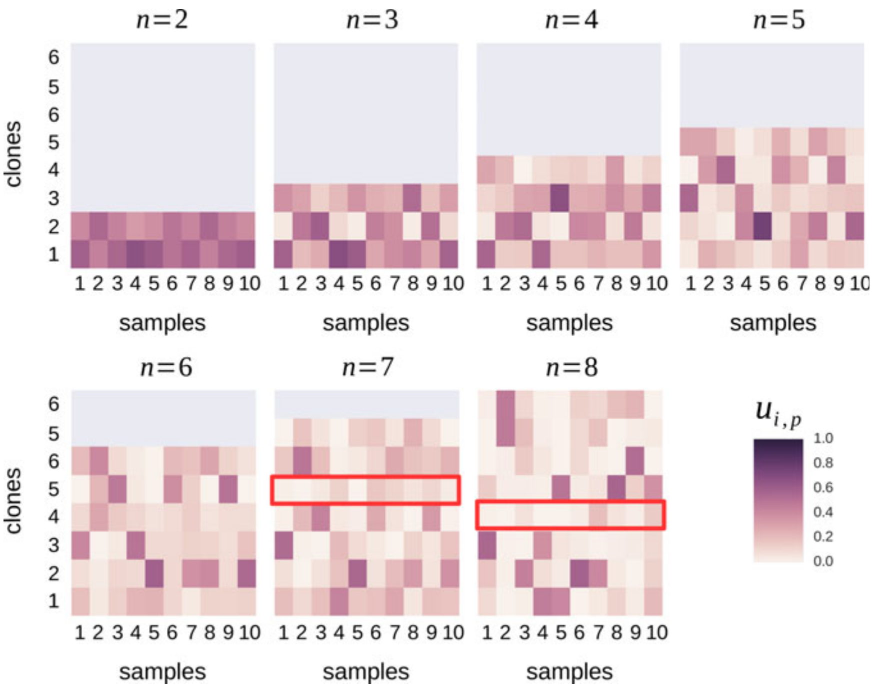

FIG. 7. Usage matrices $U = [u_{i,p}]$ with $n \leq 6$ clones for patient A22 from Gundem et al. (2015) are well supported. That is, with $n \leq 6$ clones, each inferred clone i (row) is present in at least one sample p(column) in high proportion, whereas with $n > 6$ clones some of the inferred clones are only present in low proportions in all samples (indicated by red boxes).

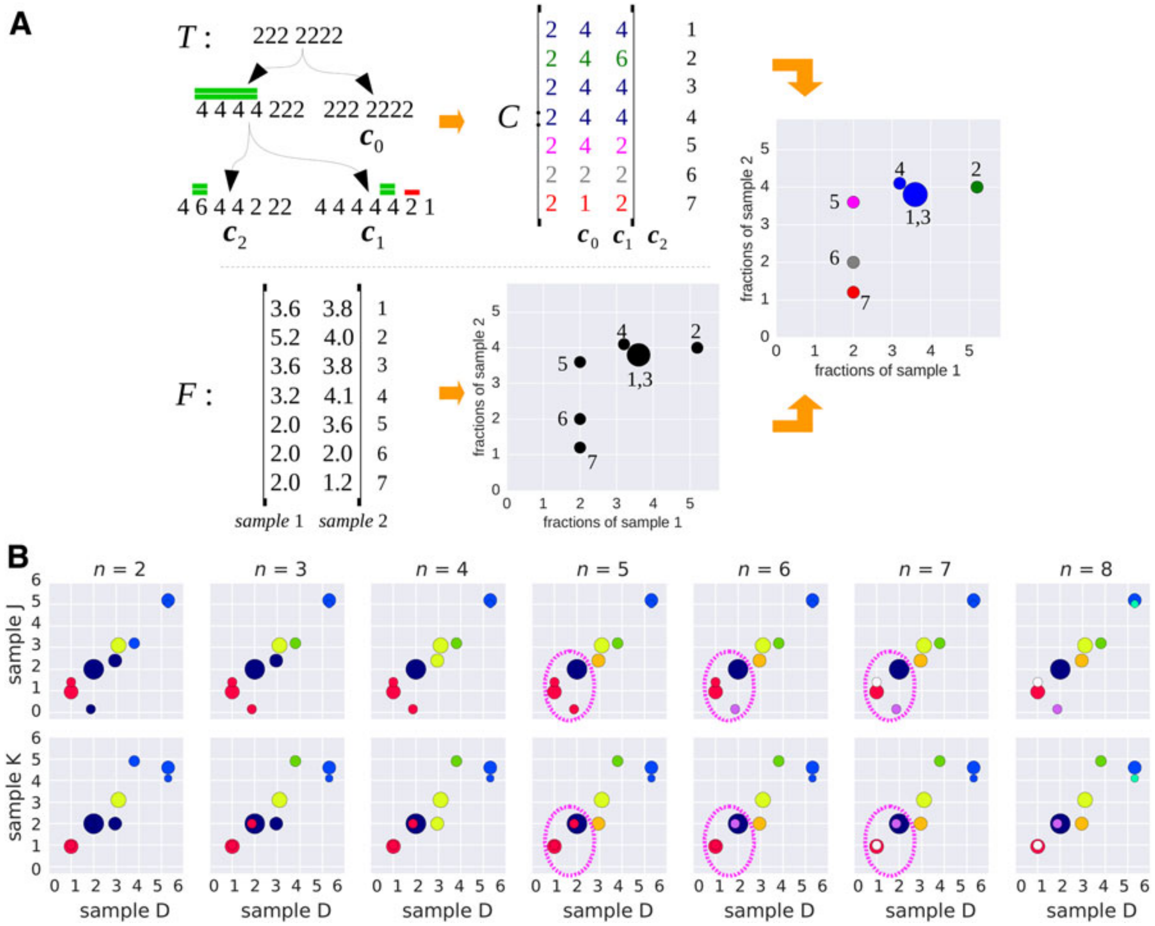

FIG. 8. Classes of segments with the same evolutionary history highlight consistency of the inferred solutions with the input data. (A) We use the inferred tree T to partition seven segments into classes. All segments that have the same copy-number change on each edge of T are in the same class (indicated by the color). Next, we consider a pair of samples and plot the fractions of the segments (size of the dot indicates the number of segments with the same class and same fractions). The class is *consistent* if all the segments of the same class are located close to each other. **(B)** Fractional copy numbers for three A22 prostate cancer samples from Gundem et al. (2015): D, K, and J. The largest dot contains 14 segments. Each column represents a value of $n \in \{2, \dots, 8\}$. The consistency of the classes improves with increasing n. The red class in $n=5$ is composed of two subsets of segments: segments that have one copy in all the considered samples, and segments that have two copies in samples D, K and zero copies in sample J. With $n=6$, these two subsets are separated into different classes (red and purple). With $n=7$, one more class (white) is introduced for these segments, potentially overfitting the data.

substantial gain in fitting the observed copy number fractions. Second, we find that the entries of the usage matrix U for $n \leq 6$ have well-supported proportions with reasonable mixing proportions for each clone in several samples (Fig. 7). In contrast, for $n > 6$, we identify clones with very low mixing proportions across samples (such as $\mathbf{c}_5$ for $n=7$ and $\mathbf{c}_4$ for $n=8$), suggesting that the additionally inferred clones are not supported by the data. Third, we consider the topologies and costs of inferred trees with a varying number of clones and find that the tree with $n=6$ clones best describes the data. We find that most of the clonal events, which are events that are shared by all tumor clones and occur on the first branch of the tree, are consistent across the majority of the trees with $n \leq 6$ clones (Fig. 10).

Moreover, the trees with $n \leq 6$ clones have a cascading topology with an additional branch for every increase in n. In contrast, with $n > 6$ clones, the trees conserve the same cascading topology and each additional clone splits a previous clone (from the tree with $n-1$ clones) into two new sibling clones, potentially overfitting the data (Fig. 6). The total number of events, $\Lambda(T)$, stabilizes between $n=5$ and $n=6$ before increasing again for $n \geq 6$. The trees with more than six clones have several edges with only a few events as opposed to the trees with $n \leq 6$ clones. In sum, these findings suggest that the tree with

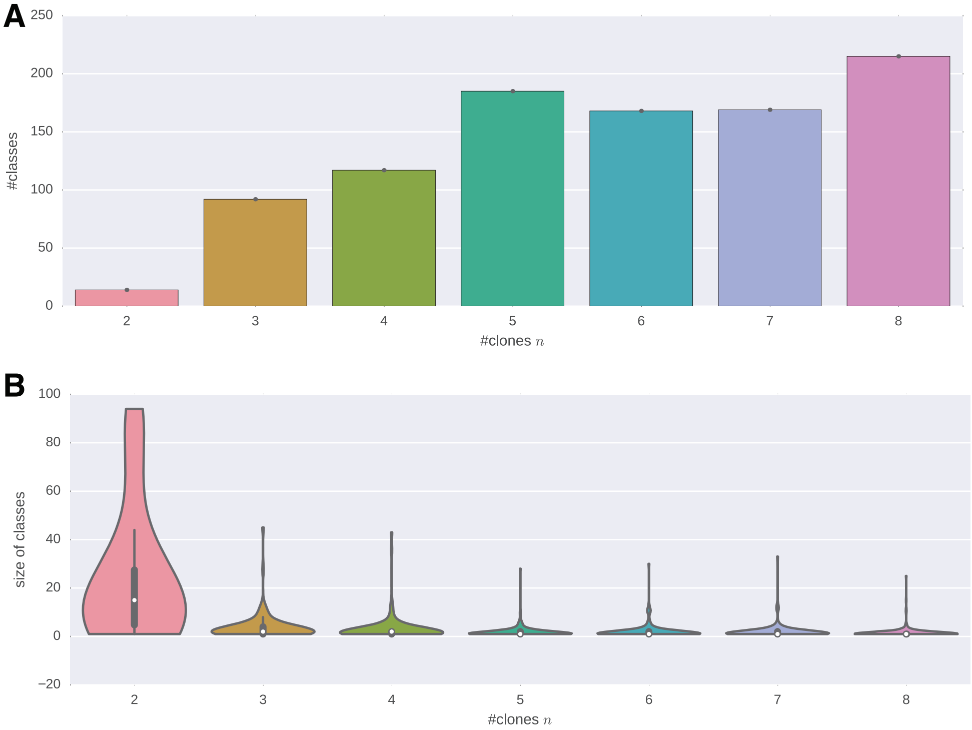

FIG. 9. Copy-number evolutionary history of segments stabilizes with increasing number of clones for patient A22 from Gundem et al. (2015). We partition the segments into classes with the same evolutionary history. Hence, **(A)** shows the number of classes for each CNT T inferred with $n \in \{2, \ldots, 8\}$ clones, while **(B)** shows the distribution of the size of classes for each CNT T inferred with $n \in \{2, \ldots, 8\}$ clones. Note that both the number and size of classes stabilize with $n > 6$ clones.

$n = 6$ clones provides a good explanation of the data in comparison with the other trees that either overfit the data ($n > 6$) or do not accurately represent the clonal structure of the data ($n < 6$).

Finally, we examine the relationship between the inferred matrix C and the observed fractional copy number matrix F, checking whether segments with close values of F across samples are assigned the same copy number values in C, as we vary the number n of clones. We do this by partitioning the segments into *classes* with the same evolutionary history in the inferred tree T (which is derived from the inferred C) as shown in Figure 8. Specifically, we define a class to be a set of segments that have the same copy-number change on all edges of T. Consequently, segments in the same class have the same copy number in all the clones. We observe that with increasing n the number of classes increases, whereas their size decreases (Fig. 9). However, the size and number of classes do not significantly change with $n \geq 6$. Next, we assess the consistency between these classes and F. For each pair p_1, p_2 of samples, we plot the fractional copy numbers of each segment in these samples, coloring segments by their class (overlapping segments with the same values result in larger dots). Figure 8 gives a schematic of this procedure. We see that for $n < 6$, segments in the same class are apart in at least one pair of samples (red, dark blue, and green clusters in Fig. 8), suggesting a poor fit to the data. On the contrary, for $n > 6$, segments with slightly different fractional copy numbers are separated (red/white clusters for $k = 7$ and light blue/cyan clusters for $k = 8$), suggesting overfitting of the data. Thus, this analysis also indicates that $n = 6$ appears to provide a reasonable partition into classes.

We also compare our inferred clonal CNAs to the published clonal CNAs in Gundem et al. (2015) (i.e., CNA labeling edges in the published trees in Gundem et al. (2015). We observe that several clonal events in our inferred T correspond to the these CNAs (Fig. 10): three inferred deletions on chr12 match the reported 12p loss of heterozygosity (LOH); a deletion with a subsequent amplification on chr13 matches the reported 13q LOH; a deletion on chr8 matches the 8p LOH; an amplification on the same chr8 matches the 8q gain; and two chr16 deletions match the reported 16q LOH. More interestingly, most of these events are clonal in the majority of the inferred trees for every n (Fig. 10). Thus, other recurrent and well-supported events in the inferred tree T are likely to be real, giving additional information about the clonal composition of these samples.

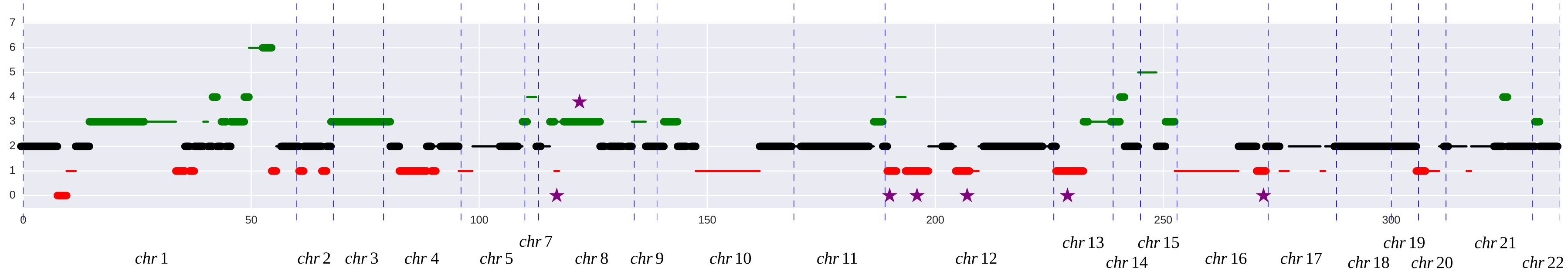

FIG. 10. The copy numbers of inferred clonal events ($n = 6$ clones) for patient A22 correspond to previously published analyses (Gundem et al., 2015). This plot shows the copy numbers corresponding to the clonal events inferred with $n = 6$ clones. We indicate separate chromosomes with dashed blue lines (chromosome 6 has been filtered out due to outliers). Green lines indicate amplifications and red lines indicate deletions. Note that the lengths are proportional to the number of segments and not to the corresponding genome length. Thick lines indicate events that are shared by the majority of the inferred trees T (with varying n). Purple stars indicate events that correspond to published clonal CNAs (Gundem et al., 2015).

5. DISCUSSION

In the article, we formulated the CNTMD Problem, and derived a coordinate-descent algorithm, with alternating ILP and LP steps, to solve this problem. CNTMD builds a phylogenetic tree describing copy number evolution directly from mixed samples, thus addressing an important issue in applying phylogenetic analysis to tumor samples. We show that CNTMD outperforms approaches that only perform deconvolution—thus ignoring the phylogenetic relationship between samples—or that build phylogenetic trees assuming that each sample is homogeneous, that is, consisting of a single clone. We also apply CNTMD to a complex metastatic prostate cancer data set and build a phylogenetic tree containing multiple distinct clones, mixed in different proportions across samples. These results demonstrate the feasibility of our approach to real-sized data sets.

There are a number of directions for future work. On the theoretical side, the hardness of CNTMD remains open. Assuming the problem is intractable, better heuristics for solving the C-step would improve the performance with increasing number of clones. An additional avenue of investigation is to incorporate uncertainty in the segmentation of the genome into the model. Finally, one could extend the approach using more sophisticated models of genome evolution, including models that include additional genome rearrangements and complex patterns of duplication—some promising work in this direction is found in Li et al. (2016), Ma et al. (2008), Oesper et al. (2012), and McPherson et al. (2015).

For practical applications, a number of improvements would be helpful. First, approaches to better address noise in the copy number fractions, using confidence intervals or posterior distributions to model the uncertainty in entries of F, are needed. Next, model selection or regularization approaches to estimate the number of clones in a tree and avoid overfitting would be helpful. For example, we report $n=6$ clones in the prostate cancer sample A22, while the original analysis (Gundem et al., 2015) reports 16 clones. This difference in the number of clones is likely due to the fact that Gundem et al. (2015) use SNVs to identify clones. Thus, methods that simultaneously identify CNAs and perform phylogeny inference from CNAs and SNVs are an important direction for future work. Finally, one could augment the phylogenetic reconstructions with single-cell measurements, including FISH (Chowdhury et al., 2014), or single-cell sequencing (Davis and Navin, 2016). Together, these improvements would enable high-fidelity phylogenetic reconstructions of tumor evolution.

ACKNOWLEDGMENTS

This work is supported by a U.S. National Science Foundation (NSF) CAREER Award (CCF-1053753) and U.S. National Institutes of Health (NIH) grants R01HG005690 and R01HG007069 to B.J.R.. B.J.R. is supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.

AUTHOR DISCLOSURE STATEMENT

B.J.R. is a cofounder and consultant at Medley Genomics.

REFERENCES

Baumbusch, L., Aarøe, J., Johansen, F., et al. 2008. Comparison of the agilent, roma/nimblegen and illumina platforms for classification of copy number alterations in human breast tumors. *BMC Genomics* 9, 379.

Carter, S.L., Cibulskis, K., Helman, E., et al. 2012. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* 30, 413–421.

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, 1–19.

Davis, A., and Navin, N.E. 2016. Computing tumor trees from single cells. *Genome Biol.* 17, 1.

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, 1.

El-Kebir, M., Oesper, L., Acheson-Field, H., et al. 2015. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics* 31, i62–i70.

El-Kebir, M., Raphael, B.J., Shamir, R., et al. 2016a. *Copy-Number Evolution Problems: Complexity and Algorithms.* In: Frith, M., Storm Pedersen, C. (eds). Algorithms in Bioinformatics. WABI 2016 Lecture Notes on Computer Science, Vol. 9838. Springer International Publishing.

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. 2016b. Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. *Cell Syst.* 3, 43–53.

Fischer, A., Vázquez-García, I., Illingworth, C.J., et al. 2014. High-definition reconstruction of clonal composition in cancer. *Cell Rep.* 7, 1740–1752.

Gawad, C., Koh, W., and Quake, S.R. 2016. Single-cell genome sequencing: Current state of the science. *Nat. Rev. Genet.* 17, 175–188.

Gerlinger, M., Rowan, A.J., Horswell, S., et al. 2012. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* 366, 883–892.

Gundem, G., Van Loo, P., Kremeyer, B., et al. 2015. The evolutionary history of lethal metastatic prostate cancer. *Nature* 520, 353–357.

Ha, G., Roth, A., Khattra, J., et al. 2014. Titan: Inference of copy number architectures in clonal cell populations from tumor whole genome sequence data. *Genome Res*. 24, 1881–1893.

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.* 113, E5528–E5537.

Li, Y., Zhou, S., Schwartz, D.C., et al. 2016. Allele-specific quantification of structural variations in cancer genomes. *Cell Syst.* 3, 21–34.

Ma, J., Ratan, A., Raney, B.J., et al. 2008. The infinite sites model of genome evolution. *Proc. Natl Acad. Sci. U. S. A.* 105, 14254–14261.

Malikic, S., McPherson, A.W., Donmez, N., et al. 2015. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics* 31, 1349–1356.

Mangiola, S., Hong, M.K., Cmero, M., et al. 2016. Comparing nodal versus bony metastatic spread using tumour phylogenies. *Sci. Rep.* 6, 33918.

McPherson, A., Roth, A., Chauve, C., et al. 2015. Joint inference of genome structure and content in heterogeneous tumor samples. *In International Conference on Research in Computational Molecular Biology*, 256–258. Springer.

McPherson, A., Roth, A., Laks, E., et al. 2016. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nat. Genet*. 48, 758–767.

Nijenhuis, A., and Wilf, H.S. 2014. *Combinatorial algorithms for computers and calculators.* Academic Press, New York.

Nik-Zainal, S., Van Loo, P., Wedge, D.C., et al. 2012. The life history of 21 breast cancers. *Cell* 149, 994–1007.

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

Oesper, L., Mahmoody, A., and Raphael, B.J. 2013. Theta: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol.* 14, R80.

Oesper, L., Ritz, A., Aerni, S.J., et al. 2012. Reconstructing cancer genomes from paired-end sequencing data. *BMC Bioinformatics* 13, S10.

Roman, T., Xie, L., and Schwartz, R. 2016. Medoidshift clustering applied to genomic bulk tumor data. *BMC Genomics* 17, 6.

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, 1–20.

Schwarz, R.F., Trinh, A., Sipos, B., et al. 2014. Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Comput Biol.* 10, 1–11.

Shamir, R., Zehavi, M., and Zeira, R. 2016. A linear-time algorithm for the copy number transformation problem. *In LIPIcs-Leibniz International Proceedings in Informatics*, vol. 54. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Sottoriva, A., Kang, H., Ma, Z., et al. 2015. A Big Bang model of human colorectal tumor growth. *Nat. Genet.* 47, 209–216.

Van Loo, P., Nordgard, S.H., Lingjærde, O.C., et al. 2010. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci.* 107, 16910–16915.

Venkatesan, S., and Swanton, C. 2015. Tumor evolutionary principles: How intratumor heterogeneity influences cancer treatment and outcome. *Am. Soc. Clin. Oncol. Educ. Book* 35, e141–49.

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.

Address correspondence to:
Prof. Benjamin J. Raphael
Department of Computer Science
Princeton University
Princeton, NJ 08540

E-mail: braphael@princeton.edu