The Copy-Number Tree Mixture Deconvolution Problem and Applications to Multi-sample Bulk Sequencing Tumor Data

Simone Zaccaria[1,2], Mohammed El-Kebir[2,3], Gunnar W. Klau[2,4,5], and Benjamin J. Raphael[2,3](✉)

[1] Dipartimento di Informatica, Univ. degli Studi di Milano-Bicocca, Milan, Italy
[2] Department of Computer Science, Brown University, Providence, RI 02912, USA
[3] Department of Computer Science, Princeton University, Princeton, NJ 08540, USA
braphael@princeton.edu
[4] Life Sciences Group, Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands
[5] Algorithmic Bioinformatics, Heinrich Heine University, Düsseldorf, Germany

Abstract. Cancer is an evolutionary process driven by somatic mutation. 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 mutational complexity of cancer and the fact that nearly all cancer sequencing is of bulk tissue, 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. CNTMD generalizes two approaches that have been researched intensively in recent years: deconvolution/factorization algorithms that aim to infer the number and proportions of clones in a mixed tumor sample; and phylogenetic models of copy number evolution that model the dependencies between copy number events that affect the same genomic loci. 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 perform either deconvolution or phylogenetic tree construction under the assumption of a single tumor clone per sample. 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.

1 Introduction

Cancer results from an evolutionary process where somatic mutations accumulate in a population of cells during the lifetime of an individual [21]. Thus, a

S. Zaccaria and M. El-Kebir—Joint first authorship.

S.C. Sahinalp (Ed.): RECOMB 2017, LNBI 10229, pp. 318–335, 2017.
DOI: 10.1007/978-3-319-56970-3_20

tumor consists of heterogeneous subpopulations of cells, or *clones*. Each clone comprises cells that share a unique complement of somatic mutations. Quantifying this intra-tumor heterogeneity has been shown to be important in cancer treatment [29]. While intra-tumor heterogeneity complicates the identification of mutations 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 [12]. Thus, a number of methods have been developed to infer phylogenetic trees from DNA sequencing data from one or more samples of a tumor [5,8,12–14,17,19,20,28].

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 intra-tumor 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 [11]. 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 [1,16]. When a sample is heterogeneous, i.e. 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 [2,9,10,16,20,23]. However, these methods have certain limitations that limit their applicability and performance. For example, ASCAT [16] and ABSOLUTE [2] 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 [23], Battenberg [20], cloneHD [9] and TITAN [10], 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 [12], as successfully done for single-nucleotide mutations [5,8,14,17] or non-integer copy numbers [24]. 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 [3,19]. 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 [25]. Recently, [27] and [7] improved the model in MEDICC. Specifically, [27] formally investigated the effects of interval events on segments of a single clone. In [7], 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 program (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 [26,28] or first attempt to infer the clones independently on each sample before performing a phylogenetic analysis of CNAs [19].

In this paper, 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 copy number aberrations modeled as interval events (Fig. 1). 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 multi-sample sequencing data of a prostate-cancer patient [13]. Our inference shows well-supported patterns that reveal the clonal composition in terms of CNAs. The software is available at http://compbio.cs.brown.edu/software/.

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 *copy-number tree*, 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 copy-number tree.

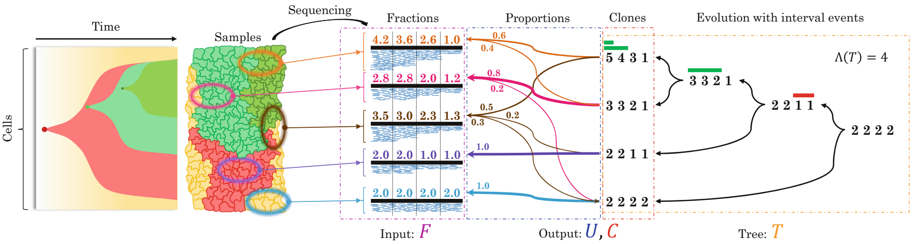

Fig. 1. Copy-Number Tree Mixture Deconvolution (CNTMD) problem. A tumor consists of heterogeneous subpopulations of cells, or clones. The normal clone is colored yellow. Five samples are bulk sequenced yielding fractional copy numbers F. We model the evolution of CNAs by a copy-number tree T (right). We combine the deconvolution of F with the inference of T. Thus, CNTMD factors F into the integer copy numbers C of the extant clones and their proportions U such that $F = CU$ and C generates a copy-number tree T with the minimum number $\Lambda(T)$ of interval events.

Following the model in [7,25,27], 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.

We consider mutations that amplify or delete contiguous segments. An *interval event*, or *event*, 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 the non-zero segments between s and t are incremented by b, whereas for negative b the events is a *deletion* and the same segments are decremented by at most $|b|$. 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, $c'_{\ell,i} = \max\{c_{\ell,i} + b, 0\}$ if $s \leq \ell \leq t$ and $c_{\ell,i} \neq 0$, or $c'_{\ell,i} = c_{\ell,i}$ otherwise. Thus, once a segment ℓ has been lost, i.e. $c_{\ell,i} = 0$, it can never be regained (or deleted).

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

Definition 1. *Given a number n of clones, a* copy-number tree *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, i.e. $c_{s,r(T)} = 2$ 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 of $\mathcal{E}(i, j)$ matters. Following a result by Shamir et al. [27], 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 $\varLambda(i, j)$ of an edge (v_i, v_j) is the total cost of the events in $\mathcal{E}(i, j)$, i.e. $\varLambda(i, j) = \sum_{(s,t,b) \in \mathcal{E}(i,j)} |b|$. The cost $\varLambda(T)$ of the tree T is the sum of the costs of all edges.

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, i.e. a minimum-cost copy-number tree T^* whose n leaves are labeled by $\mathbf{c}_1, \ldots, \mathbf{c}_n$. Previously, we have shown that this problem, the Copy-Number Tree (CNT) problem, is NP-hard and we introduced an ILP formulation for solving it [7]. 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 copy-number tree 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 copy-number tree 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, i.e. $C = (\mathbf{c}_1, \ldots, \mathbf{c}_n)$. We say that C *generates* T 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 $\varLambda(T^*)$ is minimum.

3 Method

The hardness of CNTMD is an open question. However, we suspect the problem to NP-hard, as the related unmixed version, the CNT problem, is NP-hard [7]. Moreover, other similar deconvolution problems under a tree constraint are NP-hard as well [6,8]. 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:

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

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

Problem 2 (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 $c_{s,i} \leq c_{\max}$, $\Lambda(T) \leq \Lambda_{\max}$ and $\|F - CU\|$ is 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 $\|F - C_t U_{t-1}\|$ such that $c_{s,i} \leq c_{\max}$ and C generates T with cost $\Lambda(T) \leq \Lambda_{\max}$. Conversely, in the *$U$-step* we take the matrix C_t as input and seek a usage matrix U_t such that $\|F - C_t U_t\|$ 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. 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 $\|F - C_{q,t} U_{q,t}\| \geq \|F - C_{q,t+1} U_{q,t+1}\|$. 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 copy-number tree $T_{q,K}$ with cost $\Lambda(T_{q,K}) \leq \Lambda_{\max}$ and, consequently, we have that the final matrix C' generates a copy-number tree T' with cost $\Lambda(T') \leq \Lambda_{\max}$. Thus, it may be the case that for the same usage matrix U^* there exist another copy-number matrix C'' different from C' that generates a copy-number tree 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 integer linear programming (ILP) formulation of the C-step. Figure 2 depicts the entire procedure of the coordinate-descent algorithm.

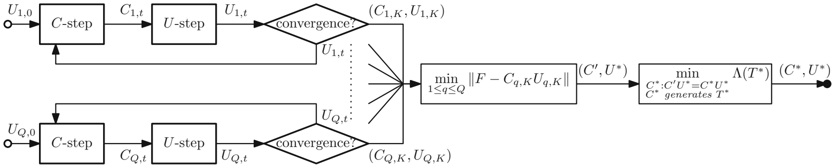

Fig. 2. Coordinate-descent algorithm. Given an initial usage matrix $U_{q,0}$, the algorithm alternatingly solves two distinct steps for at most K iterations. The C-step computes a copy-number matrix $C_{q,t}$ given the previous usage matrix $U_{q,t-1}$ and is followed by the U-step, which computes a usage matrix $U_{q,t}$ given $C_{q,t}$. We repeat the procedure using Q restarts with different initial usage matrices, yielding Q pairs $(C_{q,K}, U_{q,K})$ of matrices. Given these final matrices, the refinement step searches for a copy-number matrix that generates a copy-number tree with minimum cost.

We present a linear programming (LP) formulation for the U-step in Sect. 3.1 followed by an integer linear programming (ILP) formulation for the C-step in Sect. 3.2. Since the distance-based variant of the problem does not directly minimize the cost of the tree, we present in Sect. 3.3 an algorithm for finding the smallest maximum cost Λ^* with 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 (1), we introduce variables $\bar{f}_{s,p}$ for each segment s and sample p, and model $\bar{f}_{s,p} = |f_{s,p} - \sum_{1 \leq i \leq n} c_{s,i} u_{i,p}|$ using the following linear constraints.

$$\bar{f}_{s,p} \geq f_{s,p} - \sum_{1 \leq i \leq n} c_{s,i} u_{i,p} \qquad 1 \leq s \leq m, 1 \leq p \leq k \tag{2}$$

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

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

$$\sum_{1 \leq i \leq n} u_{i,p} = 1 \qquad 1 \leq p \leq k \tag{4}$$

Thus, we have the following LP: $\min_{\mathbf{u}, \bar{\mathbf{f}}} \sum_{1 \leq s \leq m, 1 \leq p \leq k} \bar{f}_{s,p}$ s.t. (2), (3) and (4).

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 generates 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 (2) and (3). 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 upon the model introduced in [7].

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 is a submatrix of $\tilde{C}$. Recall that T is a full binary tree (Definition 1). We construct a directed acyclic graph $G = (V, E)$ that contains all copy-number trees 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 $\{(v_i, v_j) \mid n + 1 \leq i < 2n - 1, 1 \leq j < i \leq 2n - 1\}$. 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.

$$\sum_{i \geq j, i \geq n+1} x_{i,j} = 1 \qquad 1 \leq j < 2n - 1 \tag{5}$$

$$\sum_{1 \leq j < i} x_{i,j} = 2 \qquad n < i \leq 2n - 1 \tag{6}$$

Integer variables $\tilde{C} = [c_{s,i}]$ where $c_{s,i} \in \{0, \ldots, c_{\max}\}$ encode the profiles of each vertex v_i. Since the root vertex is diploid, we add the following constraints.

$$c_{s,2n-1} = 2 \qquad 1 \leq s \leq m \tag{7}$$

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.

$$a_{s,i,j},\ d_{s,i,j} \leq c_{\max} x_{i,j} \qquad 1 \leq s \leq m, (v_i, v_j) \in E(G) \tag{8}$$

Due to these constraints, the cost of every pair (v_i, v_j) of vertices that do not form an edge of T, i.e. $x_{i,j} = 0$, is fixed to 0. Therefore, only the cost of the edges of T is computed, which significantly constraints the model and improves the performance over the formulation presented for the unmixed CNT problem [7].

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. Similarly to [7], we distinguish four different cases. Case (a) is $c_{s,i} = 0$ and

$c_{s,j} = 0$: Since both segments have value 0, we have that, following a result in [27], the number of amplifications $a_{s,i,j}$ and deletions $d_{s,i,j}$ must be between 0 and $c_{\max}$. Case (b) is $c_{s,i} \neq 0$ 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 $c_{s,j} + d_{s,i,j} = c_{s,i} + a_{s,i,j}$. Case (c) is $c_{s,i} \neq 0$ and $c_{s,j} = 0$: Since deletions precede amplifications, the number of deletions $d_{s,i,j}$ must be at least $c_{s,i}$. Case (d) is $c_{s,i} = 0$ and $c_{s,j} \neq 0$: Once a segment s has been lost it cannot be regained. As such, this case is infeasible.

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 $c_{s,i}$. We define $L := \lfloor \log_2(c_{\max}) \rfloor + 1$. 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 $c_{s,i} \neq 0$.

$$c_{s,i} = \sum_{q=0}^{L} 2^q \cdot z_{i,s,q} \qquad 1 \leq i \leq 2n-1, 1 \leq s \leq m \tag{9}$$

$$z_{i,s,q} \leq \bar{c}_{s,i} \leq \sum_{q'=0}^{L} z_{i,s,q'} \qquad 1 \leq i \leq 2n-1, 1 \leq s \leq m, 0 \leq q \leq L \tag{10}$$

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 where $(v_i, v_j) \in E(G)$.

$$c_{s,j} \leq c_{s,i} - d_{s,i,j} + a_{s,i,j} + 2c_{\max}(3 - \bar{c}_{i,s} - \bar{c}_{j,s} - x_{i,j}) \qquad 1 \leq s \leq m \tag{11}$$

$$c_{s,j} + 2c_{\max}(3 - \bar{c}_{s,i} - \bar{c}_{s,j} - x_{i,j}) \geq c_{s,i} - d_{s,i,j} + a_{s,i,j} \qquad 1 \leq s \leq m \tag{12}$$

$$d_{i,j,s} \leq c_{s,i} - 1 + (c_{\max} + 1)(2 - \bar{c}_{s,i} - \bar{c}_{s,j}) \qquad 1 \leq s \leq m \tag{13}$$

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 (11) and (12) model the equation $c_{s,j} + d_{s,i,j} = c_{s,i} + a_{s,i,j}$, whereas constraint (13) 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 [7]). Note that $d_{s,i,j}$ can be always equal to zero by constraint (13), 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.

$$c_{s,i} \leq d_{s,i,j} + c_{\max}(2 - \bar{c}_{s,i} + \bar{c}_{s,j} - x_{i,j}) \qquad 1 \leq s \leq m, (v_i, v_j) \in E(G) \tag{14}$$

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

$$1 - x_{i,j} + \bar{c}_{s,i} \geq \bar{c}_{s,j} \qquad 1 \leq s \leq m, (v_i, v_j) \in E(G) \tag{15}$$

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 $\bar{d}_{s,i,j}$ corresponds to the deletions starting at segment s and is equal to

$\max\{d_{s,i,j} - d_{s-1,i,j}, 0\}$. We model this using the following constraints.

$$\bar{a}_{s,i,j} \geq a_{s,i,j} - a_{s-1,i,j} \qquad 1 \leq s \leq m, (v_i, v_j) \in E(G) \tag{16}$$

$$\bar{d}_{s,i,j} \geq d_{s,i,j} - d_{s-1,i,j} \qquad 1 \leq s \leq m, (v_i, v_j) \in E(G) \tag{17}$$

$$a_{0,i,j} = d_{0,i,j} = 0 \qquad (v_i, v_j) \in E(G) \tag{18}$$

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.

$$\bar{a}_{s,i,j},\ \bar{d}_{s,i,j} \leq c_{\max} x_{i,j} \qquad 1 \leq s \leq m, (v_i, v_j) \in E(G) \tag{19}$$

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.

$$\sum_{(v_i,v_j) \in E(G)} \sum_{1 \leq s \leq m} (\bar{a}_{s,i,j} + \bar{d}_{s,i,j}) \leq \Lambda_{\max} \tag{20}$$

The ILP is thus: $\min_{\mathbf{c},\bar{\mathbf{f}}} \sum_{1 \leq s \leq m, 1 \leq p \leq k} \bar{f}_{s,p}$ s.t. (2), (3), (5)–(20).

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 non-increasing with larger values of Λ_t. That is, if $\Lambda_t \geq \Lambda$ then $d(\Lambda_t) \leq d(\Lambda)$, as C^{Λ} generates T with cost $\Lambda(T) < \Lambda_t$. The parameter $\Lambda_{\max}$ controls the tradeoff 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 tradeoff 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 $d(\Lambda^*) - d(\Lambda_t) \leq \varepsilon$ for any $\Lambda_t \geq \Lambda^*$. Note that in a similar fashion ε plays a role in the refinement step described previously. We use the monotonicity of the function $d(\Lambda_t)$ and employ binary search for finding the value Λ^*.

4 Results

We applied our algorithm for CNTMD to simulated data and to data from two patients from a prostate cancer dataset [13]. We ran every experiment in this

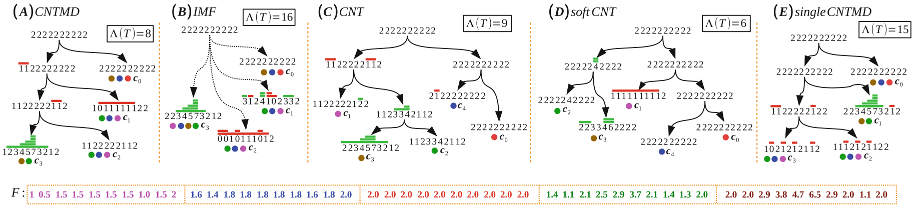

Fig. 3. Alternative methods infer trees that differ significantly from the true tree, which is inferred by our approach CNTMD. Copy-number trees inferred by the alternative methods where deletions $(s, t, -1)$ are red and amplifications $(s, t, 1)$ are green. (A) Shows the true tree composed of four 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 (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.

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 implementation of the alternative methods are available at http://compbio.cs.brown.edu/software/.

We benchmarked CNTMD on simulated data, comparing its performance to several other approaches, which we now describe. 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 [9,19,20,23]—one exception is [24], but this infers non-integer 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 fashion as the procedure described in Sect. 3.

Another class of approaches use the same copy number model as CNTMD, but assume that each sample is unmixed. One strategy is to first round the entries of F before inferring a copy-number tree. We will do this by solving the CNT problem with an ILP model [7], mimicking the strategy that has been used with MEDICC [26,28]. 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 [7].

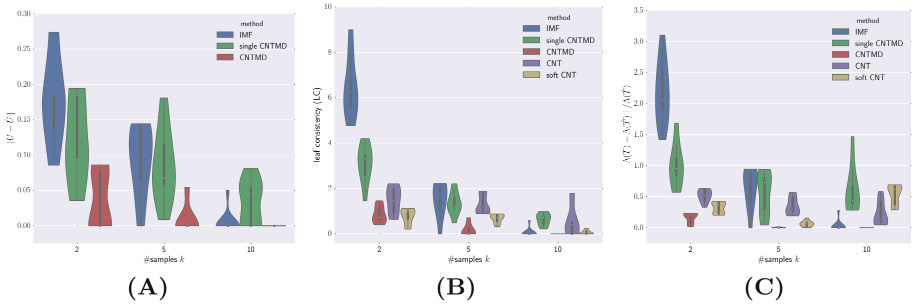

Fig. 4. CNTMD outperforms alternative methods on simulated data. Comparison of five methods across 27 simulated datasets with $k \in \{2, 5, 10\}$ samples, consisting of 4 tumor clones and a normal diploid clone, each with a total of 350 segments across 22 chromosomes. Each simulated instance was solved with n set to the true number of clones. (A) Normalized usage difference $\|U - \hat{U}\|$. (B) Leaf consistency (LC) measure. (C) Difference $|\Lambda(T) - \Lambda(\hat{T})| / \Lambda(\hat{T})$.

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 copy number aberrations in cancer, as the length distribution of somatic copy number aberrations is not simply a function of length [30]. Such a copy number model was used by [19] and [3] 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.

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 datasets available in [13]. Using a procedure similar to the one described in [7], we randomly generate three copy-number trees, 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 [6, 17], 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 $(i, \hat{\imath})$. To compare C to $\hat{C}$, 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 [7] 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 copy-number tree 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 matching weight 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 and 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 and C), demonstrating how CNTMD leverages the extra information given by the tree constraint. Moreover, by not accounting for interval events, single CNTMD results in copy-number trees that are inconsistent with the simulated trees and have many more events (Fig. 4C).

4.1 Application to Prostate Cancer Dataset

We apply our approach on prostate cancer patient A22 from the dataset of Gundem et al. [13]. Patient A22 comprises 10 samples. We use the published fractional copy numbers that were obtained by the Battenberg algorithm [20], which relies on the sample purity and tumor ploidy estimated by ASCAT [16].

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 copy number aberrations; in comparison [13] estimate 16 clones using 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 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 (data not shown). In contrast, for $n > 6$, we identify clones with very low mixing proportions across samples (such

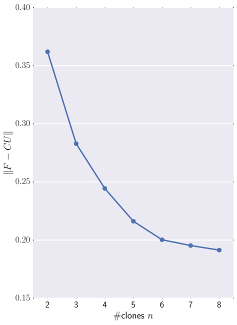

Fig. 5. The distance decreases with increasing number of clones n and stabilizes with $n > 6$ clones. The y-axis shows the normalized value of the distance $\|F - CU\|$ for each n.

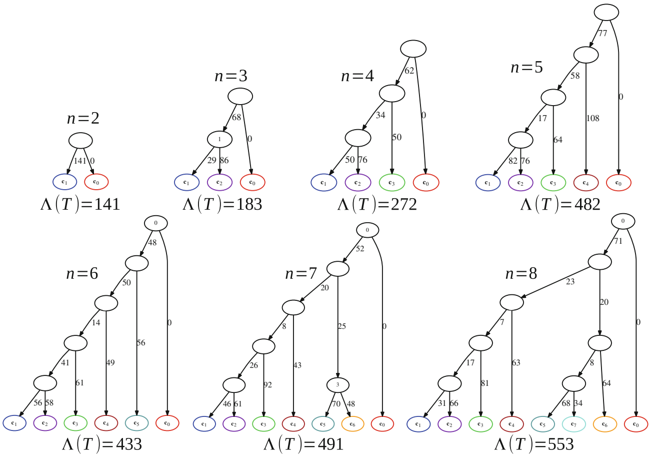

Fig. 6. Trees with $n \leq 6$ clones have a cascading topology and well-supported edges, whereas trees with $n > 6$ clones have the same cascading topology but have less-supported edges. For each copy-number tree T, we show the cost $\Lambda(T)$ and label the edges by their corresponding costs. The colors of leaves map corresponding clones in the topologies. The normal clone is red.

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 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. 8). 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 $n > 6$ 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 $n = 6$ clones provides a good explanation of the data in comparison with the other trees that either overfit ($n > 6$) 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). 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 (data not shown). 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 7 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. 7), suggesting a poor fit to the data. On the other hand, 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 copy number aberrations (CNAs) to the published clonal CNAs in [13]: We observe that several clonal events in our inferred T correspond to the these CNAs (Fig. 8): three inferred deletions on chr12 match the reported 12p 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. 8). 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.

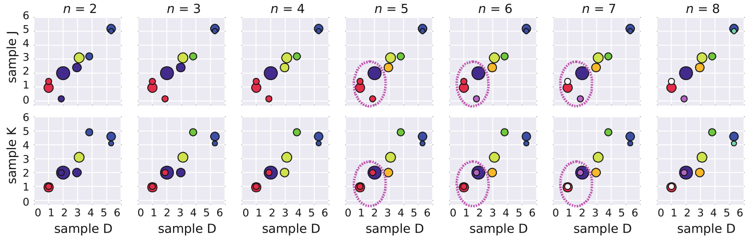

Fig. 7. Classes of segments with the same evolutionary history highlight consistency of the inferred solutions with the input data. Fractional copy numbers for three A22 prostate cancer samples: D, K and J. The largest dot contains 14 segments. The consistency of the classes improves with increasing n. The red class in $n = 5$ is composed of segments that have one copy in all the considered samples, and segments that have two copies in samples D, K and zero copies copies in sample J. With $n = 6$ these two subsets are separated into different classes (red and purple), while with $n = 7$ one more class (white) is introduced, potentially overfitting the data.

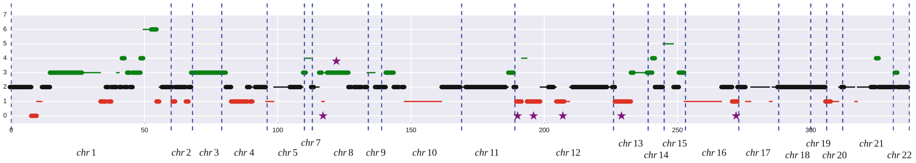

Fig. 8. Well-supported clonal events correspond to published clonal CNAs. This plot shows the copy numbers of the clonal events inferred with $n = 6$ clones. We indicate separate chromosomes with dashed blue lines. Green lines indicate amplifications and red lines indicate deletions. The lengths are proportional to the number of segments. 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 [13].

5 Discussion

In the paper, we formulated the Copy-Number Tree Mixture Deconvolution (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, i.e. consisting of a single clone. We also apply CNTMD to a complex metastatic prostate cancer dataset, 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 datasets.

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 [15,18,22]. 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 [13] reports 16 clones. This difference is likely due to the fact that [13] use single-nucleotide variants (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 [3] or single-cell sequencing [4]. Together, these improvements would enable high-fidelity phylogenetic reconstructions of tumor evolution.

Acknowledgements. This work is supported by a US National Science Foundation (NSF) CAREER Award (CCF-1053753) and US National Institutes of Health (NIH) grants R01HG005690 and R01HG007069 to BJR. BJR is supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, an Alfred P. Sloan Research Fellowship.

References

1. Baumbusch, L.O., et al.: Comparison of the agilent, ROMA/NimbleGen and Illumina platforms for classification of copy number alterations in human breast tumors. BMC Genom. **9**(1), 379 (2008)
2. Carter, S.L., et al.: Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. **30**(5), 413–421 (2012)
3. Chowdhury, S.A., et al.: Algorithms to model single gene, single chromosome, and whole genome copy number changes jointly in tumor phylogenetics. PLoS Comput. Biol. **10**(7), 1–19 (2014)
4. Davis, A., et al.: Computing tumor trees from single cells. Genome Biol. **17**, 1 (2016)
5. Deshwar, A.G., et al.: PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol. **16**(1), 1 (2015)
6. El-Kebir, M., et al.: Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. Bioinformatics **31**(12), i62–i70 (2015)
7. El-Kebir, M., Raphael, B.J., Shamir, R., Sharan, R., Zaccaria, S., Zehavi, M., Zeira, R.: Copy-number evolution problems: complexity and algorithms. In: Frith, M., Storm Pedersen, C.N. (eds.) WABI 2016. LNCS, vol. 9838, pp. 137–149. Springer, Heidelberg (2016). doi:10.1007/978-3-319-43681-4_11
8. El-Kebir, M., et al.: Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. Cell Syst. **3**(1), 43–53 (2016)
9. Fischer, A., et al.: High-definition reconstruction of clonal composition in cancer. Cell Rep. **7**(5), 1740–1752 (2014)
10. Gavin, H., et al.: Titan: inference of copy number architectures in clonal cell populations from tumor whole genome sequence data. Genome Res. **24**, 1881–1893 (2014)
11. Gawad, C., et al.: Single-cell genome sequencing: current state of the science. Nat. Rev. Genet. **17**(3), 175–188 (2016)
12. Gerlinger, M., et al.: Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N. Engl. J. Med. **366**(10), 883–892 (2012)
13. Gundem, G., et al.: The evolutionary history of lethal metastatic prostate cancer. Nature **520**(7547), 353–357 (2015)
14. Jiang, Y., et al.: Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. PNAS **113**, E5528–E5537 (2016)
15. Li, Y., et al.: Allele-specific quantification of structural variations in cancer genomes. Cell Syst. **3**(1), 21–34 (2016)

16. Van Loo, P., et al.: Allele-specific copy number analysis of tumors. PNAS **107**, 16910–16915 (2010)
17. Malikic, S., et al.: Clonality inference in multiple tumor samples using phylogeny. Bioinformatics **31**(9), 1349–1356 (2015)
18. McPherson, A., Roth, A., Chauve, C., Sahinalp, S.C.: Joint inference of genome structure and content in heterogeneous tumor samples. In: Przytycka, T.M. (ed.) RECOMB 2015. LNCS, vol. 9029, pp. 256–258. Springer, Cham (2015). doi:10.1007/978-3-319-16706-0_25
19. McPherson, A., et al.: Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. Nat. Genet. **48**, 758–767 (2016). doi:10.1038/ng.3573
20. Nik-Zainal, S., et al.: The life history of 21 breast cancers. Cell **149**(5), 994–1007 (2012)
21. Nowell, P.C.: The clonal evolution of tumor cell populations. Science **194**, 23–28 (1976)
22. Oesper, L., et al.: Reconstructing cancer genomes from paired-end sequencing data. BMC Bioinform. **13**(6), S10 (2012)
23. Oesper, L., et al.: THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. Genome Biol. **14**(7), R80 (2013)
24. Roman, T., et al.: Medoidshift clustering applied to genomic bulk tumor data. BMC Genom. **17**(1), 6 (2016)
25. Schwarz, R.F., et al.: Phylogenetic quantification of intra-tumour heterogeneity. PLoS Comput. Biol. **10**(4), 1–11 (2014)
26. Schwarz, R.F., et al.: Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. PLoS Med **12**(2), 1–20 (2015)
27. Shamir, R., et al.: A linear-time algorithm for the copy number transformation problem. In: LIPIcs, vol. 54. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2016)
28. Sottoriva, A., et al.: A Big Bang model of human colorectal tumor growth. Nat. Genet. **47**(3), 209–216 (2015)
29. Venkatesan, S., et al.: Tumor evolutionary principles: How intratumor heterogeneity influences cancer treatment and outcome. ASCO **35**, e141–e149 (2015)
30. Zack, T.I., et al.: Pan-cancer patterns of somatic copy number alteration. Nat. Genet. **45**(10), 1134–1140 (2013)