Bioinformatics, 36, 2020, i344–i352
doi: 10.1093/bioinformatics/btaa470
ISMB 2020

Copy number evolution with weighted aberrations in cancer

Ron Zeira and Benjamin J. Raphael*

Department of Computer Science, Princeton University, Princeton, NJ 08544, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Copy number aberrations (CNAs), which delete or amplify large contiguous segments of the genome, are a common type of somatic mutation in cancer. Copy number profiles, representing the number of copies of each region of a genome, are readily obtained from whole-genome sequencing or microarrays. However, modeling copy number evolution is a substantial challenge, because different CNAs may overlap with one another on the genome. A recent popular model for copy number evolution is the copy number distance (CND), defined as the length of a shortest sequence of deletions and amplifications of contiguous segments that transforms one profile into the other. In the CND, all events contribute equally; however, it is well known that rates of CNAs vary by length, genomic position and type (amplification versus deletion).

Results: We introduce a weighted CND that allows events to have varying weights, or probabilities, based on their length, position and type. We derive an efficient algorithm to compute the weighted CND as well as the associated transformation. This algorithm is based on the observation that the constraint matrix of the underlying optimization problem is totally unimodular. We show that the weighted CND improves phylogenetic reconstruction on simulated data where CNAs occur with varying probabilities, aids in the derivation of phylogenies from ultra-low-coverage single-cell DNA sequencing data and helps estimate CNA rates in a large pan-cancer dataset.

Availability and implementation: Code is available at https://github.com/raphael-group/WCND.

Contact: braphael@princeton.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Cancer is an evolutionary process where somatic mutations accumulate in a population of tumor cells (Nowell, 1976). Copy number aberrations (CNAs), the deletion or amplification of large genomic regions, are a common type of somatic mutation in many cancer types (Ciriello *et al.*, 2013). CNAs range in scale from a few kilobases to chromosome arms and even entire chromosomes (Zack *et al.*, 2013). CNAs play an important role in driving cancer development (Burrell *et al.*, 2013; McGranahan and Swanton, 2015), and thus characterization of these events is essential for disease diagnosis, prognosis and treatment (Fisher *et al.*, 2013). Moreover, CNAs provide important information for reconstructing tumor evolution (Beerenwinkel *et al.*, 2015; Schwartz, 2019).

There are two major challenges in modeling copy number evolution. First, genomes containing multiple duplicated regions are difficult to correctly reconstruct from current short-read DNA sequencing technologies (Li *et al.*, 2016; McPherson *et al.*, 2017; Oesper *et al.*, 2012). Second, genome evolution models that allow multiple genomic copies are computationally hard to solve (Fertin *et al.*, 2009). To address these difficulties, recent research has focused on modeling the evolution of copy number profiles (CNPs), a simplified representation of a genome. A CNP is a sequence of integers that indicates the number of copies of each region, or segment, from a reference genome that are present in the genome. Thus, CNPs model only the number of copies of segments of the reference genome and *not* the sequence of rearranged segments. However, they are a useful representation because they can be readily derived from DNA sequencing data or microarrays.

CNPs can be derived from DNA sequencing data of bulk tumor samples using specialized algorithms that infer the integer-valued CNPs from the mixtures of normal and cancerous cells in this data (Carter *et al.*, 2012; Fischer *et al.*, 2014; Ha *et al.*, 2014; Nik-Zainal *et al.*, 2012; Oesper *et al.*, 2013; Shen and Seshan, 2016; Zaccaria *et al.*, 2018). While earlier methods calculated the total copy number of the two alleles (Carter *et al.*, 2012), recent methods derive allele-specific (McPherson *et al.*, 2017; Zaccaria and Raphael, 2018), and even haplotype-specific CNPs (Jamal-Hanjani *et al.*, 2017). Recently, single-cell sequencing has emerging as a promising approach for assessing tumor heterogeneity and evolution (Gawad *et al.*, 2016; Wang *et al.*, 2014). Single-cell sequencing precludes the need for deconvolution of bulk samples into integer CNPs, and thus enables the detection of small populations of cells with specific aberrations (Wang *et al.*, 2014). While high-coverage whole-genome single-cell sequencing is technically and financially prohibitive, two recent technologies, Direct Library Preparation (Laks *et al.*, 2018; Zahn *et al.*, 2017) and the 10X Genomics CNV Solution (10X Genomics, 2019a; Andor *et al.*, 2018), have demonstrated the feasibility of obtaining CNPs from thousands of single cells using ultra-low coverage ($<0.05\times$ per cell). While earlier methods derived total

copy numbers from single-cell sequencing (10X Genomics, 2019b; Garvin *et al.*, 2015), a recent method called CHISEL (Zaccaria and Raphael, 2019) derives allele-specific and haplotype-specific copy number calls, thus opening new opportunities for analyzing copy number evolution in cancer.

Modeling copy number evolution using CNPs is challenging because, unlike single-nucleotide mutations, CNAs often overlap, and therefore the copy numbers of different segments are not independent (Beerenwinkel *et al.*, 2015; Schwartz, 2019). Recently, several methods have been introduced to describe the evolution of CNPs. Some methods do not rely on an evolutionary model but instead use distance measures such as the Euclidean distance to reconstruct phylogenies from CNPs (Navin *et al.*, 2011; Pennington *et al.*, 2007). Other methods consider only events that alter the copy number of single segments independently (McPherson *et al.*, 2016) or include only single position events as well as whole-chromosome and whole-genome duplication events (Chowdhury *et al.*, 2014). An extension to the latter model allows for events of different weights (Chowdhury *et al.*, 2015), but both the unweighted and weighted models lack efficient algorithms to compute the distance between profiles; thus, applications of this model have been limited to very short profiles.

An alternative model of CNA evolution is the copy number transformation (CNT) model (Schwarz *et al.*, 2014). In this model, amplifications and deletions of contiguous intervals are counted as single events. The copy number distance (CND) between two profiles is the length of a shortest sequence of amplifications and deletions that transform one profile into the other. MEDICC (Schwarz *et al.*, 2014), the first algorithm to compute the CND, uses a heuristic to reconstruct a phylogenetic tree from CNPs, and has been used successfully in several cancer studies (Mangiola *et al.*, 2016; Schwarz *et al.*, 2015; Sottoriva *et al.*, 2015). More recently, Zeira *et al.* (2017) showed that the CND between a pair of profiles can be computed in linear time and El-Kebir *et al.* (2017) gave an integer linear programming formulation for reconstructing a phylogenetic tree between CNPs with the minimum number of events.

The CND is useful because it can be computed efficiently, but the CND has the disadvantage that it gives all events equal weight, regardless of their length, position or type (amplification versus deletion). This limitation has several drawbacks. First, CNAs are reported to have different rates in different cancers depending on their location, length and type (Beroukhim *et al.*, 2010; Ciriello *et al.*, 2013; Macintyre *et al.*, 2018; Zack *et al.*, 2013). Second, some events, such as those affecting oncogeneic regions, may have more profound effect on cancer development and thus are more important than others (Mermel *et al.*, 2011). Third, the CND is sensitive to errors in copy number calls as each change is counted as a single event. The discrete nature of CND makes it hard to distinguish small focal events that are possibly errors from large scale events such as chromosome losses.

Generalization of the CND to a model that weighs events differently is not straightforward. Moreover, it is not immediately apparent that a weighted generalization retains the combinatorial properties of the CNT model that enable its efficient computation. For example, there is no change in computational complexity between computing edit distance and weighted edit distance for sequences of independent characters. However, this is not the case for other models with non-independent characters. Most famously, while reversal distance can be computed in linear time (Hannenhalli and Pevzner, 1995a, b), weighted reversal distance is NP-hard (Bader and Ohlebusch, 2007; Pinter and Skiena, 2002). A recent generalization of the CNT model allows each event to modify the CNP with any amplitude at the same unit cost (Cordonnier and Lafond, 2020). But, computing the optimal transformation under this cost framework is NP-hard. Finally, the existing algorithms that compute the CND efficiently restrict the order of aberrations to specific *ordered* CNTs that have identical distances (El-Kebir *et al.*, 2017; Zeira *et al.*, 2017). This restriction is not appropriate for weighted CND. While the MEDICC algorithm (Schwarz *et al.*, 2014) computes CND without any assumption on the order of events, its algorithmic complexity has not been analyzed and is suggested to be exponential (Zeira *et al.*, 2017).

In this work, we derive a weighted CND and provide an efficient algorithm to compute this distance. The weighted CND allows for different weights (or probabilities) to be assigned to segmental events according to their genomic positions, lengths and/or types (amplification versus deletion). This is the first efficient algorithm for weighted CND and relies on two key results: (i) a generalization of the *ordered* CNTs used in the derivation linear-time algorithm for unweighted CND (Zeira *et al.*, 2017) to *semi-ordered* CNTs and (ii) formulation of weighted CNTs as a linear program (LP) with a totally unimodular constraint matrix, implying that integer solutions are obtained with a polynomial time algorithm. In addition to the distance, the algorithm also provides a minimum weight transformation, i.e. a likely series of amplifications and deletions between a pair of profiles.

We demonstrate the utility of the CND on three applications. First, we show that weighted CND produces more accurate phylogenetic trees on simulated CNPs generated with events with varying probabilities. Next, we use the weighted CND to derive phylogenetic trees from single-cell whole-genome sequencing data of a breast tumor obtained using the 10X Genomics CNV Solution. We show that the weighted CND improves the inference of tumor clones and cell lineages. Finally, we use the weighted CND to infer CNA rate signatures across cancer types and chromosomes on the Cancer Genome Atlas (TCGA) pan-cancer dataset.

2 Materials and methods

We start by reviewing CNTs and the CND (Section 2.1). We then describe the solution space of optimal CNTs, generalizing from ordered CNTs to semi-ordered CNTs (Section 2.2). Finally, we show how to compute optimal weighted CNTs where events have a weight determined by their position, length or type (Section 2.3). Additional details and proofs are in the Appendix (Supplementary Section S1).

2.1 CNT distance

We review the CNT model (Zeira *et al.*, 2017).

We model chromosomes and CNAs as follows. A *CNP* $C = (c_1, \ldots, c_n)$ is a vector of non-negative integers. A *segmental event* is a triplet $e = (i, j, \tau)$ where $1 \leq i \leq j \leq n$ are the start and end positions of the event and $\tau = \pm 1$ is the *type* of the event. An event with $\tau = 1$ is an *amplification* and an event with $\tau = -1$ a *deletion*. Segmental event $e = (i, j, \tau)$ transforms a profile C into a new profile $C' = e(C) = (c_1, \ldots, c_{i-1}, \max(c_i + \tau, 0), \ldots, \max(c_j + \tau, 0), c_{j+1}, \ldots, c_n)$; i.e. *positive* values between $c_i, \ldots, c_j$ are increased by τ.

A *CNT* from a source CNP S to a target CNP T is a sequence $E = (e_1, \ldots, e_l)$ of events such that $T = E(S) = e_l(\ldots e_1(S))$. Given a source CNP S and a target CNP T, the *copy number transformation distance* (*CND*) (note that this measure is not a true metric as it is not symmetric and does not obey the triangle inequality) $d(S, T)$ is the length of the shortest CNT from S to T. Note that if a pair S, T of CNPs has $s_i = 0$ but $t_i > 0$ for some i, then there is no CNT from S to T; we say that $d(S, T) = \infty$ in this case. The CND $d(S, T)$ between a pair of profiles can be computed in linear time (Zeira *et al.*, 2017).

2.2 Semi-ordered CNTs

Both the linear time algorithm (Zeira *et al.*, 2017) and the integer linear programming (ILP) algorithm (El-Kebir *et al.*, 2017) for computing the CND restrict to *ordered* CNTs, where all deletions come before all amplifications. Formally, let $E = (e_1, \ldots, e_l)$ be a CNT from S to T. Suppose, we partition E into maximal contiguous sequences of events of the same type. Thus, $E = (E_1, \ldots, E_k)$ where each *phase* E_j is a contiguous sequence in E, each E_j is composed of events of the same type $\tau(E_j)$, and no two consecutive subsequences are of the same type. In this case, we say that E is composed of k *phases* $E_1, \ldots, E_k$. Let $op(E_j, i) = |\{(\ell, h, \tau(E_j)) \in E_j | \ell \leq i \leq h\}|$ be the number of events of type $\tau(E_j)$ that affect the ith position in the profile in phase E_j. CNT E from S to T is *phase-bounded* provided

$op(E_j, i) \leq B$ for all $i \in \{1, \ldots, n\}$ and every phase E_j, where $B = \max(\max(S), \max(T))$ is the maximum copy number. Zeira *et al.* (2017) showed that for any pair S, T of CNPs with $d(S,T) < \infty$ there exists a shortest phase-bounded CNT $E = (E_-, E_+)$ with two phases: E_- having only deletions, and E_+ having only amplifications. A transformation of this form is called an *ordered* transformation.

As a shortest ordered CNT always exists, it is algorithmically sufficient to restrict attention to ordered CNTs in order to compute the CND. However, unordered CNTs may yield the same distances, and in some cases may be more biologically relevant. For example, the CNPs $S = (1,1,1,1,1)$ and $T = (2,2,1,2,2)$ have CND $d(S,T) = 2$ (Fig. 1). A shortest ordered transformation $E = ((1,2,+1),(4,5,+1))$ consists of an amplification of the first two positions followed by an amplification of the last two positions. On the other hand, the unordered transformation $E = ((1,5,+1),(3,3,-1))$ also has two events: an amplification of the entire chromosome followed by a deletion of the middle segment. As whole-chromosome duplications and deletions are common in cancer, the unordered transformation may be more plausible than the first transformation. Thus, restricting to ordered CNTs may preclude other optimal transformations that better explain the profiles.

We define a *semi-ordered* CNT from S to T as a CNT $E = (E_1, E_2, E_3)$ with three phases where E_1 and E_3 have only deletions, E_2 has only amplifications, and $E_1(S)_i = 0$ for all i where $t_i = 0$. In other words, a semi-ordered CNT has three phases: deletions, amplifications and deletions, and every zero position in T reaches zero after the first phase of deletions. While there is no specific biological rationale for restricting transformations to three phases, this restriction provides a richer space of transformations than ordered transformations while remaining computationally tractable (Fig. 1). In Supplementary Section S1.1, we show that while finding a shortest semi-ordered transformation between a pair of profiles can be written as an ILP formulation, the corresponding constraint matrix is totally unimodular (TUM). As a result, the ILP can be converted to a linear programming formulation (**LP S2**) without integrality constraints that is guaranteed to have integer optimum (Hoffman and Kruskal, 2010). In addition, we give a graph-theoretic characterization of the space of shortest CNTs that can be generated from a solution to the LP.

2.3 Weighted CNTs

In this section, we derive a weighted CNT model and describe an efficient algorithm to compute the weighted CNT distance.

The CND counts all events equally in the distance, regardless of the length or type of event. This is problematic for two reasons. First, it has been observed that CNAs of different lengths occur at different rates in cancer (Beroukhim *et al.*, 2010; Ciriello *et al.*, 2013; Zack *et al.*, 2013). Second, CNPs inferred from real data often have uncertainty, and this uncertainty is generally length dependent. For instance, consider the following pairs of CNPs: $S = (1,1,1,1,1,1)$, $T = (1,2,1,1,2,1)$, $E = ((2,2,+1),(5,5,+1))$ and $S' = (1,1,1,2,2,2)$, $T' = (2,2,2,1,1,1)$, $E' = ((4,7,-1),(1,4,+1))$. Both pairs of profiles have CND $d(S,T) = d(S',T') = 2$. However, the first transformation E includes two focal amplifications which might be less likely in cases where the CNPs have errors. The second transformation E' also has a distance of 2, but the events are chromosome arm gain and loss. While the CND gives both pairs the same distance, if there is uncertainty in the CNPs, then arguably S' and T' should be less similar than S and T.

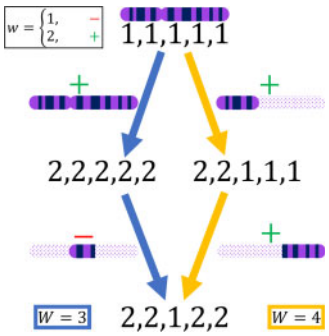


Fig. 1. Weighted and semi-ordered CNTs from $(1,1,1,1,1)$ to $(2,2,1,2,2)$. The right (yellow) CNT is ordered with two amplifications while the left (blue) CNT is semi-ordered and has one amplification and one deletion. Given a weight function that assigns a weight of 1 to deletions and 2 to amplifications, the left CNT has a weight of 3 while the right CNT has a weight of 4

To model differences in events, we introduce an *event weight function* $w : \{1 \ldots n\} \times \{1 \ldots n\} \times \{+,-\} \to \mathbb{R}^+$ that maps events on CNPs of length n to positive weights. Namely, $w(i,k,\tau)$ is the weight of an operation starting at position i and ending at position k of type τ. The event weight accounts for the position of the event along the chromosome, the length of the event and its type. For example, deletions and amplifications can have different weights depending the rate of these events in a specific cancer type. Longer events may have a higher weight than shorter ones, and the weight $w(1,n,+)$ of a whole-chromosome duplication may have a different weight regardless of the chromosome length.

We define the following weighted CNT model:

Weighted CNT model: Let S and T be CNPs, let E be a CNT from S to T and let w be a weight function for events. We define the *weight* $W(E) = \sum_{e \in E} w(e)$ of the CNT E to be the sum of weights of events in E.

The weighted CNT model distinguishes transformations based on their weight and not just the number of events. For example, there are two shortest CNTs from $S = (1,1,1,1,1)$ to $T = (2,2,1,2,2)$: (i) an amplification of the entire chromosome followed by a deletion of the middle segment, or (ii) an amplification of the first two positions followed by an amplification of the last two positions (Fig. 1). Suppose that the weight function is $w(i,j,+1) = 2$ and $w(i,j,-1) = 1$. Then, the weight of the first CNT is 3 whereas the weight of second CNT is 4.

The weighted CNT model can also be interpreted as a probability model. Suppose that each event $e = (i,k,\tau)$ occurs with probability p_e, and that events in a transformation E from S to T are independent. Then, the probability of observing a transformation E is $\prod_{e \in E} p_e$ and $\min_{E:E(S)=T}(-\sum_{e \in E} \log p_e)$ gives a maximum likelihood CNT between S and T. Therefore, setting the weight $w(e) = -\log p_e$ for each event e will make the weight of a transformation proportional to its likelihood.

The goal of the event weight function is to distinguish between CNTs. In Supplementary Section S1.1, we show how weights can be used to determine a shortest semi-ordered CNT consistent with a single solution to the LP (Supplementary Problem S2). However, there may be multiple optimal solutions to the LP. Moreover, while a shortest CNT has the minimum number of events, the true biological transformation need not be parsimonious. For instance, a shortest transformation of $(1,1,1,1,1,1)$ to $(1,1,1,2,2,2)$ involves one chromosome arm amplification. Yet, a biological explanation for gaining an arm could be first gaining a whole chromosome and then losing an arm. The next problem generalizes the CND by finding a minimum weight transformation between a pair of profiles.

Problem 1 (Minimum weight semi-ordered CNT). Given a source CNP S, a target CNP T and a weight function w, find semi-ordered phase-bounded CNT E having a minimum weight $W(E)$.

We now give our main result; an LP formulation to find a minimum weight semi-ordered CNT. Let x^j_{lk} be a variable indicating the number of events from position l to position k in phase j.

Minimum weight semi-ordered CNT (LP 1): $\min \sum_j \sum_{l \leq k} w(l,k,j) x^j_{lk}$ subject to

$$s_i \leq \sum_{l \leq i \leq k} x^1_{lk} 1 \leq i \leq n, \quad \text{if} \quad t_i = 0, \tag{1.1}$$

$$\sum_{l \leq i \leq k} x^1_{lk} \leq s_i - 1 1 \leq i \leq n, \quad \text{if} \quad t_i > 0, \tag{1.2}$$

$$s_i - \sum_{l \le i \le k} x^1_{lk} - x^2_{lk} + x^3_{lk} = t_i 1 \le i \le n, \quad \text{if} \quad t_i > 0, \tag{1.3}$$

$$\sum_{l \le i \le k} x^j_{lk} \le B 1 \le i \le n, \quad j \in \{1,2,3\}, \tag{1.4}$$

$$0 \le x^j_{lk} 1 \le l \le k \le n, \quad j \in \{1,2,3\}. \tag{1.5}$$

LP 1 has a quadratic number of variables and a linear number of constraints. Next, we show that **LP 1** yields a minimum weight transformation.

Theorem 1 *The constraint matrix of* ***LP 1*** *is totally unimodular. Thus,* ***LP 1*** *has an integer solution corresponding to a minimum weight semi-ordered CNT between a given pair S, T of CNPs and any weight function w.*

Note that as the minimal weight CNT problem does not find a shortest transformation, it may produce long transformations. Therefore, we may want to find a transformation that balances both its weight and its length. In Problem 2 we find a transformation that minimizes a linear combination of the weight and the length of the transformation, while in Problem 3 we find the minimum weight transformation only among the set of shortest transformations.

Problem 2 (Minimum regularized semi-ordered CNT). Given a source CNP S, a target CNP T, a weight function w and a non-negative number λ, find a phase-bounded semi-ordered CNT E that minimizes $W(E) + \lambda|E|$.

Problem 3 (Minimum weight shortest semi-ordered CNT). Given a source CNP S, a target CNP T and a weight function w, find a shortest phase-bounded semi-ordered CNT E having a minimum weight, $\min_{E:|E|=d(S,T)} W(E)$.

We show that **LP 1** solves both Problems 2 and 3. First, for Problem 1, let E be a semi-ordered CNT from S to T and denote by x^j_{lk} the number of events in E from position l to k in phase j. The objective $W(E) + \lambda|E|$ of Problem 2 can be written as $\sum_j \sum_{l \le k} w(l,k,j) x^j_{lk} + \lambda \sum_j \sum_{l \le k} x^j_{lk}$. Hence, Problem 2 is equivalent to Problem 1 with modified weights $w'(l,k,j) = w(l,k,j) + \lambda$ and is solved with **LP 1**.

We reduce Problem 3 to Problem 2 with $\lambda = B \sum_j \sum_{l \le k} w(l,k,j)$. As $W(E) \le \lambda$ for any phase-bounded CNT E, in order to minimize $W(E) + \lambda|E|$, the length $|E|$ of the CNT must be minimized first and only then the weight $W(E)$ of the CNT should be minimized. We note that Problem 3 can also be solved directly by modifying **LP 1** with a constraint $\sum_j \sum_{l \le k} x^j_{lk} \le d(S,T)$, where $d(S, T)$ is the CND from S to T. Though by adding this constraint, the constraint matrix of the modified formulation would not be TUM, the first approach shows that that Problem 3 has integer optimum regardless.

3 Results

We present three applications of the weighted CND. First, we show on simulated data that the weighted CND provides better estimates of the evolutionary distance between CNPs that evolve with non-uniform length distribution compared to the unweighted CND and the Euclidean distance (Section 3.1). Next, we show how the weighted CND helps recover cell populations in tumors using noisy CNPs derived from low-coverage single-cell DNA sequencing data (Section 3.2). Finally, we use the weighted CND to estimate CNA rates on TCGA data (Section 3.3).

3.1 Reconstruction of simulated copy number trees

In this section, we compare distance-based tree reconstruction on simulated CNPs using three distances: Euclidean distance, unweighted CND and weighted CND. We simulate CNPs from a directed tree via copy number events that occur with different probabilities. We assume that the tree is unknown but the probability distribution of events is known. We further assume that all profiles, including inner nodes are used to reconstruct the evolutionary tree and estimate the events along the edges. Obviously, these assumptions do not hold in real data; rather, the goal of these simulations is to show that using the weighted CND with prior knowledge of the distribution of events gives better estimates of distance than the other distance measures.

In this setting, a minimum spanning arborescence (MSA) (Edmonds, 1967) of the simulated profiles corresponds to a maximum parsimony tree when using unweighted CND between nodes. Conversely, we define the likelihood of a tree as the product of all transformation probabilities along its edges. Therefore, if we set the weight of an event e as $-\log(p(e))$, where $p(e)$ is the probability of e in the simulation, then an MSA corresponds to a maximum likelihood tree when using the weighted CND.

We generate a rooted, directed binary tree T with n nodes and a designated root node r having a CNP of length m with all entries having the same value b. The length of each edge in T, corresponding to the number of events between nodes, is $1 + X$, where X is drawn from a Poisson(λ) distribution. Thus, each edge has a minimum of one event and an average of $1 + \lambda$ events. Each event is an amplification with probability p and a deletion with probability $1 - p$. For each event, we draw the length l of an event from a distribution having $\Pr(l) \propto e^{-\beta \frac{l}{m}}$. The position an event acts along the genome is selected uniformly among $m - l + 1$ possible positions. Profiles in the tree are simulated from the root downwards. Throughout the simulations we fix $n = 61$, $\lambda = 1$, $b = 2$, $m = 50$, $p = 0.6$. We tested different distribution of event lengths $\beta \in \{10, 5, 1, -1, -5, -10\}$ (Supplementary Fig. S2). For each β, 50 trees and corresponding profiles are simulated.

Let $\mathcal{G}$ be the set of simulated profiles in the tree T. Given a distance measure d, we calculate a pairwise distance matrix D. Note that D is not necessarily symmetric as both the weighted CND and unweighted CND are not symmetric. Moreover, for some ordered pairs of profiles there may not exist a transformation between them and in this case the distance in undefined. We build a directed weighted graph $G = (V = \{1 \ldots n\}, E = \{(u, v, D(u, v)) | \forall u, v \in V, D(u, v) < \infty\})$, i.e. G has a directed edge from u to v of weight $D(u, v)$ if the distance from u to v exists. To make a fair comparison, when building the graph based on the Euclidean distance, we remove edges where the CND will not exist. Finally, we find a minimum weight spanning arborescence $\hat{T}$ rooted at r in G (Edmonds, 1967).

We evaluate the difference between the inferred tree $\hat{T} = (V, E_{\hat{T}})$ and the true tree $T = (V, E_T)$ using two measures. First, we define the *true positive edge rate* $P(\hat{T}, T) = \frac{E_{\hat{T}} \cap E_T}{n-1}$ as the fraction of edges common to both trees. Second, we calculate the difference $\Delta(\hat{T}, T) = N(T) - N(\hat{T})$ between $N(T)$, the total number of events along the edges of T, and $N(\hat{T})$, the corresponding quantity for $\hat{T}$. $\Delta(\hat{T}, T)$ quantifies how well the inferred tree recapitulates the events of the simulated tree. As events can overlap and cancel one another over time, a tree that correctly captures the events ($\Delta(\hat{T}, T) \approx 0$) is a better estimation of the true evolution.

We find that both the unweighted CND (CND) and weighted CND (WCND) outperform the Euclidean distance (EUC) in reconstructing the true tree across all values of β (Fig. 2a). In addition, the weighted CND shows significant improvement over the unweighted CND ($p \le 3 \times 10^{-5}$ in paired t-test) over all values of β. The average $P(\hat{T}, T)$ improvement increases from 0.03 when $\beta = 10$ to 0.11 when $\beta = -10$. Smaller values of β correspond to distributions where longer events are more likely than short ones (Supplementary Fig. S2). In this case, there is a higher probability for events to overlap, creating by chance similar profiles on different branches of the tree. As there are more similar profiles, it is more difficult to correctly recover the true tree topology. We indeed see that the reconstruction performance improves in all methods when β increases. The weighted

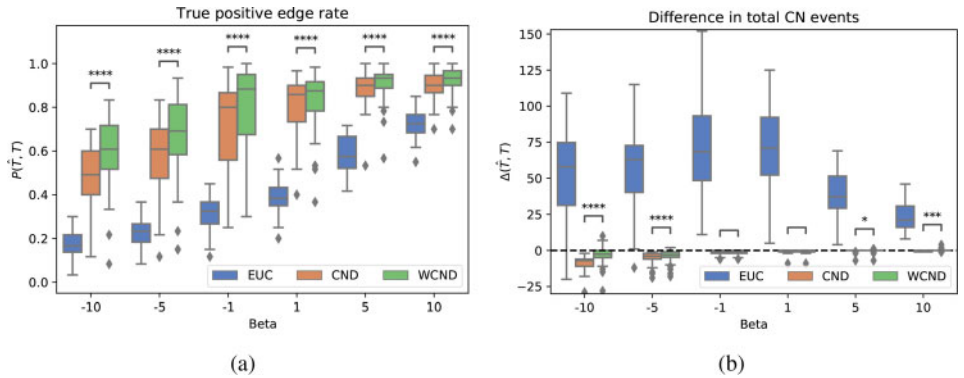


Fig. 2. Comparison of trees constructed using Euclidean distance (EUC), unweighted CND (CND) and weighted CND (WCND) on simulated profiles with length-based distribution of CNAs. (**a**) The proportion $P(\hat{T}, T)$ of edges common to the simulated tree and the tree inferred by the MSA. (**b**) The difference $\Delta(\hat{T}, T)$ in the total number of events between the simulated and the tree inferred by the MSA. *P*-value in paired *t*-test: * ≤ 0.05, ** $\leq 10^{-1}$, *** $\leq 10^{-3}$, **** $\leq 10^{-4}$

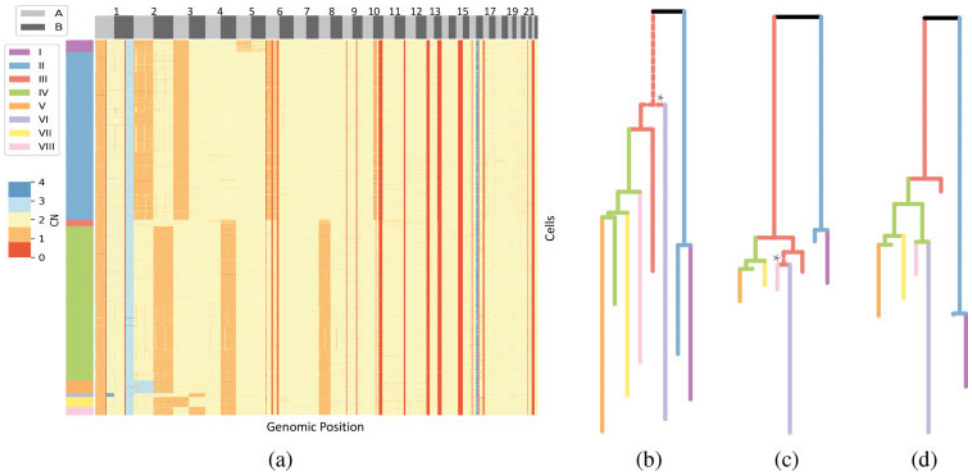


Fig. 3. (**a**) Haplotype-specific CNPs obtained from whole-genome single-cell sequencing of a breast tumor. Cells were previously clustered into eight clones I–VIII (Zaccaria and Raphael, 2019). Copy numbers were limited to 4 for simpler presentation. Trees constructed using neighbor joining on clone consensus CNPs using (**b**) Euclidean distance, (**c**) unweighted CND and (**d**) weighted CND. Dashed branches marked with an asterisk indicate differences from previous CHISEL (Zaccaria and Raphael, 2019) analysis

CND shows most improvement exactly in those hard cases ($\beta < 0$).

We also find that the weighted CND infers trees with $\Delta(\hat{T}, T) \approx 0$ for all values of β (Fig. 2b). In contrast, the Euclidean distance yields trees that grossly overestimate the true number of events in the tree ($\Delta(\hat{T}, T) \gg 0$), while the unweighted CND yields trees that underestimate the true number of events in the tree ($\Delta(\hat{T}, T) < 0$). The latter is not surprising as the unweighted CND minimizes the total number of events in the tree. Although for $\beta \in \{-1, 1\}$, both the weighted and unweighted CND produce trees with similar number of events, the weighted CND has higher $P(\hat{T}, T)$. This shows that the weighted CND is able to recover the tree topology among multiple trees with the same total number of events along the edges. For other β, the weighted CND produces trees with total number of events closer to the true tree, having an average of 0.36–3.78 more events than the unweighted CND.

As in real data the true distribution of events is unknown, we repeated the simulations using different probabilities for tree inference with the weighted CND. Specifically, we simulated trees with $\beta \in \{10, 5, -5, -10\}$ except we assigned the weight of an event of length l from a distribution with $\Pr(l) \propto e^{-\text{sign}(\beta)\frac{l}{m}}$. We find that even with an incorrect weight distribution, the WCND still infers trees that are significantly better than the unweighted CND (Supplementary Fig. S3a). Thus, having prior knowledge of the event distribution is still superior to using an unweighted CND. On the other hand, both weighted and unweighted CND find trees with the exact minimum number of events (Supplementary Fig. S3b).

3.2 Single-cell cancer sequencing data

We analyzed CNPs derived from ultra-low-coverage (0.02× to 0.05×) single-cell whole-genome sequencing data of a breast tumor obtained using the 10X CNV Solution (10X Genomics, 2019a). The dataset includes five tumor sections, each comprising ≈2k cells. CHISEL (Zaccaria and Raphael, 2019) was used to infer a haplotype-specific CNP (separating integer copy numbers of the two homologous chromosomes) for each cell in 5-MB bins across all autosomes. Thus, each cell has 44 CNPs corresponding to the 22 pairs of chromosomes. We analyzed cells jointly across all five sections, focusing on cells that CHISEL classified as tumor cells (Zaccaria and Raphael, 2019); we excluded centromere bins that had highly variable calls and discarded duplicate cells with identical CNPs, resulting in a final dataset of 4012 cells by 1052 CN entries. We arbitrarily designated one allele as A and the other as B (Fig. 3a).

Due to the ultra-low coverage, copy number calls in individual cells are prone to errors. While whole-chromosome and arm-level CNs can be more reliably called, focal (single bin) changes in the CNPs are more likely to be errors. The CHISEL analysis (Zaccaria and Raphael, 2019) addressed this issue by clustering cells using Hamming distance and creating consensus CNPs for each cluster of cells. This procedure resulted in eight groups of cells, or *clones*, labeled I–VIII and characterized by different large-scale CNAs, including whole-chromosome and chromosome–arm level events and an early whole-genome duplication (Fig. 3a and Supplementary Fig. S4). This analysis also built a phylogeny relating these clones with two main branches: one branch containing clones I and II harboring deletions of chromosome 2A and 3B, and arm deletions of

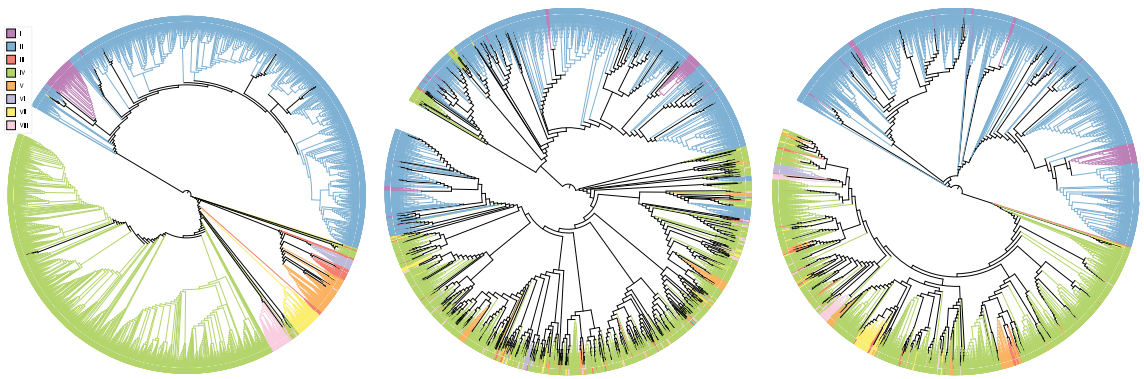
6A.p and 10B.p; the other branch containing clones III–VIII harboring deletions of chromosomes 4B and 8A, and also a deletion of chromosome 2B in most of the clones in this lineage. The CHISEL tree analysis also suggested that clone III is the ancestor of clone IV which is the ancestor of clones V–VIII.

As the individual cell CNPs are noisy, we first analyzed the clone CNPs. Similar to CHISEL, we created consensus CNPs for the cells of each clone (Supplementary Fig. S4). While cells in the dataset had on average 81 break points, i.e. positions where consecutive copy numbers do not match, clone profiles had only 39 breakpoints on average. We then calculated a symmetric distance matrix between clones using Euclidean distance, unweighted CND and weighted CND. Because both CNDs are not symmetric, we defined the symmetric distance between a pair of profiles as the average of the distances in both directions. We constructed a phylogenetic tree of the clones using neighbor joining. As the major clonal events in this sample are large chromosomal aberrations, we gave whole-chromosome and arm level events a high weight corresponding to $(-\log)$ a probability of 10^{-7}. On the other hand, we suspect that focal changes are more likely to be errors, and thus assigned these events a weight corresponding to a probability of 0.9. We gave amplifications and deletions equal probabilities. While all distances are able to separate the two main lineage of the tumor, only the weighted CND produces a tree concordant with previous analysis (Fig. 3b–d). Notably, the Euclidean distance mistakenly places clones III and VI close together on the tree because the events that distinguish these two clones are shorter relative to other chromosomal events (1A.q gain and 3B loss). Similarly, the unweighted CND also misplaces a branch, placing clones VI and VIII as descendants of clone III. This placement would imply that chromosome 2B was lost twice independently in two separate branches, which is unlikely since most cells on this lineage contain this mutation. This change is caused by a single bin in chromosome 1A.p having a different copy number than the rest of the arm. This change is shared by clones III and VIII making them one event closer in terms of the unweighted CND. On the other hand, the weighted CND gives this change a low weight in comparison to chromosome/arm level events, thus correctly resolving the clone topology.

To show that the weighted CND is able to cope with errors, we next computed phylogenies on the 4012 *individual* cells using neighbor joining. The tree computed using Euclidean distance (Fig. 4a) has clades that largely recapitulate the clone assignment given in Zaccaria and Raphael (2019), which is not surprising because (i) Hamming distance was used to cluster cells into clones in CHISEL and (ii) whole-chromosome and arm level events are the major events in this tumor and since Euclidean distance weighs events proportional to their lengths, it is more robust to small changes in the profiles. However, we find that the Euclidean distance tree has strange evolutionary relationships placing clones III, V and VI together in the same clade although previous analysis (Zaccaria and Raphael, 2019) suggested that clones V and VI descended from clone IV. This arrangement would again imply that chromosome 2B was lost twice independently, which is not likely. The unusual placements are not surprising, because Euclidean distance has no underlying evolutionary model for CNAs. In contrast, the tree based on the unweighted CND (Fig. 4b) mixes cells from different clones in the same clade and even cannot separate the two main lineages (I–II and III–IIIV) that are the most distinct in their CNPs. This is likely because the unweighted CND is susceptible to noise in the CNPs in individual cells, and small differences in CNPs are weighted equally as large events. Finally, the tree inferred using the weighted CND clearly divides the two main lineages, preserves most of the clonal structure and recapitulates the evolutionary relationships from the clone tree in Zaccaria and Raphael (2019). Importantly, the weighted CND tree shows that clones V–VIII descend from clone IV, which is the more reasonable as these cells are distinguished by a 2B loss. The weighted CND tree does group some cells from different clones into the same clade. Closer inspection of cells that were reported to be from the same clone in Zaccaria and Raphael, (2019) but were split in the weighted CND tree sheds light on smaller sub-populations of cells. For instance, a small group of 13 cells that were classified as clone VIII but separated from the rest of the cells from this clone have one fewer copy of the p arm of chromosome 16B compared to the other cells in this clone. As we use higher weights for arm events, these cells were split from the rest of the clone. Similarly, a group of 14 cells of clone V containing a loss of chromosome 11 were separated from the rest of the cells in this clone.

As the true weights of events are unknown, we explored how the choice of weights in the weighted CND affects the phylogeny. We restricted our analysis to a single tumor section (E) with 1062 cells and calculated cell pairwise distance matrices using weighted CND with various weights. As the space of possible event weights is large, we reduced the weight function to three parameters: amplification/deletion probability P, local/chromosome event probability q and focal/segmental probability r. Therefore, the probability of a segmental deletion, for example, would be $(1-p)q(1-r)$. Each of these parameters was assigned either 0.25 or 0.75 resulting in a total of eight parameter configurations. In addition, we also considered another parameter combination $p = 0.5, q = 1-10^{-6}, r = 0.8$ to reflect the fact that chromosomal events are the major events in this tumor's evolution.

We find that varying the weights yields neighbor-joining trees that are quite different from one another, showing that the selection of weights can have a considerable effect of the resulting trees (Supplementary Fig. S6). We see that when perturbing only the probability p of amplifications while keeping other parameters fixed, the resulting trees seem a lot more similar. This can be explained by the fact that when symmetrizing the distance, some of the information on the direction of events is lost. Interestingly, weighted CND trees are more similar to the unweighted tree when



(a) Euclidean distance (b) Unweighted CND (c) Weighted CND

Fig. 4. Trees built on single-cell CNPs for different distance measures (a-c) visualized using iTOL (Letunic and Bork, 2019) where edge lengths are scaled for visualization and not proportional to the distance

the chromosome/arm level events and focal events are more likely ($q = 0.25, r = 0.75$) and least similar when segmental events are more likely ($q = 0.75, r = 0.25$). Conversely, the tree is more similar to the Euclidean tree when focal events have lower weights ($q = 0.75, r = 0.75$). This corroborate with our expectation as Euclidean distance gives a higher weight to longer events. Finally, the additional parameter combination that assigns a high weight to chromosome event gives a tree that is the most similar to the Euclidean distance tree than the other parameter combinations, as expected.

3.3 Estimating CNA rates in TCGA pan-cancer data

A key challenge in applying the weighted CND to real data is how to select biologically realistic weights. Estimating CNA rates is substantially more difficult than estimating SNV rates: as CNA overlap with each other, it is difficult to conclude which CNAs have occurred from pairs of moderately diverged CNPs. In this section, we illustrate how the weighted CNT can help in the inference of CNA rates using the inferred transformations between pairs of CNPs. We apply this procedure to CNPs from 26 cancer types from TCGA. It has been observed that different cancer types have different rates of CNAs (Ciriello *et al.*, 2013; Zack *et al.*, 2013), and even within a cancer type, different chromosomes show varying patterns of aneuploidy (Taylor *et al.*, 2018).

We infer CNA rates from a collection of CNPs using the following model. Let $D = \{(S_1, T_1), \ldots, (S_n, T_n)\}$ be a set of independent pairs of source and target profiles. Our goal is to find a probability distribution P over CNAs that will maximize the likelihood of observing the set of CNP pairs, namely $\max_P \sum_i \Pr(S_i \to T_i|P)$. Here, we assume that there are m classes of events $C_1, \ldots, C_m$ with unknown probabilities $p_1, \ldots, p_m$. Event classes may be determined by length, genomic location, or type of event: e.g. whole-chromosome amplifications, whole p-arm deletions, local q-arm amplifications, etc.

To find the most likely CNA probability distribution P, we use an EM-like approach. Previous analysis of CNA used a similar approach based on a heuristic for deconstruction of CNPs to identify segmental events (Mermel *et al.*, 2011; Zack *et al.*, 2013). We start with a random probability distribution $P^{(0)}$ over CNAs. At each iteration t, we find the most likely transformation $E_i = \text{argmax}_{E(S_i)=T_i} \Pr(S_i \to T_i|E, P^{(t-1)})$ for each pair i of profiles given the probabilities $P^{(t-1)}$ at the previous iteration. Then we re-estimate the event probabilities $p_j^{(t)} = \frac{\sum_i |\{e: e \in E_i, e \in C_j\}|}{\sum_i |E_i|}$ by the proportion of events in each class j in the entire cohort. We continue until convergence or a predefined number of iterations.

We obtained total CNPs and whole-genome doubling statuses for 10180 samples from the 26 cancer types from the TCGA pan-cancer dataset that had at least 100 samples (Taylor *et al.*, 2018). For each cancer type and each chromosome, we created a set $\{(S_i, T_i)\}$ of CNP pairs where T_i is the CNP of the ith sample and $S_i = b_i, \ldots, b_i$ is a profile with the same CN b_i across segments, where $b_i = 2$, 4 or 6 depending on the sample's genome doubling status. For each such cohort, we estimated CNA rates considering the following classes of events. As it has been noted that CNA tend to be localized on each chromosome arm, we classify events based on their start and end points in relation to the centromer (p-arm, q-arm, cross). An event is classified as affecting the whole chromosome/arm if its length consists of at least 80% of the chromosome/arm length (Taylor *et al.*, 2018). Finally, each event is either an amplification or a deletion giving 12 classes of events overall.

We limited our analysis to the 17 non-acrocentric chromosomes resulting in 442 distributions over 12 CNA classes (Supplementary Fig. S7). We clustered the inferred CNA rates into eight groups with k-means using the Silhouette method to determine the number of clusters. The cluster centroids represent CNA rate signatures common to cancers and chromosomes (Fig. 5). The different signatures show variability in CNA rates. For instance, in Cluster 1 about 30% of events are whole-chromosome deletions while in Cluster 4 about 30% of events are chromosome amplifications. Similarly, Cluster 0 shows more deletions on the q-arm and amplifications on the p-arm while Cluster 6 shows more deletions on the p-arm and amplifications on the q-arm. Cluster 2 has around 50% focal amplifications on the q-arm whereas Cluster 3 has around 50% focal deletions on the q-arm.

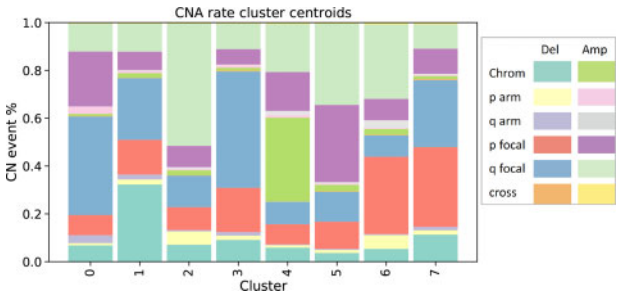


Fig. 5. Inferred CNA rates in TCGA pan-cancer CN data. Colors correspond to 12 classes of events: {deletions(−), amplifications(+)} × {whole chromosome, whole p-arm, whole q-arm, focal p-arm, focal q-arm, crossing centromere}

Table 1. Significantly enriched chromosome and cancer type in each CNA rate cluster

Cluster	Chromosome	Cancer
0	5 (1.22e-6)	
1	18 (5.6e-5)	TGCT (4.04e-7), KIRC (6.36e-5)
2	17 (4.6e-22)	
3	4 (3.77e-6)	
5		LAML (1.25e-10)
6	1 (3.99e-7), 3 (3.21e-6), 8 (3.99e-7)	
7	6 (1.3e-5), 9 (1.3e-5)	

Note: Hyper-geometric P-value is presented in parentheses. The P-values were thresholded with Bonferroni correction for multiple testing.

We tested whether the clusters were enriched for certain cancer types or chromosomes (Table 1). With the exceptions of Cluster 1 having a high number chromosomes from TGCT and KIRC, and Cluster 5 having a high number of LAML chromosomes, clusters were not enriched to specific cancer types. On the other hand, almost half of the chromosomes were enriched in some cluster. Notably, 24 (out of 26) chromosome 17 CNA rates were included in Cluster 2 suggesting that the distribution of events on chromosome 17 is consistent across cancer types. Cluster 6, having a high proportion of deletions on the p-arm and amplifications on the q-arm, was enriched with Chromosomes 1, 3 and 8. Chromosome 3 p-arm loss and q-arm gain have been shown to be a dominant feature of squamous cell carcinomas (Taylor *et al.*, 2018). Indeed, we see that HNSC, LUSC, CESC have similar CNA rates characterized by 3p-arm loss and 3q-arm gain (Supplementary Fig. S8).

4 Discussion

In this paper, we introduce a weighted CND. The weighted CND allows for segmental events to have different weights, or probabilities, based on type, length and location, enabling more detailed studies of the rates of copy number events compared to the unweighted CND that is currently in use. We give the first efficient polynomial-time algorithm to compute the weighted CND. This algorithm is based on the observation that computation of the weighted CND is an optimization problem with totally unimodular constraint matrix. In addition to computing the minimum weighted distance, the algorithm also provides an explicit transformation that achieves this distance. We illustrate the utility of the weighted CND in three different applications: distance-based phylogenetic tree reconstruction, analysis of noisy CNPs from single-cell DNA

sequencing data and estimation of CNA rates. We show on simulated data that the weighted CND outperforms Euclidean distance and unweighted CND in inferring ancestral relations between profiles when events are generated with different rates. We analyze CNPs from single-cell DNA sequencing of a breast tumor and show that the weighted CND overcomes errors in copy number calls and improves tree reconstruction. Finally, we use the weighted CND to infer CNA rate signatures from the TCGA pan-cancer data-set.

An important question in applications of the weighted CND is how to determine the weights. We showed that it is possible to estimate CNA rates from cancer cohorts, but there is substantial room improving this process. First, larger cohorts are needed to estimate the distribution of sub-arm focal events as a function of their length and/or position along the chromosome. Second, while we assumed that samples from the same tumor type share the similar CNA rate distribution, there may actually be different mutagenic processes that affect the rate of CNAs in different samples (Alexandrov *et al.*, 2020; Macintyre *et al.*, 2018; Shah, 2018). There may be multiple CNA rate distributions both within a cohort and even within a single sample. Third, CNPs from bulk tumors are a mixture of multiple cells with potentially different CNPs. Thus, one must rely on accurate deconvolution of bulk samples into integer copy numbers (Gerstung *et al.*, 2020; Salcedo *et al.*, 2020; Zaccaria and Raphael, 2018) or obtain larger cohorts of single-cell sequencing data. Finally, an alternative approach for finding event weights is by examining the relative least-squares fit of the distances to the tree [e.g. using the Fitch–Margoliash method (Fitch and Margoliash, 1967)] for different weight parameters.

The probabilistic model used in the weighted CND can be extended in several ways. First, the model assumes that events are independent given a transformation. This may not be the case in general as the probability of events may depend on previous events. Second, the model does not directly estimate errors in the CNPs. Both a probabilistic model and a simulation framework that accurately model real events and errors are important directions for future work. In addition, the number of CNAs and not just their relative proportions should be modeled, especially because some cancers are characterized by a higher number of CNAs (Ciriello *et al.*, 2013). Finally, extending the model to more complex CNAs such as whole-genome duplications (Bielski *et al.*, 2018) or breakage–fusion–bridge cycles (Zakov *et al.*, 2013) remains open.

Acknowledgement

We thank Simone Zaccaria for his help with the single-cell CN data.

Funding

This work was supported by the US National Institutes of Health (NIH) [U24CA211000]; US National Science Foundation (NSF) CAREER Award [CCF-1053753]; O'Brien Family Fund for Health Research; Wilke Family Fund for Innovation; and Chan Zuckerberg Initiative DAF [2018-182608 to B.J.R.].

Conflict of Interest: B.J.R. is a founder of Medley Genomics and a member of its board of directors.

References

10X Genomics (2019a). Assessing tumor heterogeneity with single cell CNV. https://www.10xgenomics.com/solutions/single-cell-cnv.

10X Genomics (2019b). What is cell ranger DNA? https://support.10xgenomics.com/single-cell-dna/software/pipelines/latest/what-is-cell-ranger-dna.

Alexandrov,L.B. *et al.*; PCAWG Mutational Signatures Working Group (2020) The repertoire of mutational signatures in human cancer. *Nature*, **578**, 94–101.

Andor,N. *et al.* (2018) Joint single cell DNA-seq and RNA-seq of gastric cancer cell lines reveals rules of *in vitro* evolution. *NAR Genomics and Bioinformatics*, Volume 2, Issue 2, June 2020, lqaa016.

Bader,M. and Ohlebusch,E. (2007) Sorting by weighted reversals, transpositions, and inverted transpositions. *J. Comput. Biol.*, **14**, 615–636.

Beerenwinkel,N. *et al.* (2015) Cancer evolution: mathematical models and computational inference. *Syst. Biol.*, **64**, e1–e25.

Beroukhim,R. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.

Bielski,C.M. *et al.* (2018) Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat. Genet.*, **50**, 1189–1195.

Burrell,R.A. *et al.* (2013) The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, **501**, 338–345.

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

Chowdhury,S.A. *et al.* (2014) Algorithms to model single gene, single chromosome, and whole genome copy number changes jointly in tumor phylogenetics. *PLoS Comput. Biol.*, **10**, e1003740.

Chowdhury,S.A. *et al.* (2015) Inferring models of multiscale copy number evolution for single-tumor phylogenetics. *Bioinformatics*, **31**, i258–i267.

Ciriello,G. *et al.* (2013) Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.*, **45**, 1127–1133.

Cordonnier,G. and Lafond,M. (2020) Comparing copy-number profiles under multi-copy amplifications and deletions. *BMC Genomics,* **21**, 198.

Edmonds,J. (1967) Optimum branchings. *J. Res. Natl Bur. Stand. B*, **71B**, 233–240.

El-Kebir,M. *et al.* (2017) Complexity and algorithms for copy-number evolution problems. *Algorithms Mol. Biol.*, **12**, 13.

Fertin,G. *et al.* (2009) *Combinatorics of Genome Rearrangements*. Cambridge, Massachusetts, MIT Press.

Fischer,A. *et al.* (2014) High-definition reconstruction of clonal composition in cancer. *Cell Rep.*, **7**, 1740–1752.

Fisher,R. *et al.* (2013) Cancer heterogeneity: implications for targeted therapeutics. *Br. J. Cancer*, **108**, 479–485.

Fitch,W.M. and Margoliash,E. (1967) Construction of phylogenetic trees. *Science*, **155**, 279–284.

Garvin,T. *et al.* (2015) Interactive analysis and assessment of single-cell copy--number variations. *Nat. Methods*, **12**, 1058–1060.

Gawad,C. *et al.* (2016) Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.*, **17**, 175–188.

Gerstung,M. *et al.*; PCAWG Evolution & Heterogeneity Working Group (2020) The evolutionary history of 2,658 cancers. *Nature*, **578**, 122–128.

Ha,G. *et al.* (2014) Titan: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.*, **24**, 1881–1893.

Hannenhalli,S. and Pevzner,P.A. (1995a) Transforming cabbage into turnip. In: *Proceedings of Annual ACM Symposium on the Theory of Computing*, Vol. **46**, pp. 178–189, New York, NY, USA.

Hannenhalli,S. and Pevzner,P.A. (1995b) Transforming men into mice (polynomial algorithm for genomic distance problem). In: *Proceedings of IEEE Symposium on Foundations of Computer Science*, Vol. **36**, pp. 581–592.

Hoffman,A.J. and Kruskal,J.B. (2010) Integral boundary points of convex polyhedra. In: *50 Years of Integer Programming 1958-2008*. Springer: Heidelberg, Berlin, pp. 49–76.

Jamal-Hanjani,M. *et al.* (2017) Tracking the evolution of non–small-cell lung cancer. *N. Engl. J. Med.*, **376**, 2109–2121.

Laks,E. *et al.* (2018) Resource: Scalable whole genome sequencing of 40,000 single cells identifies stochastic aneuploidies, genome replication states and clonal repertoires. bioRxiv, p. 411058.

Letunic,I. and Bork,P. (2019) Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.*, **47**, W256–W259.

Li,Y. *et al.* (2016) Allele-specific quantification of structural variations in cancer genomes. *Cell Syst.*, **3**, 21–34.

Macintyre,G. *et al.* (2018) Copy number signatures and mutational processes in ovarian carcinoma. *Nat. Genet.*, **50**, 1262–1270.

Mangiola,S. *et al.* (2016) Comparing nodal versus bony metastatic spread using tumour phylogenies. *Sci. Rep.*, **6**, 33918.

McGranahan,N. and Swanton,C. (2015) Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell*, **27**, 15–26.

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

McPherson,A.W. *et al.* (2017) Remixt: clone-specific genomic structure estimation in cancer. *Genome Biol.*, **18**, 140.

Mermel,C.H. *et al.* (2011) Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.*, **12**, R41.

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

Nik-Zainal,S. *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. *et al*. (2012) Reconstructing cancer genomes from paired-end sequencing data. *BMC Bioinformatics*, **13**, S10.

Oesper,L. *et al*. (2013) Theta: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol*., **14**, R80.

Pennington,G. *et al*. (2007) Reconstructing tumor phylogenies from heterogeneous single-cell data. *J. Bioinform. Comput. Biol*., **05**, 407–427.

Pinter,R.Y. and Skiena,S. (2002) Genomic sorting with length-weighted reversals. *Genome Inform*., **13**, 103–111.

Salcedo,A. *et al*.; DREAM SMC-Het Participants (2020) A community effort to create standards for evaluating tumor subclonal reconstruction. *Nat. Biotechnol*., **38**, 97–107.

Schwartz,R. (2019) Computational models for cancer phylogenetics. In: *Bioinformatics and Phylogenetics*. Springer: Heidelberg, Berlin, pp. 243–275.

Schwarz,R.F. *et al*. (2014) Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Comput. Biol*., **10**, e1003535.

Schwarz,R.F. *et al*. (2015) Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. *PLoS Med*., **12**, e1001789.

Shah,S.P. (2018) Copy number signatures in ovarian cancer. *Nat. Genet*., **50**, 1208–1209.

Shen,R. and Seshan,V.E. (2016) Facets: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res*., **44**, e131.

Sottoriva,A. *et al*. (2015) A big bang model of human colorectal tumor growth. *Nat. Genet*., **47**, 209–216.

Taylor,A.M. *et al*. (2018) Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell*, **33**, 676–689.

Wang,Y. *et al*. (2014) Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, **512**, 155–160.

Zaccaria,S. and Raphael,B.J. (2018) Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. bioRxiv.

Zaccaria,S. and Raphael,B.J. (2019) Characterizing the allele- and haplotype-specific copy number landscape of cancer genomes at single-cell resolution with chisel. bioRxiv.

Zaccaria,S. *et al*. (2018) Phylogenetic copy-number factorization of multiple tumor samples. *J. Comput. Biol*., **25**, 689–708.

Zack,T.I. *et al*. (2013) Pan-cancer patterns of somatic copy number alteration. *Nat. Genet*., **45**, 1134–1140.

Zahn,H. *et al*. (2017) Scalable whole-genome single-cell library preparation without preamplification. *Nat. Methods*, **14**, 167–173.

Zakov,S. *et al*. (2013) An algorithmic approach for breakage-fusion-bridge detection in tumor genomes. *Proc. Natl. Acad. Sci. USA*, **110**, 5546–5551.

Zeira,R. *et al*. (2017) A linear-time algorithm for the copy number transformation problem. *J. Comput. Biol*., **24**, 1179–1194.