A zero-agnostic model for copy number evolution in cancer

Henri Schmidt[1], Palash Sashittal[1], and Benjamin J. Raphael[1,†]

[1]Department of Computer Science, Princeton University, NJ, USA
[†]Correspondence: braphael@princeton.edu

Abstract

Motivation: New low-coverage single-cell DNA sequencing technologies enable the measurement of copy number profiles from thousands of individual cells within tumors. From this data, one can infer the evolutionary history of the tumor by modeling transformations of the genome via copy number aberrations. A widely used model to infer such *copy number phylogenies* is the *copy number transformation* (CNT) model in which a genome is represented by an integer vector and a copy number aberration is an event that either increases or decreases the number of copies of a contiguous segment of the genome. The CNT distance between a pair of copy number profiles is the minimum number of events required to transform one profile to another. While this distance can be computed efficiently, no efficient algorithm has been developed to find the most parsimonious phylogeny under the CNT model.

Results: We introduce the *zero-agnostic copy number transformation* (ZCNT) model, a simplification of the CNT model that allows the amplification or deletion of regions with zero copies. We derive a closed form expression for the ZCNT distance between two copy number profiles and show that, unlike the CNT distance, the ZCNT distance forms a metric. We leverage the closed-form expression for the ZCNT distance and an alternative characterization of copy number profiles to derive polynomial time algorithms for two natural relaxations of the small parsimony problem on copy number profiles. While the alteration of zero copy number regions allowed under the ZCNT model is not biologically realistic, we show on both simulated and real datasets that the ZCNT distance is a close approximation to the CNT distance. Extending our polynomial time algorithm for the ZCNT small parsimony problem, we develop an algorithm, *Lazac*, for solving the large parsimony problem on copy number profiles. We demonstrate that *Lazac* outperforms existing methods for inferring copy number phylogenies on both simulated and real data.

Availability: *Lazac* is implemented in C++17 and is freely available at: github.com/raphael-group/lazac-copy-number.

1 Introduction

Tumor evolution is characterized by both small and large genomic alterations that alter the fitness of cancer cells [31]. *Copy number aberrations*, i.e. modifications to the number of copies of a genomic segment, are an important and frequent sub-class of such alterations that drive prognostic and metastatic outcomes [1]. Deriving the evolutionary history of copy number aberrations, herein referred to as *copy number phylogenies*, is thus important for understanding the emergence of primary tumors and the development of subpopulations of cells that evade treatment and/or metastasize to other anatomical sites.

Recent technological and computational improvements in single-cell sequencing have enabled the mapping of high resolution copy number profiles in single cells. For example, the high-throughput 10x Genomics Single-cell Copy Number Variation solution [2, 48] produces ultra-low coverage ($< 0.05\times$) whole genome sequencing data from $\approx$ 2000 individual cells. Other recent technologies, including DLP/DLP+ [49, 23, 15] and ACT [28], produce similar data. Multiple computational methods [48, 45, 44, 10, 20, 24] have been introduced to infer high resolution *copy number profiles*, integer vectors that contain the number of copies of each genomic segment, from this type of data. Other recent methods can infer copy number profiles from thousands of cells or spatial locations from single-cell RNA sequencing (scRNA-seq) [16], scATAC-seq [47], or spatial transcriptomics data [12].

The increasing availability of technologies to measure genomic copy number in thousands of cell motivates development of methods to infer the cellular phylogenies from copy number profiles. However, there are multiple challenges in inferring phylogenies from copy number profiles. First, copy number aberrations are diverse, ranging from small duplications and deletions [3] to whole chromosome shattering and reconstruction events [41]. Second, a single copy number aberration can alter the number of copies of a large section of the genome *simultaneously*. This means that loci on the genome cannot be treated as independent phylogenetic characters, a widely-used assumption in phylogenetics [18, 19, 39, 46] Finally, the increasing size ($> 10,000$ cells) and resolution ($<$ 5Kb bins) of copy number profiles require increasingly scalable algorithms.

One widely used model of copy number evolution is the *copy number transformation* (CNT) model [38]. In the CNT model, a genome is represented as a vector of non-negative integers and copy number aberrations correspond to the increase or decrease of the entries in a contiguous *interval* of coordinates in the vector, explicitly modeling the non-independence of copy number amplifications and deletions. The *CNT distance* is the minimum number of copy number events needed to transformation one profile to another. The CNT distance is computable in linear time [51] and has been used to define an evolutionary distance between profiles. Since the CNT distance is not symmetric, a variety of symmetrized CNT distances have also been used to construct copy number phylogenies using distance-based phylogenetic methods [38, 11, 50, 22]. Further, owing to its effectiveness, the CNT model has become the basis of a variety of distinct models [50, 22, 8] for copy number evolution.

While the CNT model is described by specific events – or mutations – there has been little work on constructing phylogenetic trees under the CNT model using the method of maximum parsimony. Even the small parsimony problem – where the topology of the tree is given and one aims to infer the ancestral profiles that minimizes the total number of copy number events on the tree – has no known efficient solution. For example, for the special case of a two leaf tree, the best algorithm for the CNT small parsimony problem [51] runs in $O(nB^7)$ time where B is the largest allowed copy number and n is the number of loci [11]. Without an efficient algorithm for the small parsimony problem under the CNT model, one cannot hope to solve the large parsimony problem, where the topology of the tree is unknown.

We introduce a relaxation of the CNT model, called the *zero-agnostic copy number transformation* (ZCNT) model, that approximates the CNT model and has a number of desirable properties. Unlike the CNT model,

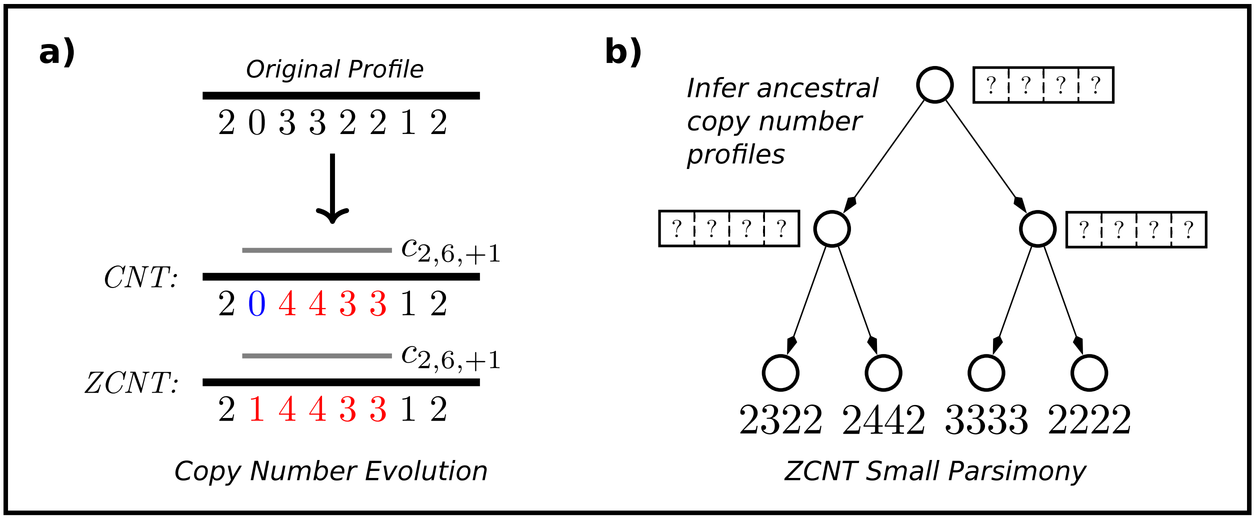


Figure 1: (a) The results of applying a copy number transformation $c_{2,6,+1}$ under both the CNT model and ZCNT model. Under the CNT model a zero cannot be increased via the amplification, but the zero-agnostic CNT model allows the zero to increase to one copy. **(b)** An instance of the ZCNT small parsimony problem: given a tree with copy number profiles labeling the leaves, the goal is to infer the ancestral copy number profiles that minimize the total ZCNT distance across all edges.

the ZCNT model allows for the amplification of zero copy number regions. While such an operation is not biologically realistic, we show that this relaxation makes the ZCNT distance a metric, in contrast to the CNT distance. Moreover, we derive a closed form expression for the ZCNT distance between two profiles. We use this closed form expression as well as an alternative characterization of copy number profiles to solve two relaxations of the small parsimony problem in polynomial time. To our knowledge, this is the first attempt to solve the small parsimony problem for a segment-based (i.e. non-independent) model of copy number evolution. We then use our efficient algorithm for the (relaxed) small parsimony problem to design an algorithm, *Lazac* (Large-scale Analysis of Zero Agnostic Copy number), for inferring copy number phylogenies by solving the large parsimony problem. We show on simulated data that *Lazac* is $> 100\times$ faster than other phylogenetic methods and also more accurate in recovering the ground truth phylogeny. On single-cell whole-genome sequencing data from human breast and ovarian tumors, *Lazac* finds phylogenies that are more consistent with both copy number clones and single-nucleotide variants (SNVs).

2 Copy number transformations

A *copy number profile* $p = [p_1, \ldots, p_m]$ is a vector of non-negative integers where $p_j \in \{0, \ldots, B\}$ is the the number of copies of locus j. Suppose we measure the copy number profile of n cells of a tumor across m loci in a single-cell DNA sequencing experiment. We encode the copy number profiles in a $n \times m$ *copy number matrix* $M = \left[M_{i,j}\right]$ where $M_{i,j} \in \{0, \ldots, B\}$ is the copy number of cell i at locus j. The copy number profile p of cell i is then the i^{th} row of this matrix, and is denoted M_i.

One of the most basic phylogenetic principles is that nearly perfect measurement of evolutionary distances enables exact recovery of the evolutionary history [40]. It is thus not surprising that many of the successful attempts at inferring copy number phylogenies focus on finding *good* methods to compute an evolutionary distance between a pair of copy number profiles. Early methods for computing evolutionary distances on copy number data [3, 7] employed simple measures of distance such as the Hamming, weighted Hamming, and ℓ_1 distance between copy number profiles. However, these distances do not account for dependencies between loci caused by long CNAs spanning contiguous segments of the genome, leading to inaccurate phylogenetic reconstruction [38, 22].

In this section, we describe and investigate the copy number transformation (CNT) model, one of the most well-known and successful evolutionary models for copy number evolution in cancer. The CNT model was originally introduced in MEDICC [38] and extended in subsequent studies [51, 11, 50, 8, 22]. Since the CNT model only allows intrachromosomal copy number events, it is sufficient to consider the case of a single chromosome, and thus for ease of exposition we will describe the model using a single chromosome.

The fundamental operation in the CNT model is a *copy number event* which increases or decreases (by one) the entries in a contiguous interval of a copy number profile, defined formally as follows.

Definition 1 (Copy number event)**.** *A* copy number event $c_{s,t,b} : \mathbb{Z}_+^n \to \mathbb{Z}_+^n$ *is a function that maps a copy number profile $p \in \mathbb{Z}_+^n$ to a profile $c_{s,t,b}(p)$ described by its entries as*

$$c_{s,t,b}(p)_i = \begin{cases} p_i + b & \text{if } s \le i \le t \text{ and } p_i \ne 0, \\ p_i & \text{otherwise,} \end{cases}$$

where $s \le t$ and $b \in \{+1, -1\}$. We denote such a function as c when clear by context.

That is, an amplification (resp. deletion) increases (resp. decreases) the copy number of all *non-zero* entries in the interval between positions s and t, or alternatively a copy number event *skips* the zero entries (Figure 1). Thus, once a locus is lost (i.e. $p_i = 0$), the locus cannot be regained or deleted further. A *copy number transformation* C is the composition of multiple copy number events and we denote this function as $C = (c_1, \ldots, c_n)$ when $C(p) = c_n(\cdots(c_2(c_1(p))))$.

Several copy number problems have been previously studied to compute evolutionary distances under the CNT model. The first, and simplest, is the *copy number transformation problem*, originally introduced in [38], which defines a distance, $\sigma(u, v)$, between two copy number profiles. Put simply, the distance between two profiles is the length of the shortest copy number transformation needed to transform one profile to another.

Definition 2 (Copy number transformation distance)**.** *Given two copy number profiles u and v, the* copy number transformation distance *is*

$$\sigma(u, v) := \min_{C(u)=v} |C|,$$

where $C = (c_1, \ldots, c_n)$ is a CNT. Alternatively, $\sigma(u, v) = \infty$ if no such transformation exists.

[51, 43] show there is a (non-trivial) strongly linear time algorithm (i.e. time complexity $O(|u| + |v|)$) for computing the CNT distance $\sigma(u, v)$. Unfortunately, the CNT distance $\sigma(u, v)$ is not symmetric (i.e. $\sigma(u, v) \ne \sigma(v, u)$), which makes it difficult to use in distance based phylogenetic methods such as neighbor joining [34].

In order to apply distance-based phylogenetic methods, multiple approaches to symmetrize the distance $\sigma(u, v)$ have been introduced. [51] use a mean correction replacing the asymmetric $\sigma(u, v)$ with a symmetric distance $\sigma'(u, v)$ defined as

$$\sigma'(u, v) = \frac{\sigma(u, v) + \sigma(v, u)}{2}.$$

Alternatively, several authors [38, 22, 11] define the distance between two profiles in terms of a closely related, median profile w. Specifically, the *median distance* between two profiles u and v is defined to be the smallest value of

$$\sigma(w, u) + \sigma(w, v)$$

over all profiles w. Computing this median distance is called the *copy number triplet problem* in [11]. Unfortunately, no efficient algorithm is known for the copy number triplet problem. The fastest algorithm uses $O(mB^7)$ time and $O(mB^4)$ space where B is the maximum allowed copy number [11].

3 Small and large copy number parsimony

The small parsimony problem for copy number profiles is the following: given a tree $\mathcal{T}$ whose leaves are labeled by copy number profiles, infer ancestral copy number profiles that minimize the total dissimilarity between profiles across all edges (Figure 1). For evolutionary models in which each character evolves independently and has finitely many states (e.g. single nucleotide substitution models), the small parsimony problem is solved in polynomial time via Sankoff's algorithm, a dynamic programming algorithm [36]. Unfortunately, the CNT model presents two major challenges in solving the small parsimony problem. First, since copy number events affect multiple loci simultaneously, the loci cannot be analyzed independently, in contrast to most phylogenetic characters. Second, the space of possible copy number profiles is a priori unbounded, since the maximum copy number of a segment in a genome is unknown. Thus, it is not surprising that there is no published solution to the small parsimony problem for CNT dissimilarity, with the exception of the special case of two-leaf trees [11]. Here, we formalize both the CNT small parsimony problem and the corresponding large parsimony problem, the latter of which was previously described in [11].

A *copy number phylogeny* $(\mathcal{T}, \ell)$ is a rooted tree $\mathcal{T}$ and leaf labeling ℓ. Let $V(\mathcal{T})$, $E(\mathcal{T})$, and $L(\mathcal{T})$ denote the edges, vertices, and leaves of $\mathcal{T}$, respectively. In our applications below, each leaf of $\mathcal{T}$ represents one of the n cells (or bulk samples) from a tumor. An ancestral labeling $\hat{\ell}$ of a copy number phylogeny is a vertex labeling of $\mathcal{T}$ that agrees with ℓ on the leaves of $\mathcal{T}$, i.e. $\ell(v) = \hat{\ell}(v)$ when $v \in L(\mathcal{T})$. We say that $(\mathcal{T}, \ell)$ is a copy number phylogeny for copy number matrix M if $\mathcal{T}$ has n leaves such that ℓ labels each leaf by a row of M. Formally, if $(\mathcal{T}, \ell)$ is a copy number phylogeny for a copy number matrix M, then there exists a cell assignment $\pi : [n] \to L(\mathcal{T})$ that assigns each cell i to a leaf v such that $\ell(v) = M_i$.

We define the cost $J(\mathcal{T}, \hat{\ell})$ of a vertex labeled, copy number phylogeny as the total number of copy number events required to explain the phylogeny:

$$J(\mathcal{T}, \hat{\ell}) := \sum_{(u,v) \in E(\mathcal{T})} \sigma(\hat{\ell}(u), \hat{\ell}(v)).$$

We now introduce the small parsimony problem [14] under the copy number transformation model.

Problem 1 (CNT Small parsimony). *Given a copy number phylogeny $(\mathcal{T}, \ell)$ find an ancestral labeling $\hat{\ell} : V(\mathcal{T}) \to \{0, \dots, B\}^m$ such that (i) $\hat{\ell}(v) = \ell(v)$ for all leaves $v \in L(\mathcal{T})$ and (ii) $J(\mathcal{T}, \hat{\ell})$ is minimized.*

The *parsimony score* is defined as the cost $J(\mathcal{T}, \hat{\ell})$ of the solution $\hat{\ell}$ to the CNT small parsimony problem. To the best of our knowledge the CNT small parsimony problem (Problem 1) has not been analyzed in the literature. We believe this is due to the difficulty of solving the CNT small parsimony problem. That is, for even a special case of two-leaf trees, referred to as the *copy number triplet* problem [11], no strongly polynomial time algorithm is known (Section 2).

The CNT large parsimony problem defined in [11], aims to find a vertex labeled, copy number phylogeny $(\hat{\mathcal{T}}, \ell)$ for a matrix M with minimum cost.

Problem 2 (CNT Large parsimony). *Given copy number matrix M, find a copy number phylogeny $(\mathcal{T}, \ell)$ for M and ancestral labeling $\hat{\ell}$ such that $J(\mathcal{T}, \hat{\ell})$ is minimized.*

Unsurprisingly, [11] showed that the above large parsimony problem (Problem 2) is NP-hard. They also formulated an integer linear program (ILP) to solve the problem exactly. However, this ILP consists of

$O(n^2m + nm \log B)$ variables and does not scale to the size of current real data sets with thousands of cells.

4 The zero-agnostic CNT model

The copy number transformation (CNT) model imposes the constraint that once a locus is lost (has zero copy number), the locus remains with zero copies for all time. While this constraint is biologically realistic, the constraint also makes the inference problems – including the CNT small (and large) parsimony problems – computationally hard to solve. Here, we show that *relaxing* the constraint that copy number events do not alter zero entries leads to a simpler model with favourable mathematical properties. We call this the *zero-agnostic* copy number model (Figure 1) to indicate that the model allows the amplification and deletion of loci with zero copies. Formally, we define a *zero-agnostic copy number event* as follows.

Definition 3 (Zero-agnostic copy number event). *A* zero-agnostic copy number event $c_{s,t,b} : \mathbb{Z}^n \to \mathbb{Z}^n$ *is a function that maps a profile* $p \in \mathbb{Z}^n$ *to a profile written* $c_{s,t,b}(p)$ *described by its entries as*

$$c_{s,t,b}(p)_i = \begin{cases} p_i + b, & \text{if } s \leq i \leq t, \\ p_i & \text{otherwise,} \end{cases}$$

where $s \leq t$ *and* $b \in \{+1, -1\}$. *We denote such a function as* c *when clear by context.*

Thus, a zero-agnostic copy number event either increases or decreases the number of copies of all loci in the interval (s, t) regardless of whether the loci have zero copies. While our formulation allows for the number of copies of a locus to decrease below zero, one can show that given two profiles with non-negative entries, it is always possible to find an optimal ZCNT such that no intermediate profile has negative entries. Specifically, as a corollary to the commutativity of zero-agnostic copy number transformations (Proposition 2), one can re-order events such that amplifications always occur first.

Due to space constraints, we do not include all proofs in the main text. Any proof not present in the main text can be found in Supplementary Proofs C.

4.1 Delta profiles

We simplify our analysis of zero-agnostic copy number events by examining their effect on the differences between the copy number of adjacent loci. In particular, while a zero-agnostic copy number event $c_{s,t,b}$ increments (or decrements) all entries p_i where $i \in \{s, \ldots, t\}$, $c_{s,t,b}$ only alters two *differences* between adjacent loci, namely the difference $p_s - p_{s-1}$ and the difference $p_{t+1} - p_t$. To formalize this idea, we first define the *delta profile*, vectors obtained by taking the differences in copy number between adjacent loci.

Definition 4. *A* delta profile *is any vector* $p \in \mathbb{Z}^n$ *that satisfies the* balancing condition*:*

$$\sum_{i=1}^{n} p_i = 0. \tag{1}$$

Or equivalently, $\sum_{p_i > 0} |p_i| = \sum_{p_i < 0} |p_i|$. *We denote the set of delta profiles in* $\mathbb{Z}^n$ *as* $\mathcal{D}_n$.

The above definition provides us with a convenient (and useful) description of the image of the following difference transformation, which we call the *delta map*.

Definition 5. *The delta map* $\Delta : \mathbb{Z}^n \to \mathcal{D}_{n+1}$ *maps a copy number profile* p *to a delta profile* $\Delta(p)$ *by taking the differences in adjacent copy number loci. Specifically,*

$$\Delta(p)_1 = p_1 - 2, \ \Delta(p)_i = p_i - p_{i-1}, \ \text{and} \ \Delta(p)_{n+1} = 2 - p_n$$

where the constant 2 *represents a normal, diploid copy number.*

A basic property of the delta map $\Delta : \mathbb{Z}^n \to \mathcal{D}_{n+1}$ is that it is invertible.

Proposition 1. *The delta map $\Delta : \mathbb{Z}^n \to \mathcal{D}_{n+1}$ is invertible.*

Since Δ is one-to-one and onto with respect to $\mathcal{D}_{n+1}$, each delta profile p' then corresponds to a unique copy number profile $p = \Delta^{-1}(p')$.

Interestingly, a copy number event $c_{s,t,b}$ applied to a copy number profile p only affects two entries of the delta profile $\Delta(p)$, meaning that loci of the corresponding delta profile are (nearly) independent. We formalize this in the following definition of a *delta event.*

Definition 6 (Delta event)**.** *A* delta event $\delta_{s,t,b} : \mathcal{D}_n \to \mathcal{D}_n$ *is a function that maps a delta profile $p \in \mathcal{D}_n$ to a delta profile $\delta_{s,t,b}(p)$ described by its entries as*

$$\delta_{s,t,b}(p)_i = \begin{cases} p_i + b & \text{if } i = s \\ p_i - b & \text{if } i = t+1 \\ p_i & \text{otherwise,} \end{cases}$$

where $s \leq t$ and $b \in \{+1, -1\}$. We denote such a function as δ when clear by context.

A *delta transformation* $D = (\delta_1, \ldots, \delta_n)$ is the composition of multiple delta events, where $D(p) = \delta_n(\cdots(\delta_2(\delta_1(p))))$. We now state the connection between delta events and zero-agnostic copy number (ZCNT) events in the following theorem and corollary.

Theorem 1. *Let $c_{s,t,b}$ be a zero-agnostic copy number event and $\delta_{s,t,b}$ be a delta event. Then,*

$$p' = c_{s,t,b}(p) \quad \text{if and only if} \quad \Delta(p') = \delta_{s,t,b}(\Delta(p)).$$

Corollary 1. *Let $C = (c_{s_1,t_1,b_1}, \ldots, c_{s_n,t_n,b_n})$ be a zero-agnostic copy number transformation and $D = (\delta_{s_1,t_1,b_1}, \ldots, \delta_{s_n,t_n,b_n})$ be the corresponding delta transformation. Then,*

$$p' = C(p) \quad \text{if and only if} \quad \Delta(p') = D(\Delta(p))$$

Proof. The corollary follows by induction on $|C|$ and repeated application of (Theorem 1). ■

4.2 Computing the ZCNT distance

Let $d(u, v)$ be the minimum number of *zero-agnostic* copy number events needed to transform the copy number profile u to v. In this section we derive a closed form expression for $d(u, v)$.

We begin by noting that $d(u, v)$ is equal to the minimum number $d'(\Delta(u), \Delta(v))$ of delta events needed to transform delta profile $\Delta(u)$ to $\Delta(v)$. This follows from the equivalence between the copy number transformations and the corresponding delta transformation (Corollary 1). Thus, it suffices to only consider delta profiles and delta events; for the rest of the section all profiles u and v are delta profiles unless otherwise specified.

We start by observing two basic facts: delta transformations are commutative and $d'(u, v)$ forms a metric.

Proposition 2. *A delta transformation $D = (\delta_1, \ldots, \delta_n)$ is commutative. That is, the application of D to a profile is identical to the application of $D_\sigma = (\delta_{\sigma_1}, \ldots, \delta_{\sigma_n})$ where σ is any permutation of $\{1, \ldots, n\}$.*

Proposition 3. *$d'(u, v)$ is a distance metric. Further,*

$$d'(u, v) = d'(v - u, 0) = d'(u - v, 0).$$

Note that this also implies that zero-agnostic copy number transformations are commutative and that $d(\cdot,\cdot)$ is a distance metric. To see this, let $C = (c_{s_1,t_1,b_1},\ldots,c_{s_n,t_n,b_n})$ be a zero-agnostic copy number transformation and $D = (\delta_{s_1,t_1,b_1},\ldots,\delta_{s_n,t_n,b_n})$ be the corresponding delta transformation, then for any vector p and permutation σ,

$$C(p) = p' \iff D(\Delta(p)) = \Delta(p') \iff D_\sigma(\Delta(p)) = \Delta(p') \iff C_\sigma(p) = p',$$

where the first equivalence follows from Corollary 1, the second from Proposition 2, and the third from Corollary 1. This implies that $C(p) = C_\sigma(p)$, which proves that a zero-agnostic copy number transformation is commutative. To see that $d(\cdot,\cdot)$ is a distance metric, it suffices to observe that $d(u,v) = d'(\Delta(u),\Delta(v))$ implies symmetry and reflexivity. Then, the triangle inequality is satisfied since the composition of a zero-agnostic copy number transformation from u to w and w to v yields a copy number transformation from u to v.

From our characterization of delta profiles, we derive our expression for the distance between delta profiles.

Theorem 2. *For delta profiles u and v, $d'(u,0) = \frac{1}{2}\|u\|_1$. Thus, $d'(u,v) = \frac{1}{2}\|u-v\|_1$.*

Proof. Since each event decreases the total magnitude of $\|\Delta(u)\|_1$ by at most two, to transform $\Delta(u)$ to the 0 profile requires at least $\frac{1}{2}\|\Delta(u)\|_1$ events.

We prove the other direction by induction on $\sum_{u_i>0}|\Delta(u)_i|$. Clearly, if the sum is zero, the claim holds. Otherwise, by (Proposition 1), we can choose c to be any event that decrements $i \in \{i : \Delta(u)_i > 0\}$ and increments $j \in \{j : \Delta(u)_j < 0\}$. Applying c to $\Delta(u)$ results in a delta profile $\Delta(u')$ such that $\|\Delta(u')\|_1 = \|\Delta(u)\|_1 - 2$. Invoking the induction hypothesis then yields a sequence of $\frac{1}{2}\|\Delta(u)\|$ events to transform $\Delta(u)$ to the 0 profile.

The second statement follows from Proposition 3. ■

As a corollary to the above theorem and the equivalence between zero-agnostic copy number transformations and delta transformations (Corollary 1), we have our closed form expression for the ZCNT distance between copy number profiles.

Corollary 2. *For copy number profiles p and p',*

$$d(p,p') = d'(\Delta(p),\Delta(p')) = \frac{1}{2}\|\Delta(p) - \Delta(p')\|_1.$$

Further, as a corollary to the fact that $d(u,v)$ is a distance metric, the following median distance is trivially computed in linear time:

Corollary 3. *Given two copy number profiles u and v, both u and v minimize the median distance $d(w,u) + d(w,v)$ over all choices of copy number profiles w. Thus,*

$$\min_{w\in\mathbb{Z}_+^m} \{d(w,u) + d(w,v)\} = d(u,v).$$

5 Algorithms

5.1 ZCNT small parsimony

We show below that the special form of the ZCNT model enables us to solve two natural relaxations of the small parsimony problem in polynomial time. First, using the equivalence between copy number profiles and delta profiles described above, we formulate the small parsimony problem (Problem 1) using the ZCNT model as follows.

Problem 3 (ZCNT Small Parsimony). *Given a copy number matrix M, a tree $\mathcal{T}$ and cell assignment $\pi : [n] \longrightarrow L(\mathcal{T})$, find a vertex labeling $\ell : V(\mathcal{T}) \longrightarrow \mathbb{Z}^m$ minimizing*

$$\sum_{(u,v)\in E(\mathcal{T})} \frac{1}{2}\|\ell(u) - \ell(v)\|_1$$

such that the following two conditions are satisfied:

i. $\ell(\pi(i)) = \Delta(M_i)$ for all cells $i \in [n]$,

ii. $l(u)$ satisfies the balancing condition (1) for all vertices $u \in V(\mathcal{T})$.

To solve the above problem, we recall the general form of the Sankoff-Rousseau recurrence [9, 36] for solving the small parsimony problem. Let $c(\mathcal{T}; x)$ be the cost of the optimal labeling $\hat{\ell}$ of $V(\mathcal{T})$ that agrees with copy number matrix M and has label x for the root. Let $\mathcal{T}_w$ denote the sub-tree rooted at w and suppose that w has children u and v. Then, by condition (ii), and the requirement that the ancestral labeling lies in $\mathbb{Z}^m$, we have the following recurrence relation [9]:

$$c(\mathcal{T}_w; x) = \min_{y,z\in\mathcal{D}_n} \left\{\frac{1}{2}\|x - y\|_1 + \frac{1}{2}\|x - z\|_1 + c(\mathcal{T}_u; y) + c(\mathcal{T}_v; z)\right\}, \quad (2)$$

This recurrence has several difficulties. First, y and z are unbounded and can take on any value in $\mathbb{Z}^m$. Thus, it is impossible to store a dynamic programming table for $c(T_w; x)$ without imposing bounds on the maximum copy number. Further, even when the entries are constrained to a bounded interval $\{0, \ldots, B\}^m$, the dynamic programming table has size $(B + 1)^m$, exponentially large. Second, because of the balancing conditions (1), $\sum_{i=1}^n y_i = \sum_{i=1}^n z_i = 0$, one cannot analyze the loci independently.

Despite these challenges, the recurrence (2) is a substantial improvement over the analogous recurrence under the CNT model. In fact, if we remove *either the balancing* (1) *or the integrality condition*, we can solve this recurrence in (resp. strong or weak) polynomial time.

Theorem 3. *If the balancing condition* (1) *is dropped, the ZCNT small parsimony problem can be solved in $O(mn)$ time. If the integrality condition is dropped, the ZCNT small parsimony problem can be solved in (weakly) polynomial time using a linear program with $O(mn)$ variables and constraints.*

Both of these facts derive from our closed form expression for the ZCNT distance between two copy number profiles in terms of the ℓ_1 norm. We sketch the ideas here, and refer to Supplementary Results B.2, B.1 for proofs of these claims.

For the first case when we drop the balancing condition, we can analyze the loci independently as there is no constraint on the entries of the ancestral profiles. Then, it suffices to observe that since the distance corresponds to the absolute difference, the function $c(\mathcal{T}_w; x)$ has a nice structure and we do not have to store an infinitely large dynamic programming table. When the integrality is removed, then, since both (i) and (ii) are *linear* constraints on the profiles and because ℓ_1 norm minimization can be written as a linear program (LP), there is an LP formulation of the ZCNT small parsimony problem. As it is well known we can solve LPs in (weakly) polynomial time, this concludes the second case.

5.2 *Lazac* algorithm for ZCNT large parsimony

We develop a tree-search algorithm, *Lazac*, to find approximate solutions to the ZCNT large parsimony problem (Problem 2). Our procedure searches the space of copy number trees for a given copy number matrix C using sub-tree interchange operations [27] and relies heavily on the efficient algorithm we developed for the small parsimony problem (Problem 3) when the balancing condition (1) is dropped. The

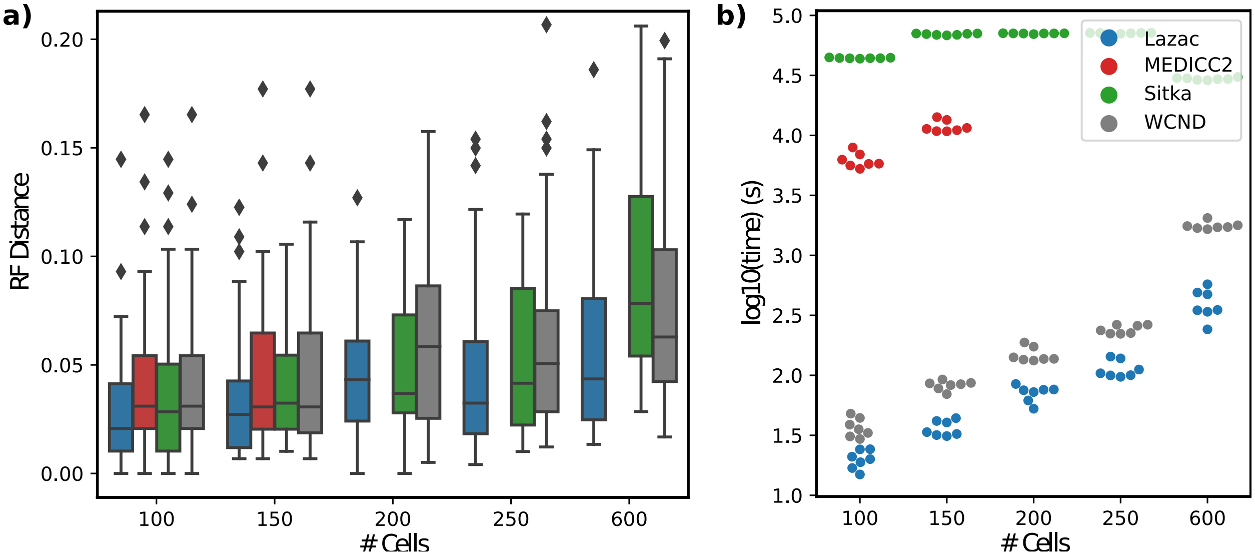


Figure 2: **(a)** Comparison of reconstruction accuracy (RF distance) on simulated data for several state-of-the art methods for copy number tree reconstruction with varying number of cells $n = 100, 150, 200, 250, 600$ across four sets of loci $l = 1000, 2000, 3000, 4000$ and seven random seeds $s = 0, 1, 2, 3, 4, 5, 6$. **(b)** Timing results for varying number of cells $n = 100, 150, 250, 600$ and fixed number of loci $l = 4000$. As MEDICC2 was too slow to run on more than 150 cells (with a 2 hour time limit), we exclude it from comparisons where the number of cells $n > 150$.

procedure is similar to the tree search procedure we developed for lineage tracing data [37]. Complete details on our tree search procedure are in Supplementary Methods A.1. *Lazac* is implemented in C++17 and is freely available at: `github.com/raphael-group/lazac-copy-number`.

6 Results

6.1 Comparison of copy number distances and phylogenies on prostate cancer data

We first investigated the differences between the CNT and ZCNT distances on copy number profiles inferred from bulk whole-genome sequencing data from ten metastatic prostate cancer patients [17]. We analyzed the copy number profiles for these patients published in [22]. For each pair of copy number profiles from distinct samples (e.g. anatomical sites) from the same patient, we computed the CNT distance d_{CNT} and ZNCT distance d_{ZCNT}. We found that for all ten patients, the median relative difference $|d_{\text{CNT}}/d_{\text{ZCNT}} - 1|$ over all pairs of samples was less than 5% – and for most patients the relative difference was even smaller (Supplementary Figure 5, 6).

6.2 Evaluation on simulated data

We compared *Lazac* to several state-of-the-art methods for inferring copy number phylogenies – namely MEDICC2 [22], MEDALT [43], Sitka [35], and WCND [50] – on simulated data.

Lazac inferred the most accurate phylogenetic trees across varying number of cells ($n = 100, 150, 200, 250, 600$) and loci ($l = 1000, 2000, 3000, 4000$). In particular, we found that on all but one parameter setting, *Lazac* had the lowest median RF (Figure 2) and Quartet distance (Supplementary Methods A.5) on simulated instances (Supplementary Figure 1). Further, on large phylogenies containing $n = 600$ cells, *Lazac* showed an even larger improvement in median RF (Figure 2) and Quartet distance (Supplementary Methods A.5) over other methods, owing to its scalability. In terms of speed, *Lazac* was the fastest method

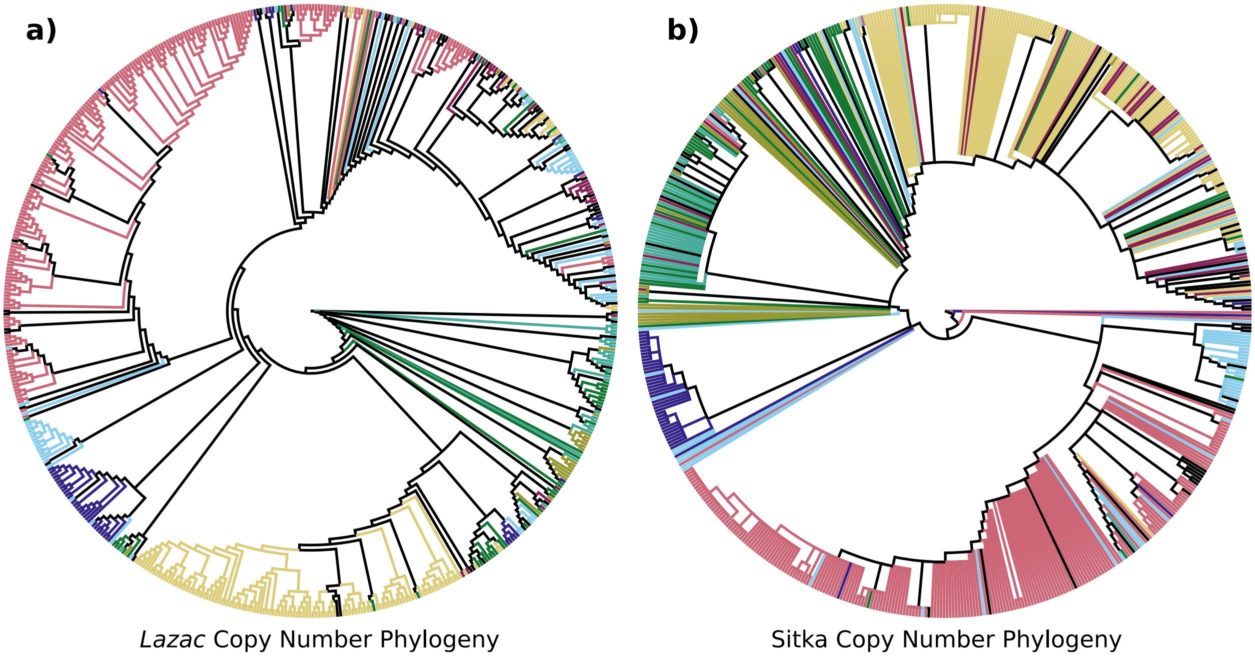


Figure 3: The copy number phylogenies inferred by *Lazac* **(a)** and Sitka **(b)** on sample SA1184 with the leaves colored by the corresponding clone labels, as visualized using Iroki [29]. The normalized RF distance between the two trees is 0.9869.

on every instance, taking less than ~250 seconds to run on the largest simulated dataset containing 600 cells. Further, it was ~100 times faster than the other top performing methods Sitka and MEDICC2 (Figure 2b).

As a further evaluation of the differences between the ZCNT and CNT distances, we compared the trees obtained using distance-based phylogenetic methods with the ZCNT and CNT distances. Specifically, we compared the performance of applying neighbor joining on the ZCNT distances, referred to as *Lazac-NJ*, to three distance-based methods for reconstructing copy number phylognies: MEDICC2 [22], WCND [50], and MEDALT [43] on simulated data. MEDICC2 and WCND compute distances based on extensions of the CNT model and then apply neighbor joining to infer phylogenies. As such, they allow for a natural benchmark with which to compare our simpler, ZCNT distance. *Lazac-NJ* had nearly identical (within 1%) median RF and Quartet distance compared to other distance based methods (Supplementary Figure 2, 3) and was often the top performer. This provides evidence that even by itself, the ZCNT distance is useful for phylogenetic reconstruction.

6.3 Single-cell DNA sequencing data

We used *Lazac* to analyze single-cell whole genome sequencing (WGS) data from 25 human breast and ovarian tumor samples [15]. This dataset was generated using the DLP+ [23], single-cell whole-genome sequencing technology which produces ≈ 0.04× coverage from a median (resp. mean) of 636 (resp. 1457) cells per sample. The original study used Sitka [35], a method that uses the breakpoints between copy number segments as phylogenetic markers, to construct copy number phylogenies using this data.

We found that the phylogenies inferred by *Lazac* are substantially different than the phylogenies constructed by Sitka. Specifically, the normalized RF distance between pairs of phylogenies was greater than 0.90 in all cases (Supplementary Figure 9). In many cases, the normalized RF distance was 1, indicating that the phylogenies completely disagree.

To investigate these differences, we analyzed the concordance between the phylogenies and the assignments of cells to copy number clones, reported in the original publication [15]. Specifically, we defined the *clonal discordance score* as the parsimony score of the clonal labeling; i.e. the minimum number of changes in clone label that are required to label the leaves with the published clone labels. Thus, for a dataset with k clone labels the minimum possible clonal discordance score for a tree is $k-1$ corresponding to the case each clone label is a clade in the tree. We find that on 18/25 of the samples, the *Lazac* phylogenies had substantially lower clonal discordance scores than the Sitka phylogenies (Supplementary Figure 4, 10) showing that the *Lazac* phylogenies were more concordant with the copy number clones compared to the published phylogenies. As a qualitative example, the phylogeny constructed for sample SA1184 by *Lazac* also appears more concordant with the copy number clones than that inferred by Sitka (Figure 3). Further details on the clonal discordance analysis are in Supplementary Methods A.3.

As a further evaluation of the *Lazac* and Sitka phylogenies, we examined whether somatic single-nucleotide variants (SNVs) supported the splits in each phylogeny, following the approach of [48]. Note that these SNVs were not used in the inference of either phylogeny, and thus they provide independent validation of the phylogeny. Given the extremely low sequence coverage (0.04× per cell), it is not possible to reliably measure SNVs of individual cells. Thus, we performed this analysis on the three samples (SA039, SA604, SA1035) with the largest number of cells. We identified subtrees in the phylogeny with at least 5% and at most 15% of the cells and identified SNVs present in the subpopulation of cells in these subtrees. Following the approach in [48], we perform a permutation test to determine whether the subtree is supported by more SNVs than expected (Supplementary Methods A.4). For all three samples, we found that the *Lazac* phylogenies had a greater fraction of supported subtrees ($P < 0.05$) than the *Sitka* phylogenies (Supplementary Figure 11). On the largest sample, SA1035, we identified five out of six supported subtrees (supported by 3175, 3334, 3799, 3435, and 3402 SNVs) for the *Lazac* phylogeny compared to only three of eight statistically significant subtrees (supported by 3426, 3129, and 3362 SNVs) for the Sitka phylogeny.

6.4 Approximation error of ZCNT small parsimony relaxations

We investigated the approximation error produced by the relaxations (Section 5.1) used in our two polynomial time algorithm's for the ZCNT small parsimony problem. To perform this investigation, we first generated a set of 200 copy number phylogenies by stochastically perturbing a phylogeny inferred by Sitka [35] from single-cell whole genome sequencing data (Section 6.3). Then, for each phylogeny, we computed the optimal solution to the ZCNT small parsimony problem and its two relaxations using (integer) linear programming.

Importantly, we found that the exact solution to the ZCNT small parsimony problem and the solution obtained by relaxing the integrality condition were identical in every case. This leads us to believe that the relaxed linear program has a special structure, which we state as the following conjecture:

Conjecture 1. *The constraint matrix A of the linear program obtained by relaxing the integrality condition of the ZCNT small parsimony problem is totally unimodular.*

If true, this would imply that ZCNT small parsimony can be solved *exactly* in polynomial time using linear programming.

In contrast, dropping the balancing condition resulted in solutions with a lower score, implying that the balancing condition does meaningfully constrain the solution space. Specifically, the Pearson correlation between the score produced by dropping the balancing condition and the exact ZCNT small parsimony score was $R^2 = 0.972$ ($p < 10^{-100}$) across the 200 copy number phylogenies (Supplementary Figure 7).

Further, as the number of stochastic perturbations increased, both the parsimony score of the relaxed and the exact solutions increased (Supplementary Figure 8). This serves as a sanity check for the ZCNT small parsimony score since the phylogeny generally becomes further from ground truth as the number of stochastic perturbations increase.

7 Discussion

We introduced the *zero-agnostic copy number transformation* (ZCNT) model, a relaxation of the CNT model that allows for modification of zero copy number regions. We derived a closed-form expression for the ZCNT distance and presented polynomial time algorithms to solve two natural approximations of the small parsimony problem for copy number profiles. We used our efficient algorithm for the small parsimony problem to derive a method *Lazac*, to solve the large parsimony problem for copy number profiles. We demonstrated that on both real and simulated data, *Lazac* found better copy number phylogenies than existing methods.

There are multiple directions for future work. First, the complexity of the small parsimony problems for both the CNT and ZCNT models remains open, though we conjecture, and provide empirical evidence, that the latter is polynomial. Second, the algorithm we developed for the ZCNT large parsimony problem relies on a simple, hill climbing search using nearest-neighbor interchange operations. We expect that a more advanced approach that uses state-of-the-art techniques from phylogenetics [27, 32] could substantially improve both inference speed and accuracy. Finally, is to apply *Lazac* to other large single-cell WGS datasets [48, 28]. We anticipate that the scalability and accuracy of *Lazac* will be useful in analyzing the increasing amount of single-cell WGS cancer sequencing data.

8 Acknowledgements

We thank Uthsav Chitra and Gillian Chu for their helpful comments and review of this manuscript during several stages of its preparation. This work was supported by National Cancer Institute (NCI) grant U24CA248453 to B.J.R.

References

[1] Donna G Albertson, Colin Collins, Frank McCormick, and Joe W Gray. Chromosome aberrations in solid tumors. *Nature genetics*, 34(4):369–376, 2003.

[2] Noemi Andor, Billy T Lau, Claudia Catalanotti, Anuja Sathe, Matthew Kubit, Jiamin Chen, Cristina Blaj, Athena Cherry, Charles D Bangs, Susan M Grimes, et al. Joint single cell dna-seq and rna-seq of gastric cancer cell lines reveals rules of in vitro evolution. *NAR Genomics and Bioinformatics*, 2(2):lqaa016, 2020.

[3] Niko Beerenwinkel, Roland F Schwarz, Moritz Gerstung, and Florian Markowetz. Cancer evolution: mathematical models and computational inference. *Systematic biology*, 64(1):e1–e25, 2015.

[4] Damian Bogdanowicz and Krzysztof Giaro. Matching split distance for unrooted binary phylogenetic trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(1):150–160, 2011.

[5] Damian Bogdanowicz, Krzysztof Giaro, and Borys Wróbel. Treecmp: comparison of trees in polynomial time. *Evolutionary Bioinformatics*, 8:EBO–S9657, 2012.

[6] Gabriel Cardona, Francesc Rosselló, and Gabriel Valiente. Extended newick: it is time for a standard representation of phylogenetic networks. *BMC bioinformatics*, 9(1):1–8, 2008.

[7] Salim Akhter Chowdhury, Stanley E Shackney, Kerstin Heselmeyer-Haddad, Thomas Ried, Alejandro A Schäffer, and Russell Schwartz. Phylogenetic analysis of multiprobe fluorescence in situ hybridization data from tumor cell populations. *Bioinformatics*, 29(13):i189–i198, 2013.

[8] Garance Cordonnier and Manuel Lafond. Comparing copy-number profiles under multi-copy amplifications and deletions. *BMC genomics*, 21(2):1–12, 2020.

[9] Miklós Csűrös. How to infer ancestral genome features by parsimony: Dynamic programming over an evolutionary tree. In *Models and Algorithms for Genome Evolution*, pages 29–45. Springer, 2013.

[10] Xiao Dong, Lei Zhang, Xiaoxiao Hao, Tao Wang, and Jan Vijg. Sccnv: a software tool for identifying copy number variation from single-cell whole-genome sequencing. biorxiv. *Preprint*, 10:535807, 2019.

[11] Mohammed El-Kebir, Benjamin J Raphael, Ron Shamir, Roded Sharan, Simone Zaccaria, Meirav Zehavi, and Ron Zeira. Complexity and algorithms for copy-number evolution problems. *Algorithms for Molecular Biology*, 12(1):1–11, 2017.

[12] Rebecca Elyanow, Ron Zeira, Max Land, and Benjamin J Raphael. Starch: Copy number and clone inference from spatial transcriptomics data. *Physical Biology*, 18(3):035001, 2021.

[13] James S Farris. Methods for computing wagner trees. *Systematic Biology*, 19(1):83–92, 1970.

[14] Walter M Fitch. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, 20(4):406–416, 1971.

[15] Tyler Funnell, Ciara H O'Flanagan, Marc J Williams, Andrew McPherson, Steven McKinney, Farhia Kabeer, Hakwoo Lee, Sohrab Salehi, Ignacio Vázquez-García, Hongyu Shi, et al. Single-cell genomic variation induced by mutational processes in cancer. *Nature*, pages 1–10, 2022.

[16] Teng Gao, Ruslan Soldatov, Hirak Sarkar, Adam Kurkiewicz, Evan Biederstedt, Po-Ru Loh, and Peter V Kharchenko. Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes. *Nature Biotechnology*, pages 1–10, 2022.

[17] Gunes Gundem, Peter Van Loo, Barbara Kremeyer, Ludmil B Alexandrov, Jose Tubio, Elli Papaemmanuil, Daniel S Brewer, Heini ML Kallio, Gunilla Högnäs, Matti Annala, et al. The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547):353–357, 2015.

[18] Dan Gusfield. Efficient algorithms for inferring evolutionary trees. *Networks*, 21(1):19–28, 1991.

[19] Dan Gusfield. *ReCombinatorics: the algorithmics of ancestral recombination graphs and explicit phylogenetic networks*. MIT press, 2014.

[20] Sandra Hui and Rasmus Nielsen. Sconce: a method for profiling copy number alterations in cancer evolution using single-cell whole genome sequencing. *Bioinformatics*, 38(7):1801–1808, 2022.

[21] Matthew G Jones, Alex Khodaverdian, Jeffrey J Quinn, Michelle M Chan, Jeffrey A Hussmann, Robert Wang, Chenling Xu, Jonathan S Weissman, and Nir Yosef. Inference of single-cell phylogenies from lineage tracing data using cassiopeia. *Genome biology*, 21(1):1–27, 2020.

[22] Tom L Kaufmann, Marina Petkovic, Thomas BK Watkins, Emma C Colliver, Sofya Laskina, Nisha Thapa, Darlan C Minussi, Nicholas Navin, Charles Swanton, Peter Van Loo, et al. Medicc2: whole-genome doubling aware copy-number phylogenies for cancer evolution. *Genome biology*, 23(1):1–27, 2022.

[23] Emma Laks, Andrew McPherson, Hans Zahn, Daniel Lai, Adi Steif, Jazmine Brimhall, Justina Biele, Beixi Wang, Tehmina Masud, Jerome Ting, et al. Clonal decomposition and dna replication states defined by scaled single-cell genome sequencing. *Cell*, 179(5):1207–1221, 2019.

[24] Xian F Mallory, Mohammadamin Edrisi, Nicholas Navin, and Luay Nakhleh. Methods for copy number aberration detection from single-cell dna-sequencing data. *Genome biology*, 21(1):1–22, 2020.

[25] Magda Markowska, Tomasz Cąkała, Błażej Miasojedow, Bogac Aybey, Dilafruz Juraeva, Johanna Mazur, Edith Ross, Eike Staub, and Ewa Szczurek. Conet: Copy number event tree model of evolutionary tumor history for single-cell data. *Genome Biology*, 23(1):1–35, 2022.

[26] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[27] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt Von Haeseler, and Robert Lanfear. Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, 37(5):1530–1534, 2020.

[28] Darlan C Minussi, Michael D Nicholson, Hanghui Ye, Alexander Davis, Kaile Wang, Toby Baker, Maxime Tarabichi, Emi Sei, Haowei Du, Mashiat Rabbani, et al. Breast tumours maintain a reservoir of subclonal diversity during expansion. *Nature*, 592(7853):302–308, 2021.

[29] Ryan M Moore, Amelia O Harrison, Sean M McAllister, Shawn W Polson, and K Eric Wommack. Iroki: automatic customization and visualization of phylogenetic trees. *PeerJ*, 8:e8584, 2020.

[30] Artem S Novozhilov, Georgy P Karev, and Eugene V Koonin. Biological applications of the theory of birth-and-death processes. *Briefings in bioinformatics*, 7(1):70–85, 2006.

[31] Peter C Nowell. The clonal evolution of tumor cell populations: Acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression. *Science*, 194(4260):23–28, 1976.

[32] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490, 2010.

[33] David F Robinson and Leslie R Foulds. Comparison of phylogenetic trees. *Mathematical biosciences*, 53(1-2):131–147, 1981.

[34] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.

[35] Sohrab Salehi, Fatemeh Dorri, Kevin Chern, Farhia Kabeer, Nicole Rusk, Tyler Funnell, Marc J Williams, Daniel Lai, Mirela Andronescu, Kieran R Campbell, et al. Cancer phylogenetic tree inference at scale from 1000s of single cell genomes. 2020.

[36] David Sankoff and Pascale Rousseau. Locating the vertices of a steiner tree in an arbitrary metric space. *Mathematical Programming*, 9(1):240–246, 1975.

[37] Palash Sashittal, Henri Schmidt, Michelle M Chan, and Benjamin J Raphael. Startle: a star homoplasy approach for crispr-cas9 lineage tracing. *bioRxiv*, 2022.

[38] Roland F Schwarz, Anne Trinh, Botond Sipos, James D Brenton, Nick Goldman, and Florian Markowetz. Phylogenetic quantification of intra-tumour heterogeneity. *PLoS computational biology*, 10(4):e1003535, 2014.

[39] Charles Semple, Mike Steel, et al. *Phylogenetics*, volume 24. Oxford University Press on Demand, 2003.

[40] Mike Steel. *Phylogeny: discrete and random processes in evolution*, pages 111–145. SIAM, 2016.

[41] PJ Stephens, CD Greenman, BY Fu, FT Yang, GR Bignell, LJ Mudie, ED Pleasance, KW Lau, D Beare, LA Stebbings, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1):2740, 2011.

[42] Warren H Wagner. Problems in the classification of ferns. *Recent advances in botany*, pages 841–844, 1961.

[43] Fang Wang, Qihan Wang, Vakul Mohanty, Shaoheng Liang, Jinzhuang Dou, Jincheng Han, Darlan Conterno Minussi, Ruli Gao, Li Ding, Nicholas Navin, et al. Medalt: single-cell copy number lineage tracing enabling gene discovery. *Genome biology*, 22(1):1–22, 2021.

[44] Rujin Wang, Dan-Yu Lin, and Yuchao Jiang. Scope: a normalization and copy-number estimation method for single-cell dna sequencing. *Cell systems*, 10(5):445–452, 2020.

[45] Xuefeng Wang, Hao Chen, and Nancy R Zhang. Dna copy number profiling using single-cell sequencing. *Briefings in bioinformatics*, 19(5):731–736, 2018.

[46] Tandy Warnow. *Computational phylogenetics: an introduction to designing methods for phylogeny estimation*. Cambridge University Press, 2017.

[47] Chi-Yun Wu, Billy T Lau, Heon Seok Kim, Anuja Sathe, Susan M Grimes, Hanlee P Ji, and Nancy R Zhang. Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer. *Nature biotechnology*, 39(10):1259–1269, 2021.

[48] Simone Zaccaria and Benjamin J Raphael. Characterizing allele-and haplotype-specific copy numbers in single cells with chisel. *Nature biotechnology*, 39(2):207–214, 2021.

[49] Hans Zahn, Adi Steif, Emma Laks, Peter Eirew, Michael VanInsberghe, Sohrab P Shah, Samuel Aparicio, and Carl L Hansen. Scalable whole-genome single-cell library preparation without preamplification. *Nature methods*, 14(2):167–173, 2017.

[50] Ron Zeira and Benjamin J Raphael. Copy number evolution with weighted aberrations in cancer. *Bioinformatics*, 36(Supplement_1):i344–i352, 2020.

[51] Ron Zeira, Meirav Zehavi, and Ron Shamir. A linear-time algorithm for the copy number transformation problem. *Journal of Computational Biology*, 24(12):1179–1194, 2017.

A Supplementary Methods

A.1 Large parsimony details

Our tree-search algorithm starts with an initial set of candidate trees $S = (\mathcal{T}_1, \dots, \mathcal{T}_k)$ and iteratively improves upon the trees by stochastic perturbations followed by a hill-climbing procedure . Specifically, at each iteration we select a candidate tree uniformly at random and perturb the tree using a random number r of nearest neighbor interchange (NNI) operations. With our perturbed candidate tree, we then perform local optimization using hill climbing to minimize the cost of the tree, where we use our small parsimony algorithm to efficiently evaluate the cost of each candidate topology. Once the hill-climbing procedure reaches a local minimum, we complete the iteration and update the candidate tree set if an improvement was found. The algorithm terminates after no improvement is found for a fixed number I of iterations.

For all experiments and analysis in this paper, the number of iterations prior to termination was set to $I = 150$. The number of random NNIs to perturb the candidate tree is selected uniformly at random from the discrete interval $\{0, 1, \dots, \lfloor 2.5n \rfloor\}$ at every iteration. Our candidate tree set was generated by performing neighbor joining on the boundary insensitive distances and then randomly perturbing the the neighbor joining tree.

A.2 Simulation details

We used a modified version of CONET's [25] copy number phylogeny simulator. Specifically, we found that CONET's simulation of tree structure was non-standard and opted to use a forward-birth death model [30] to simulate our topology. Once the tree structure was generated, we then used CONET's simulator to sample copy number events on each vertex. We then took our event labeled copy number phylogeny and sampled the ground truth copy number states on the leaves of the phylogeny to obtain our copy number profiles.

To generate the tree topology, we used Cassiopeia's [21] implementation of a forward-birth death model. We performed simulations for $n = 100, 150, 200, 250, 600$ leaves with a fitness parameter of 1.3 and an initial birth scale of 0.5. We drew the birth-death waiting times from an exponential distribution. With the topology, we randomly sampled events on each vertex using CONET with $l = 1000, 2000, 3000, 4000$ loci. We performed each simulation with parameters (n, l) a total of 7 times with unique random seeds $s = 0, 1, 2, 3, 4, 5, 6$. In total, there were 140 randomly simulated instances.

A.3 Clonal concordance analysis

To analyze the concordance of the inferred phylogenetic trees with clonal information, we measured the minimum number of evolutionary events required to explain the clones. Specifically, for each sample clones were identified by clustering the GC-corrected read count profiles embedded using UMAP [26, 15]. The clone labels were then attached to the leaves of the inferred phylogenetic trees. With this clone labeled phylogenetic tree, we solved the small parsimony problem under the Wagner [42] model to obtain a parsimony score, p, which we call the *clonal discordance score*. This clonal discordance score is the minimum number of clonal transitions required to explain the cells of the phylogeny.

To compare across different phylogeny sizes, we computed the relative clonal discordance score between the *Lazac* and Sitka phylogenies as

$$r = \frac{p_2 - p_1}{p_1 + p_2},$$

where p_1, p_2 are the clonal discordance scores of the *Lazac* and Sitka phylogenies respectively. In particular, a positive score indicates that the *Lazac* phylogeny is more concordant with the clones while a negative score indicates that the Sitka phylogeny is more concordant with the clones.

A.4 Permutation test for analysis of SNV support

In this section we provide details about how subtrees of the phylogenies are identified for the analysis and the permutation test used for investigate if the subtrees are supported by the SNVs.

First, we describe how we identify subtrees in the phylogeny to analyze. Our goal is to identify subtrees that have enough cells so that pooling the reads from the cells and finding SNVs is feasible. However, if the number of cells is too large, the permutation test will not yield a significantly low p-value. To that end, we perform a breadth-first traversal of the nodes of the tree (starting from the root) to identify the desired subtrees. At each iteration, we compute the number of cells in the subtree rooted at the node (i.e. the number of leaves in the subtree). We select the subtree if (1) the number of cells in the subtree is more than 10% of the total number of cells (2) the number of cells in the subtree is less than 25% of the total number of cells and (3) the subtree is not contained in any of the subtrees selected in previous iterations.

Now, we provide details about the permutation test for a given subtree. We say that an SNV supports a subtree of a phylogeny if all the cells that yield a read harboring the SNV are contained in the subtree. We randomly permute the cell labels 500 times and count the number of SNVs supporting a given subtree. The p-value is empirically estimated by the ratio of the number of instances in which more SNVs support the subtree than with the original unpermuted cell labels with the total number of permutations tested (which is 500 in our study).

A.5 Comparison to simulated trees

We assess the accuracy of the inferred trees compared to the ground truth simulated trees by employing two distinct tree dissimilarity metrics. These metrics are implemented in the TreeCmp tool [5] and the comparisons are done in a similar manner to the comparisons in our *Startle* [37] paper. Our metrics take a ground truth tree, $\mathcal{T}^*$, and an inferred tree, $\mathcal{T}$, both in Newick format [6].

The Robinson-Foulds (RF) distance, $d_{\mathrm{RF}}(\mathcal{T}, \mathcal{T}^*)$, is a tree distance metric based on the induced bi-partitions in the input trees [33, 4]. Each edge $e \in \mathcal{T}$ is associated with a bi-partition $B_e := (X, \bar{X})$ of its leaves, using the equivalence relation $x \sim y$ if x is connected to y in $\mathcal{T}_{-e}$, the forest formed by removing edge e. The set of bi-partitions for a tree $\mathcal{T}$ is $\mathrm{Bip}(\mathcal{T}) = \{B_e : e \in E(\mathcal{T})\}$. The RF distance is then:

$$d_{\mathrm{RF}}(\mathcal{T}, \mathcal{T}*) = |\mathrm{Bip}(\mathcal{T}) \triangle \mathrm{Bip}(\mathcal{T}^*)|.$$

Similarly, the quartet distance, $d_{\mathrm{Q}}(\mathcal{T}, \mathcal{T}^*)$, is a tree distance metric based on the induced quartets in the input trees [33, 4]. We define the set of quartets $Q(\mathcal{T})$ as the set of all consistent 4-leaf sub-trees with the unrooted topology of $\mathcal{T}$. Then,

$$d_{\mathrm{Q}}(\mathcal{T}, \mathcal{T}*) = |Q(\mathcal{T}) \triangle Q(\mathcal{T}^*)|.$$

Finally, we used normalized versions of both d_{RF} and d_{Q} to enable comparison across different parameter settings. This normalization is implemented in TreeCmp [5] and described in their paper.

B Supplementary Results

B.1 ZCNT small parsimony: dropping the integrality condition

Let Q be the delta matrix obtained by applying the delta transformation to each row of the copy number matrix M (i.e. $q_{ij} = \Delta(m_i)_j$). Using the formulation of the ZCNT small parsimony problem as stated in (Problem 3), we can write the objective as the following mathematical program.

$$\begin{aligned} \min_{\ell} \quad & \sum_{(u,v)\in E(\mathcal{T})} \frac{1}{2}\|\ell(u) - \ell(v)\|_1 \\ \text{s.t.} \quad & l(u) \in \mathbb{Z}^m \quad \text{for all } u \in V(\mathcal{T}), \\ & \sum_{i=1}^{n} l(u)_i = 0 \quad \text{for all } u \in V(\mathcal{T}), \\ & l(\pi(i))_j = Q_{ij} \quad \text{for all } i \in [n], j \in [m]. \end{aligned}$$

Notice that we can rewrite the optimization objective as a linear function subject to additional constraints. Specifically, $\|\ell(u)-\ell(u)\|_1 = \sum_{j=1}^{m}(x^+_{uvj}-x^-_{uvj})$ when $x^+_{uvj} = \max\{\ell(u)_j, \ell(v)_j\}$ and $x^-_{uvj} = \min\{\ell(u)_j, \ell(v)_j\}$. And we can set x^+_{uvj} using the two linear constraints $x^+_{uvj} \geq \ell(u)_j$ and $x^+_{uvj} \geq \ell(v)_j$; a similar procedure works for x^-_{uvj}. Then, by dropping the integrality condition $\ell(u) \in \mathbb{Z}^m$, we obtain the following equivalent linear program.

$$\begin{aligned} \min_{x,\ell} \quad & \frac{1}{2}\sum_{(u,v)\in E(\mathcal{T})}\sum_{j=1}^{m}(x^+_{uvj} - x^-_{uvj}) & \\ \text{s.t.} \quad & \sum_{j=1}^{m} \ell(u)_j = 0 & \text{for all } u \in V(\mathcal{T}), \\ & \ell(\pi(i))_j = Q_{ij} & \text{for all } i \in [n] \text{ and } j \in [m], \\ & x^+_{uvj} \geq \ell(u)_j \text{ and } x^+_{uvj} \geq \ell(v)_j & \text{for all } (u,v) \in E(\mathcal{T}) \text{ and } j \in [m], \\ & x^-_{uvj} \leq \ell(u)_j \text{ and } x^-_{uvj} \leq \ell(v)_j & \text{for all } (u,v) \in E(\mathcal{T}) \text{ and } j \in [m]. \end{aligned}$$

Since we can solve linear programs in (weakly) polynomial time, this proves the following theorem.

Theorem 4. *The ZCNT small parsimony problem can be solved in (weakly) polynomial time when the constraint that $\ell(u) \in \mathbb{Z}^m$ is relaxed to $\ell(u) \in \mathbb{R}^m$ using a linear program with $O(mn)$ variables and $O(mn)$ constraints.*

B.2 ZCNT small parsimony: dropping the balancing condition

When we drop the balancing condition, our problem becomes equivalent to the fixed topology rectilinear Steiner tree problem [36] on the delta profiles where the ancestral nodes lie in $\mathbb{Z}^m$. While there are several algorithms for the unrooted variant of this problem when Steiner vertices are in $\mathbb{R}^m$ [13, 36, 9], our problem is different in that i) it assumes a rooted topology ii) the Steiner vertices are required to lie in $\mathbb{Z}^m$. Further, this problem has not been recently studied and we believe deserves a modern treatment. In this section, we present and prove the correctness of a linear time dynamic programming algorithm that solves the ZCNT small parsimony problem without the balancing condition.

We first observe that it suffices to analyze each locus independently. Let Q be the delta matrix obtained by applying the delta transformation to each row of the copy number matrix M (i.e. $q_{ij} = \Delta(m_i)_j$). Let ℓ_j minimize the quantity $\sum_{(u,v)\in E(\mathcal{T})} |\ell_j(u) - \ell_j(v)|$ and agree with the delta matrix Q on the leaves; that is,

Algorithm 1 ZCNT small parsimony without the balancing condition

Require: A binary tree $\mathcal{T}$ rooted at vertex v, a delta matrix Q, an assignment π of cells to leaves, and a locus j.

Output: A minimizer of the cost $J(\ell_j, \mathcal{T})$ for locus j that agrees with Q on the leaves of $\mathcal{T}$.

```
if $v$ is a leaf in $\mathcal{T}$ then
    set $\ell_j(v) \leftarrow [q_{\pi(v),j}, q_{\pi(v),j}]$
    return
end if

get $w, z \leftarrow$ children$(v)$
recurse at nodes $w$ and $z$
set $l_j(v) \leftarrow \text{Sank}(l_j(w), l_j(v))$
```

$\ell_j(\pi(i)) = q_{ij}$ for all cells i. Then, the labeling defined as $\ell(u) = (\ell_1(u), \ldots, \ell_m(u))$ minimizes $J(\ell, \mathcal{T})$:

$$\begin{aligned}
\min_{\hat{\ell}} J(\hat{\ell}, \mathcal{T}) &= \min_{\hat{\ell}} \sum_{(u,v) \in E(\mathcal{T})} \frac{1}{2} \left\| \hat{\ell}(u) - \hat{\ell}(v) \right\|_1 \\
&= \min_{\hat{\ell}} \left(\frac{1}{2} \sum_{i=1}^{m} \sum_{(u,v) \in E(\mathcal{T})} \left| \hat{\ell}_i(u) - \hat{\ell}_i(v) \right| \right) \\
&\geq \frac{1}{2} \sum_{i=1}^{m} \min_{\hat{\ell}} \left(\sum_{(u,v) \in E(\mathcal{T})} \left| \hat{\ell}_i(u) - \hat{\ell}_i(v) \right| \right) \\
&= J(\ell, \mathcal{T}).
\end{aligned}$$

Thus, we can compute the cost of the optimal labeling ℓ_j for each locus independently and sum them together to obtain the entire cost. We also introduce the quantity $c(\mathcal{T}; x)$ as the cost of the optimal labeling $\hat{\ell}$ of $\mathcal{T}$ that agrees with ℓ on the leaves of $\mathcal{T}$ and has root label x. Our algorithm relies on the following easy to compute function on discrete intervals denoted $[a, b] = \{a, a+1, \ldots, b-1, b\}$:

$$\text{Sank}([a,b],[c,d]) = \begin{cases} [a,b] \cap [c,d] & \text{if } [a,b] \cap [c,d] \neq \{\}, \\ [b,c] & \text{if } b \leq c, \\ [d,a] & \text{otherwise if } d \leq a. \end{cases}$$

Our algorithm then applies the Sank$(\cdot,\cdot)$ function in a top-down recursive fashion, to compute *the interval* of optimal root labelings for each node. Though this procedure is quite natural, its proof of correctness is not immediately obvious and relies on a technical (Lemma 3).

Theorem 5. *(Algorithm 1) solves the ZCNT SCNP problem in $O(n)$ time for a single locus when the balancing condition is dropped.*

Proof. Follows from induction on the size of the tree using (Lemma 3). ■

To prove the correctness of our procedure, we introduce the function dist$(x, [a, b])$ defined as the distance from x to the discrete interval $[a, b]$:

$$\text{dist}(x,[a,b]) = \begin{cases} 0 & \text{if } x \in [a,b] \\ \min\{|x-a|, |x-b|\} & \text{otherwise.} \end{cases}$$

The correctness of our algorithm relies on two technical lemmas whose proofs are in (Section C).

Lemma 1. *If $a \leq b < c \leq d$ are integers, then for all integers x the following inequality holds:*

$$\min_{y,z\in\mathbb{Z}} \left(dist(y, [a, b]) + dist(z, [c, d]) + |x - y| + |x - z| \right) \geq (c - b) + dist(x, [b, c]).$$

Lemma 2. *If $a \leq b$, $c \leq d$, $a \leq c$, and $b \geq c$ are integers, then for all integers x the following inequality holds:*

$$\min_{y,z\in\mathbb{Z}} \left(dist(y, [a, b]) + dist(z, [c, d]) + |x - y| + |x - z| \right) \geq dist(x, [c, b]).$$

Then, the correctness of our algorithm follows from induction using the following lemma.

Lemma 3. *Let $\mathcal{T}$ be a tree whose root vertex v has two children v_1 and v_2. Let $\mathcal{T}_{v_1}$ and $\mathcal{T}_{v_2}$ denote the sub-trees rooted at v_1 and v_2 respectively. Then, suppose that*

$$c(\mathcal{T}_{v_1}; x) \geq c(\mathcal{T}_{v_1}) + dist(x, [a, b])$$
$$\text{and } c(\mathcal{T}_{v_2}; x) \geq c(\mathcal{T}_{v_2}) + dist(x, [c, d])$$

where $[a, b]$ and $[c, d]$ are the set of optimal root labelings for $\mathcal{T}_{v_1}$ and $\mathcal{T}_{v_2}$. Then, $[e, f] = \textsc{Sank}([a, b], [c, d])$ is the optimal set of root labelings for $\mathcal{T}$ and

$$c(\mathcal{T}; x) \geq c(\mathcal{T}_{v_1}) + c(\mathcal{T}_{v_2}) + dist(x; [e, f]) + \mathbb{1}[b < c](c - b).$$

Proof. Without loss of generality we can assume that $a \leq c$. Otherwise, we can swap the names of vertices v_1 and v_2. Let x be the labeling of the root of $\mathcal{T}$. Then,

$$\begin{aligned} c(\mathcal{T}; x) &= \min_{y,z\in\mathbb{Z}} \left[c(\mathcal{T}_{v_1}; y) + c(\mathcal{T}_{v_2}; z) + |x - y| + |y - z| \right] \\ &\geq \min_{y,z\in\mathbb{Z}} \left[c(\mathcal{T}_{v_1}) + \text{dist}(y; [a, b]) + c(\mathcal{T}_{v_2}) + \text{dist}(z; [c, d]) + |x - y| + |y - z| \right] \end{aligned}$$

where the first equality follows from the definition of $c(\cdot; \cdot)$ and the fact that the distance is ℓ_1. And the first inequality follows from our assumptions about $c(\mathcal{T}_{v_1}; y)$ and $c(\mathcal{T}_{v_2}; z)$.

We now consider the two cases of $\textsc{Sank}([a, b], [c, d])$ separately.

Case 1: $b < c$. In this case, we know from definition of $\textsc{Sank}()$ that $\textsc{Sank}([a, b], [c, d]) = [b, c]$. We want to show that

$$\begin{aligned} &\min_{y,z\in\mathbb{Z}} \left(\text{dist}(y, [a, b]) + \text{dist}(z, [c, d]) + |x - y| + |x - z| \right) \\ &\qquad \geq (c - b) + \text{dist}(x, [b, c]). \end{aligned}$$

This will prove the desired inequality of the theorem. Then, to see that $[b, c]$ is the optimal labeling of the root it is enough to observe that the inequality is realized when $x \in [b, c]$. As the proof of this inequality is rather technical and unenlightening, it is summarized in in (Lemma 1) and proven in Supplementary Proofs.

Case 2: $\textsc{Sank}([a, b], [c, d]) = [c, b]$ and $c \leq b$. In this case, we want to show that

$$\begin{aligned} &\min_{y,z\in\mathbb{Z}} \left(\text{dist}(y, [a, b]) + \text{dist}(z, [c, d]) + |x - y| + |x - z| \right) \\ &\qquad \geq \text{dist}(x, [b, c]). \end{aligned}$$

Which will again prove the desired inequality of the theorem. Then, to see that $[c, d]$ is the optimal labeling of the root it is enough to observe that the inequality is realized when $x \in [c, d]$. The proof of this inequality is given in (Lemma 2). ■

C Supplementary Proofs

Theorem 1. *Let $c_{s,t,b}$ be a zero-agnostic copy number event and $\delta_{s,t,b}$ be a delta event. Then,*

$$p' = c_{s,t,b}(p) \quad \textit{if and only if} \quad \Delta(p') = \delta_{s,t,b}(\Delta(p)).$$

Proof. ($\Rightarrow$) Let p' be the result of applying the zero-agnostic copy number event $c_{s,t,b}$ to the profile p. Then, if $i, i-1 \in \{s, \ldots, t\}$:

$$\Delta(p')_i = (p_i + b) - (p_{i-1} + b) = p_i - p_{i-1} = \Delta(p)_i.$$

Similarly, if $i, i-1 \notin \{s, \ldots, t\}$

$$\Delta(p')_i = p_i - p_{i-1} = \Delta(p)_i.$$

The remaining two cases occur when either $i = s$ or $i - 1 = t$.

$$\begin{aligned} \Delta(p')_i &= (p_i + b) - p_{i-1} = \Delta(p)_i + b && \text{if} \quad i = s, \\ \Delta(p')_i &= p_i - (p_{i-1} + b) = \Delta(p)_i - b && \text{if} \quad i - 1 = t \end{aligned}$$

Thus, $\Delta(p')$ is the result of applying the delta event δ to the profile $\Delta(p)$.

($\Leftarrow$) This case is handled symmetrically. ■

Proposition 3. *$d'(u, v)$ is a distance metric. Further,*

$$d'(u, v) = d'(v - u, 0) = d'(u - v, 0).$$

Proof. To see that $d'(\cdot, \cdot)$ satisfies the triangle inequality, it suffices to observe that the composition of delta sequences taking u to v and v to w transforms u to w. It is clearly reflexive since no delta event needs to be applied to map u to itself.

To see symmetry and the above equality, we observe that something stronger holds. Let

$$\gamma(D)_i := \sum_{i=1}^{n} (b_i * \mathbb{1}[i = s_i] - b_i * \mathbb{1}[i = t_i])$$

be the net change of coordinate i induced by the delta sequence $D = (\delta_1, \ldots, \delta_n)$ where $\delta_i = (s_i, t_i, b_i)$. Then, D takes u to v if and only if $\gamma(D)_i = u_i - v_i$ for all $i \in \{1, \ldots, n\}$. This follows from the definition of our delta event and the fact that applying D to u results in a profile v defined by its entries as $v_i = u_i + \gamma(C)_i$. ■

Proposition 4. *For a vector $p' = \Delta(p)$, the sum of the magnitudes of the positive entries equals the sum of the magnitudes of the negative entries. That is,*

$$\sum_{p'_i > 0} |p'_i| = \sum_{p'_i < 0} |p'_i|. \tag{3}$$

Proof. We first notice that $\sum_{i=0}^{n+1} \Delta(p)_i$ expands to a telescoping sum after applying our definition of a delta profile. Specifically,

$$\begin{aligned} \sum_{i=0}^{n+1} \Delta(p)_i &= (p_0 - 2) + (2 - p_n) + \sum_{i=1}^{n} (p_i - p_{i-1}) \\ &= (p_0 - p_n) + (p_n - p_0) = 0. \end{aligned}$$

Then, we rewrite the left hand side of the sum

$$\begin{aligned}\sum_{i=0}^{n+1} \Delta(p)_i &= \sum_{\Delta(p)_i>0} \Delta(p)_i + \sum_{\Delta(p)_i<0} \Delta(p)_i \\ &= \sum_{\Delta(p)_i>0} \Delta(p)_i - \sum_{\Delta(p)_i<0} -\Delta(p)_i \\ &= \sum_{\Delta(p)_i>0} |\Delta(p)_i| - \sum_{\Delta(p)_i<0} |\Delta(p)_i|,\end{aligned}$$

where the last equality follows from the definition of absolute value. ■

Proposition 1. *The delta map* $\Delta : \mathbb{Z}^n \to \mathcal{D}_{n+1}$ *is invertible.*

Proof. By (Proposition 4) the range of the map Δ lies in the space of vectors in $\mathbb{Z}^{n+1}$ satisfying the balance condition. Thus, it suffices to show that Δ is injective and can reach all vectors (i.e. is surjective) in this space.

To see that the map is injective, let $x, y \in \mathbb{Z}^n$ be two distinct vectors. Then, define i as the first coordinate in which these vectors differ. Since i is the first coordinate in which these vectors differ, $x_{i-1} = y_{i-1}$ but $x_i \neq y_i$. Thus,

$$\Delta(x)_i = x_i - x_{i-1} \neq y_i - y_{i-1} = \Delta(y)_i,$$

proving that Δ is injective.

To see the map is surjective, let $y \in \mathbb{Z}^{n+1}$ be a vector satisfying the balance condition. Define a vector $x \in \mathbb{Z}^n$ such that $x_0 = y_0 + 2$, and $x_i = y_i + x_{i-1}$ for $i \in \{1, \ldots, n\}$. Then, $\Delta(x)$ agrees with y on the first n coordinates, but since y and $\Delta(x)$ both satisfy the balance condition by (Proposition 4), their last coordinate is also determined, proving that $\Delta(x) = y$. ■

Corollary 3. *Given two copy number profiles* u *and* v*, both* u *and* v *minimize the median distance* $d(w, u) + d(w, v)$ *over all choices of copy number profiles* w*. Thus,*

$$\min_{w \in \mathbb{Z}_+^m} \{d(w, u) + d(w, v)\} = d(u, v).$$

Proof. By the triangle inequality, $d(u, w) + d(w, v) \geq d(u, v)$ for all profiles w. Setting $w = u$ or $w = v$ achieves equality, proving that setting w to either u or v minimizes the expression $d(w, u) + d(w, v)$ as $d(u, w) = d(w, u)$. ■

Lemma 1. *If* $a \leq b < c \leq d$ *are integers, then for all integers* x *the following inequality holds:*

$$\min_{y,z \in \mathbb{Z}} (dist(y, [a, b]) + dist(z, [c, d]) + |x - y| + |x - z|) \geq (c - b) + dist(x, [b, c]).$$

Proof. The statement follows from a case analysis on the locations of the variables we are optimizing over: y, z. However, before any case analysis, we prove that

$$\text{dist}(x, [a, b]) + \text{dist}(x, [c, d]) \geq \text{dist}(x, [b, c]) + (c - b). \tag{4}$$

To see this, we perform a case analysis on x. If $x \leq b$ we have

$$\text{dist}(x, [c, d]) = \text{dist}(x, [b, c]) + (c - b),$$

since x is to the left of the interval $[b, c]$ which is to the left of the interval $[c, d]$. If $x \geq c$ we have

$$\text{dist}(x, [a, b]) = \text{dist}(x, [b, c]) + (c - b),$$

since x is to the right of the interval $[b, c]$ which is to the right of the interval $[a, b]$. And finally, if $x \in [b, c]$ then,

$$\text{dist}(x, [a, b]) + \text{dist}(x, [c, d]) = (c - b) = \text{dist}(x, [b, c]) + (b - c),$$

since $\text{dist}(x, [b, c]) = 0$.

Case 1: $y \in [a, b]$ and $z \in [c, d]$

In this case, $\text{dist}(y, [a, b]) + \text{dist}(z, [c, d]) = 0$, so

$$\begin{aligned}
&\min_{\substack{y\in[a,b]\\ z\in[c,d]}} (\text{dist}(y, [a, b]) + \text{dist}(z, [c, d]) + |x - y| + |x - z|)\\
&= \min_{\substack{y\in[a,b]\\ z\in[c,d]}} (|x - y| + |x - z|)\\
&= \text{dist}(x, [a, b]) + \text{dist}(x, [c, d])\\
&\geq (c - b) + \text{dist}(x, [b, c]),
\end{aligned}$$

where the second equality follows from the fact that $\text{dist}(x, [a, b]) = \min_{y\in[a,b]} |x - y|$ and the inequality follows from (4).

Case 2: $y \in [a, b]$ and $z \notin [c, d]$ OR $y \notin [a, b]$ and $z \in [c, d]$

Since the two cases are symmetric, we only need to consider the former situation where $y \in [a, b]$ and $z \notin [c, d]$. In this case,

$$\begin{aligned}
&\min_{\substack{y\in[a,b]\\ z\notin[c,d]}} (\text{dist}(y, [a, b]) + \text{dist}(z, [c, d]) + |x - y| + |x - z|)\\
&= \min_{\substack{y\in[a,b]\\ z\notin[c,d]}} (\text{dist}(z, [c, d]) + |x - y| + |x - z|)\\
&= \text{dist}(x, [a, b]) + \min_{z\notin[c,d]} (\text{dist}(z, [c, d]) + |x - z|)\\
&= \text{dist}(x, [a, b]) + \min_{z\notin[c,d]} (\min\{|z - c|, |z - d|\} + |x - z|)
\end{aligned}$$

where the second equality follows the fact that $\text{dist}(x, [a, b]) = \min_{y\in[a,b]} |x - y|$ and the third equality from the definition of $\text{dist}(\cdot, \cdot)$. We now analyze two sub-cases separately.

Sub-case 1: $x \in [c, d]$ In this case, $\text{dist}(x, [a, b]) = \text{dist}(x, [b, c]) + (c - b)$ and we are done.

Sub-case 2: $x \notin [c, d]$

Then,

$$\min_{z\notin[c,d]} (\min\{|z - c|, |z - d|\} + |x - z|) = \text{dist}(x, [c, d]),$$

since the minimizer is found by setting $z = x$. From the above equality,

$$\begin{aligned}
&\min_{\substack{y\in[a,b]\\ z\notin[c,d]}} (\text{dist}(y, [a, b]) + \text{dist}(z, [c, d]) + |x - y| + |x - z|)\\
&= \text{dist}(x, [a, b]) + \text{dist}(x, [c, d])\\
&\geq \text{dist}(x, [b, c]) + (c - b),
\end{aligned}$$

where the inequality again follows from (4).

Case 3: $y \notin [a,b]$ and $z \notin [c,d]$

In this case, by the definition of $\text{dist}(\cdot,\cdot)$

$$\begin{aligned}
&\min_{\substack{y\notin[a,b]\\ z\notin[c,d]}} (\text{dist}(y,[a,b]) + \text{dist}(z,[c,d]) + |x-y| + |x-z|) \\
&= \min_{\substack{y\notin[a,b]\\ z\notin[c,d]}} (\min\{|y-a|,|y-b|\} + \min\{|z-c|,|z-d|\} \\
&\qquad\qquad + |x-y| + |x-z|).
\end{aligned}$$

We now analyze two sub-cases separately.

Sub-case 1: $x \notin [a,b]$ and $x \notin [c,d]$

Now, by the same reasoning as before,

$$\begin{aligned}
&\min_{y\notin[a,b]} (\min\{|y-a|,|y-b|\} + |x-y|) = \text{dist}(x,[a,b]) \\
\text{and } &\min_{z\notin[c,d]} (\min\{|z-c|,|z-d|\} + |x-z|) = \text{dist}(x,[c,d]),
\end{aligned}$$

since the minimizer is found by setting $y = x$ and $z = x$. Thus, by independence of the terms y and z,

$$\begin{aligned}
&\min_{\substack{y\notin[a,b]\\ z\notin[c,d]}} (\text{dist}(y,[a,b]) + \text{dist}(z,[c,d]) + |x-y| + |x-z|) \\
&= \text{dist}(x,[a,b]) + \text{dist}(x,[c,d]) \\
&\geq \text{dist}(x,[b,c]) + (c-b),
\end{aligned}$$

where the second inequality follows from (4).

Sub-case 2: $x \in [a,b]$ OR $x \in [c,d]$

By symmetry, without loss of generality we assume that $x \in [a,b]$. Since the two intervals are disjoint $x \in [a,b]$ implies $x \notin [c,d]$. Then since $x \notin [c,d]$,

$$\min_{z\notin[c,d]} (\min\{|z-c|,|z-d|\} + |x-z|) = \text{dist}(x,[c,d])$$

since the minimizer is realized by setting $z = x$. Finally since $\text{dist}(x,[a,b]) = 0$,

$$\begin{aligned}
&\min_{\substack{y\notin[a,b]\\ z\notin[c,d]}} (\text{dist}(y,[a,b]) + \text{dist}(z,[c,d]) + |x-y| + |x-z|) \\
&\geq \text{dist}(x,[c,d]) = \text{dist}(x,[a,b]) + \text{dist}(x,[c,d]) \\
&\geq \text{dist}(x,[b,c]) + (c-b),
\end{aligned}$$

where the second inequality follows from (4). As this is the final case, we have proven the original claim. ■

Lemma 2. *If $a \leq b$, $c \leq d$, $a \leq c$, and $b \geq c$ are integers, then for all integers x the following inequality holds:*

$$\min_{y,z\in\mathbb{Z}} (\mathit{dist}(y,[a,b]) + \mathit{dist}(z,[c,d]) + |x-y| + |x-z|) \geq \mathit{dist}(x,[c,b]).$$

Proof. The proof again follows by a case analysis, but it is much simpler than in (Lemma 1).

Case 1: $x \in [c, b]$

This case is trivial since $\text{dist}(x, [c, b]) = 0$ and the left hand side of the inequality is always non-negative.

Case 2: $x \leq c$ or $x \geq b$

As these cases are symmetric, it suffices to only consider the former case where $x \leq c$. First, if $z \leq c$,

$$\begin{aligned} \text{dist}(z, [c, d]) + |x - z| &\geq |x - c| \\ &\geq \text{dist}(x, [c, b]) \end{aligned}$$

since the minimizer is found when $z \in [x, c]$. Second, if $z \geq c$,

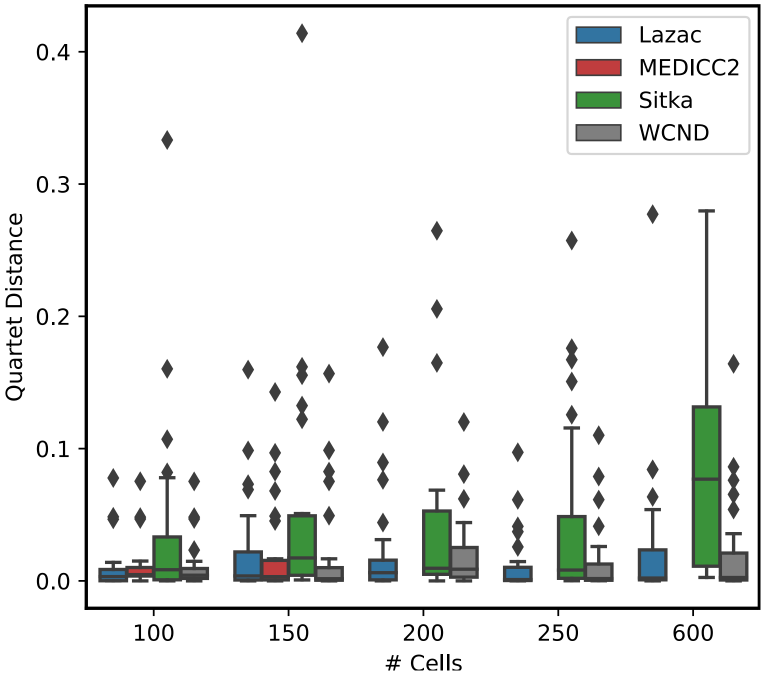
$$|x - z| = z - x \geq |c - x| \geq \text{dist}(x, [c, b]),$$

since z is to the right of c while x is to the left of c. This completes the proof. ■

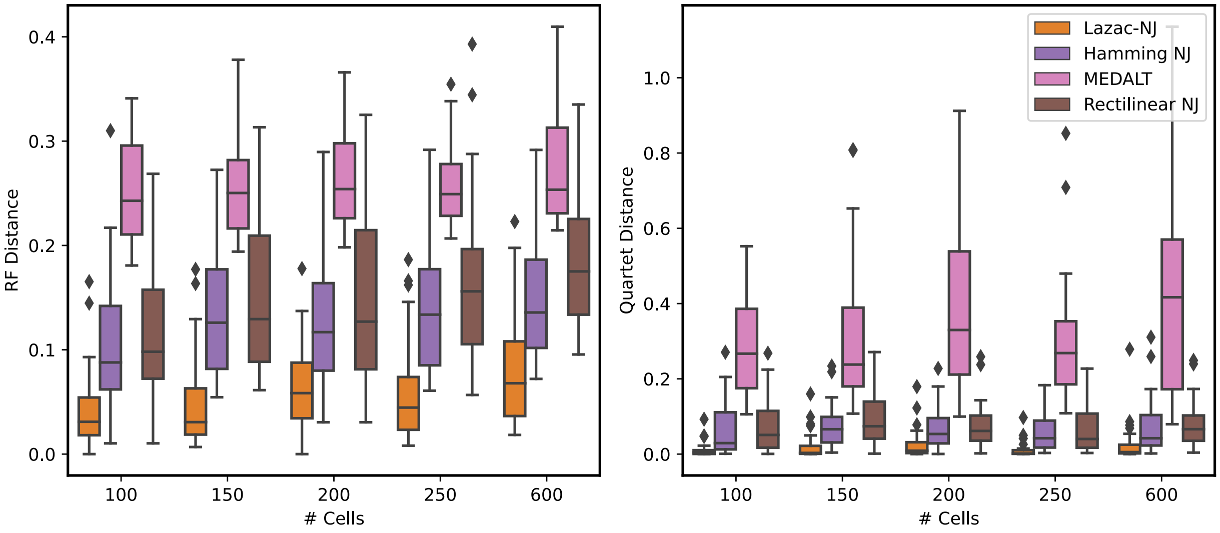
D Supplementary Figures

We have the following supplementary figures and tables:

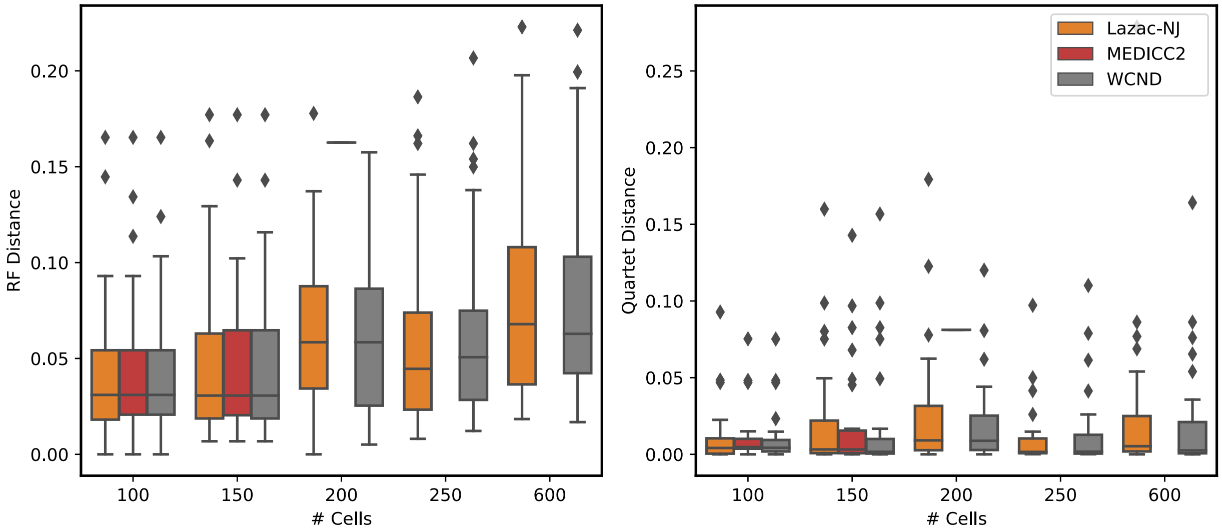
- (Supplementary Figure 1) compares the reconstruction accuracy in terms of Quartet distance on CONET simulated data across state-of-the-art methods.
- (Supplementary Figure 2) compares the baseline reconstruction accuracy for a variety of extremely simple distance based methods including our distance method, *Lazac NJ*.
- (Supplementary Figure 3) compares the reconstruction accuracy on CONET simulated data for a variety of distance based methods including our distance method, *Lazac NJ*.
- (Supplementary Figure 4) compares the clonal discordance scores between *Lazac* and Sitka inferred phylogenies from an in vitro cell line system modeling instability in human and breast ovarian tumours.
- (Supplementary Figure 5) displays the relative difference between the ZCNT and CNT distance for patient 008 from a metastatic prostate cancer tumor sample [17].
- (Supplementary Figure 6) displays the relative difference between the ZCNT and CNT distance for patient 012 from a metastatic prostate cancer tumor sample [17].
- (Supplementary Figure 7) compares the solutions to the relaxed and unrelaxed ZCNT small parsimony problem across 200 phylogenies.
- (Supplementary Figure 8) compares the solutions to the relaxed and unrelaxed ZCNT small parsimony problem across 200 phylogenies as a function of the number of stochastic perturbations used to obtain that phylogeny.
- (Supplementary Figure 9) shows the distribution of normalized RF distance between trees inferred by *Lazac* and trees inferred by Sitka on human breast and ovarian tumour data [15].
- (Supplementary Figure 10) a table displaying the results of our concordance analysis on trees inferred by *Lazac* and trees inferred by Sitka on human breast and ovarian tumour data [15].
- (Supplementary Figure 11) Results of somatic SNV analysis on subset of samples from an in vitro cell line system modeling instability in human breast and ovarian tumours [15].



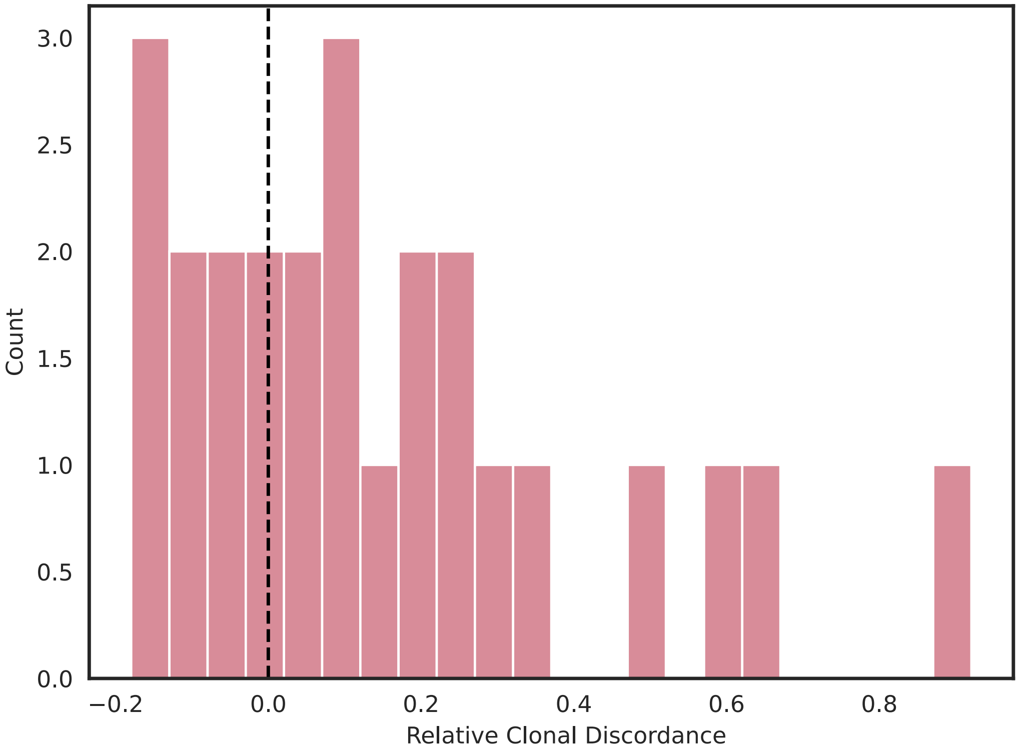
Supplementary Figure 1: Baseline reconstruction accuracy (Quartet distance) on CONET simulated data for copy number tree reconstruction with varying number of cells $n = 100, 150, 200, 250, 600$ across four sets of loci $l = 1000, 2000, 3000, 4000$ and seven random seeds $s = 0, 1, 2, 3, 4, 5, 6$.



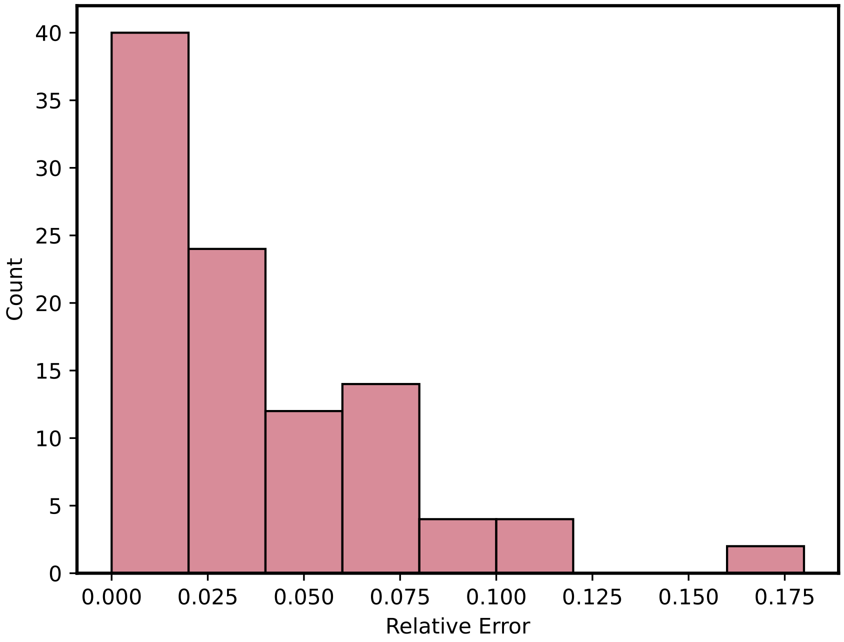
Supplementary Figure 2: Baseline reconstruction accuracy on (left: RF distance; right: Quartet distance) CONET simulated data across simple NJ methods for copy number tree reconstruction with varying number of cells $n = 100, 150, 200, 250, 600$ across four sets of loci $l = 1000, 2000, 3000, 4000$ and seven random seeds $s = 0, 1, 2, 3, 4, 5, 6$.



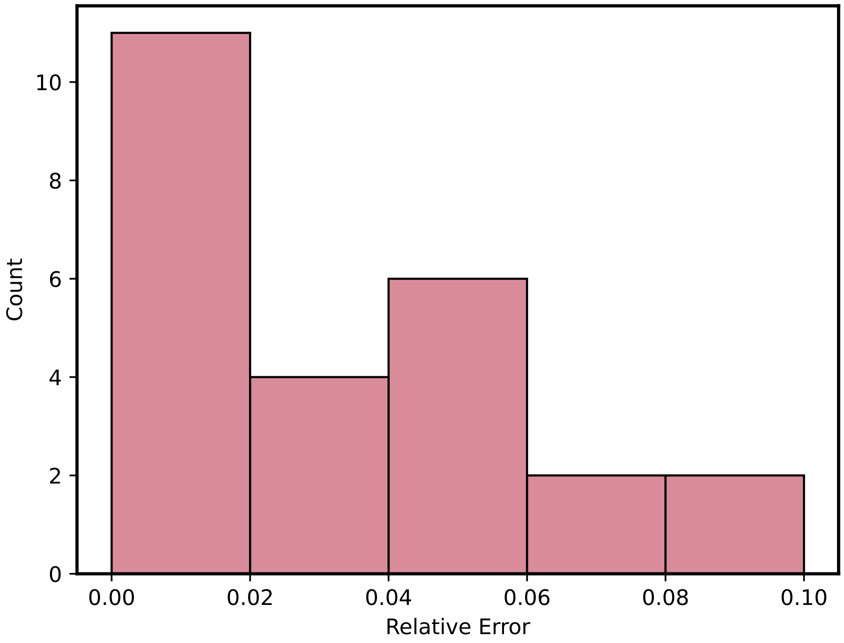
Supplementary Figure 3: Comparison of reconstruction accuracy (left: RF distance; right: Quartet distance) CONET simulated data across *distance based* methods for copy number tree reconstruction with varying number of cells $n = 100, 150, 200, 250, 600$ across four sets of loci $l = 1000, 2000, 3000, 4000$ and seven random seeds $s = 0, 1, 2, 3, 4, 5, 6$. As MEDICC2 was too slow to run on more than 150 cells, we exclude it from comparisons where the number of cells $n > 150$.



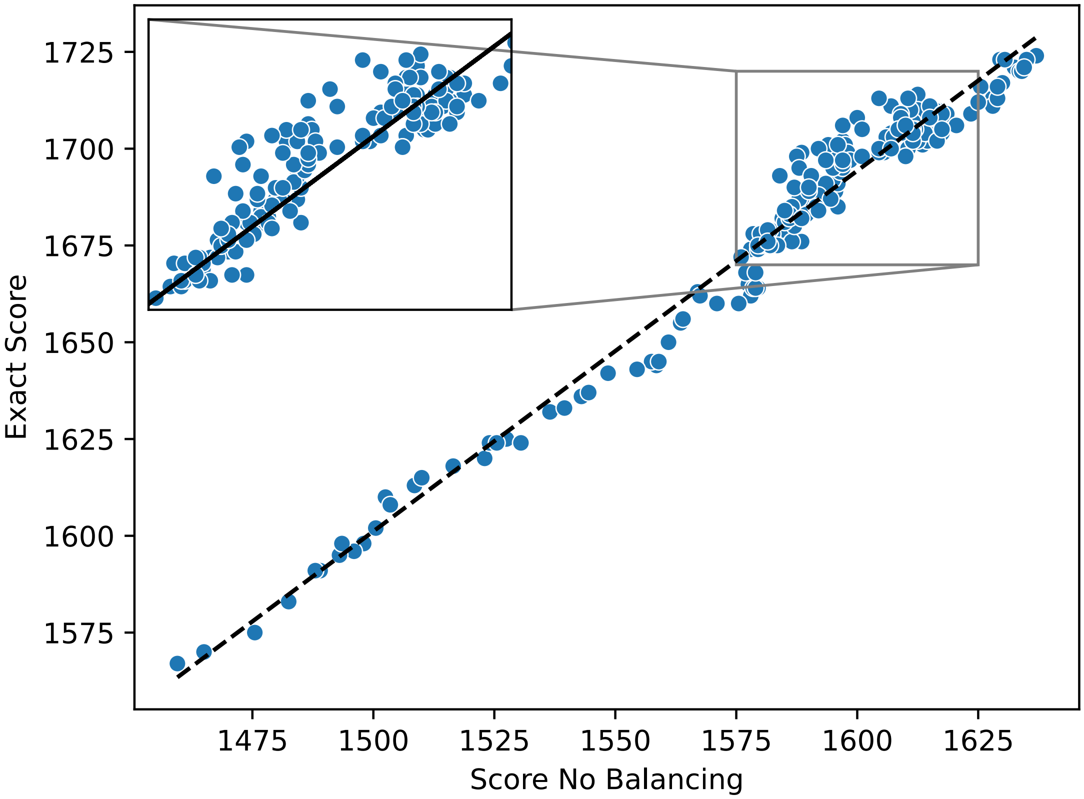
Supplementary Figure 4: The relative clonal discordance score $\frac{p_2 - p_1}{p_1 + p_2}$ where p_1, p_2 are the clonal discordance scores of the *Lazac* and *Sitka* inferred phylogenies respectively.



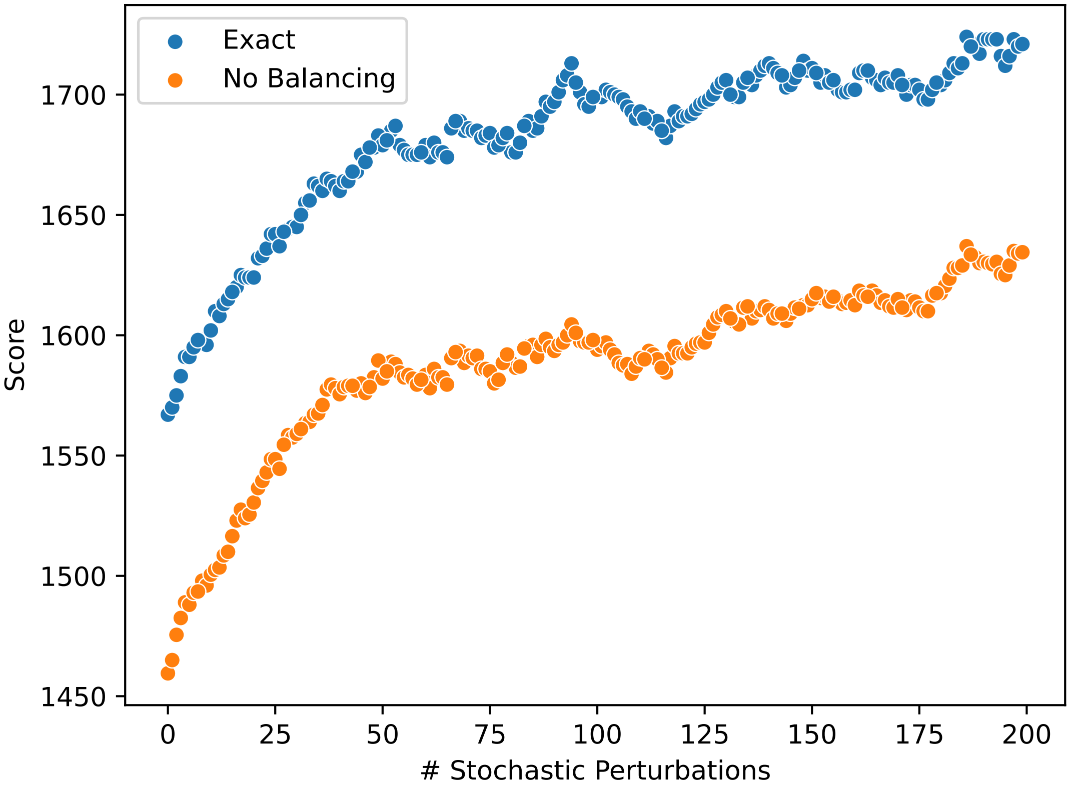
Supplementary Figure 5: Relative difference between the ZCNT distance $d(p, p')$ and the CNT distance computed for patient 8 from a metastatic prostate cancer tumor sample [17].



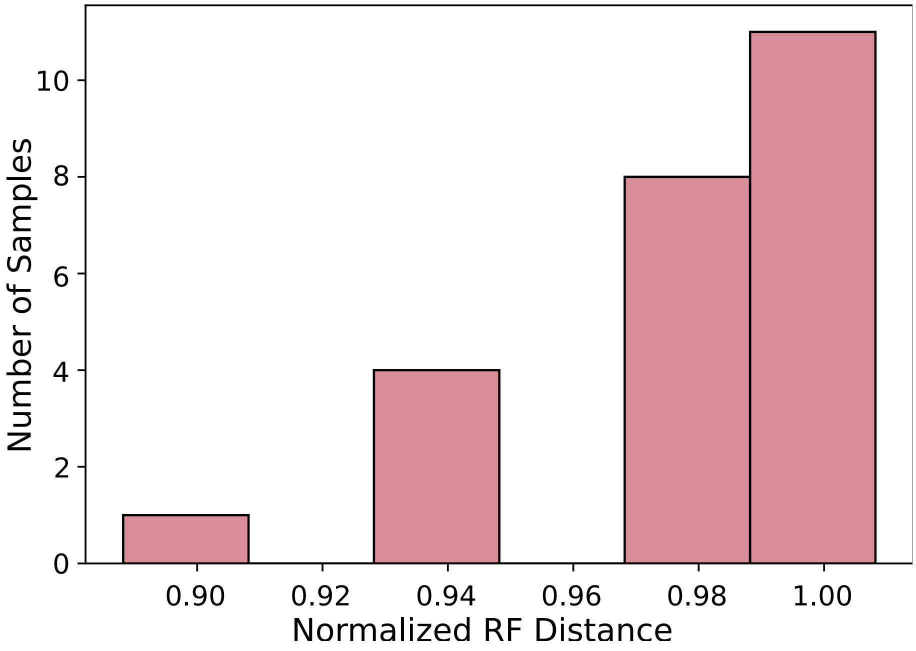
Supplementary Figure 6: Relative difference between the ZCNT distance $d(p, p')$ and the CNT distance for patient 12 from a metastatic prostate cancer tumor sample [17].



Supplementary Figure 7: The exact versus the relaxed score of the optimal solution to the ZCNT small parsimony problem when the balancing condition is removed across 200 phylogenies. The 200 phylogenies were obtained by stochastic perturbation of the phylogeny inferred Sitka [35] on sample SA1053. The dotted line is computed by performing linear regression and is defined by $y = 0.9313 * x + 204.1$ with an $R^2 = 0.972$ and a $p = 1.05 * 10^{-156}$.



Supplementary Figure 8: The exact and relaxed scores of the optimal solution to the ZCNT small parsimony problem as a function of the number of stochastic perturbations applied to the phylogeny inferred by Sitka [35] on sample SA1053.



Supplementary Figure 9: RF distances between *Lazac* and Sitka inferred trees on copy number profiles from human breast and ovarian tumour data [15].

Sample	*Lazac* Clonal Discordance	Sitka Clonal Discordance
SA039	157.0	632.0
SA1035	352.0	1667.0
SA1049	89.0	178.0
SA1050	154.0	142.0
SA1051	53.0	47.0
SA1052	45.0	34.0
SA1053	36.0	29.0
SA1054	35.0	35.0
SA1055	77.0	96.0
SA1056	43.0	30.0
SA1091	83.0	110.0
SA1093	60.0	167.0
SA1096	46.0	32.0
SA1142	17.0	25.0
SA1162	57.0	70.0
SA1180	135.0	219.0
SA1181	85.0	95.0
SA1182	15.0	12.0
SA1184	95.0	117.0
SA501	88.0	101.0
SA530	83.0	83.0
SA535	56.0	923.0
SA604	601.0	1141.0
SA605	7.0	10.0
SA610	8.0	13.0

Supplementary Figure 10: Comparison of *Lazac* and Sitka clonal discordance scores across 25 samples from an in vitro cell line system modeling instability in human breast and ovarian tumours [15].

Patient	# Cells	*Lazac*		Sitka	
		# Significant Subtrees	# Subtrees	# Significant Subtrees	# Subtrees
SA039	1963	4	5	0	1
SA604	2139	7	11	3	5
SA1035	4750	5	6	3	8

Supplementary Figure 11: Results of somatic SNV analysis on subset of samples from an in vitro cell line system modeling instability in human breast and ovarian tumours [15]. Subtrees were called significant if the SNV permutation test p-value was below 0.05 for that clone.