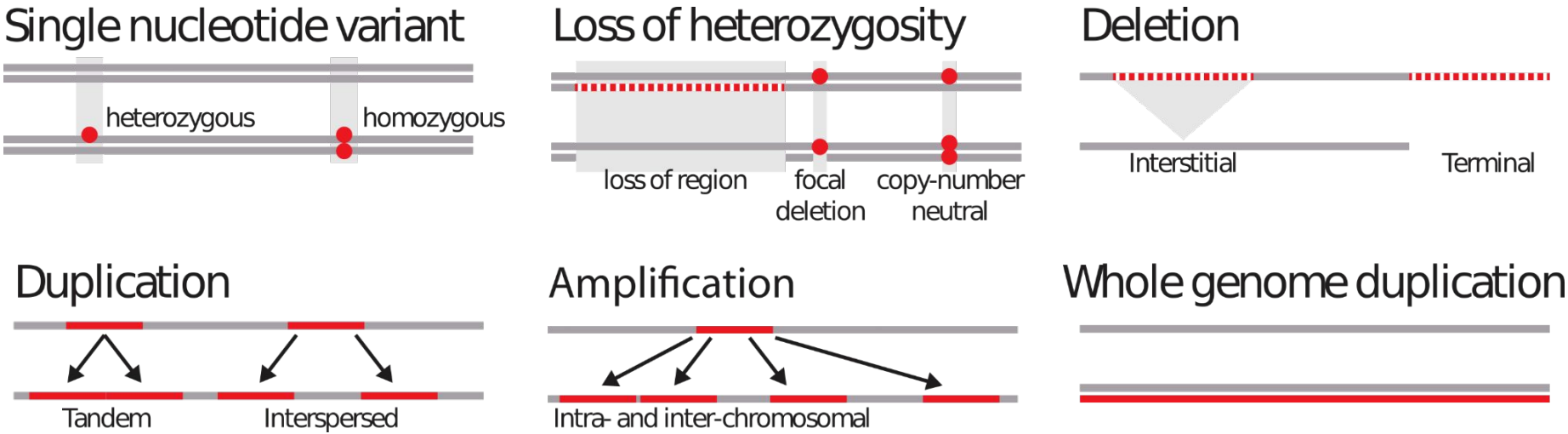
A zero agnostic model for copy number evolution in cancer

Henri Schmidt, Palash Sashittal, Ben Raphael
Department of Computer Science

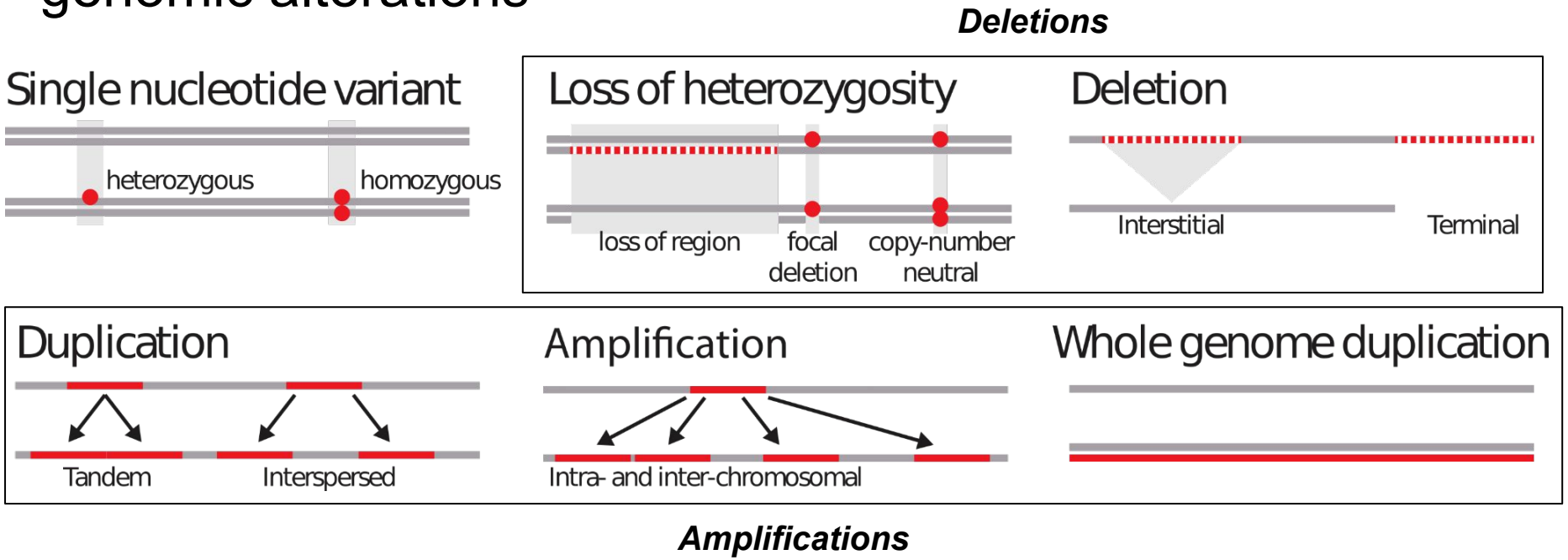
PRINCETON
UNIVERSITY

Cancer is characterized by both small and large scale genomic alterations



[1] Beerenwinkel, Niko, et al. "Cancer evolution: mathematical models and computational inference." *Systematic biology* 64.1 (2015): e1-e25.

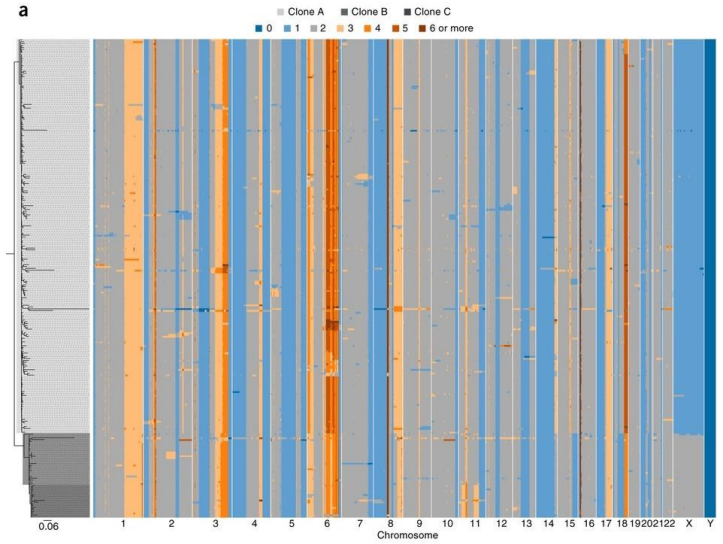
Cancer is characterized by both small and large scale genomic alterations



[1] Beerenwinkel, Niko, et al. "Cancer evolution: mathematical models and computational inference." *Systematic biology* 64.1 (2015): e1-e25.

Measuring copy number aberrations

Recent technological and computational improvements in DNA sequencing have led to **high resolution copy number profiles for thousands of cells.**

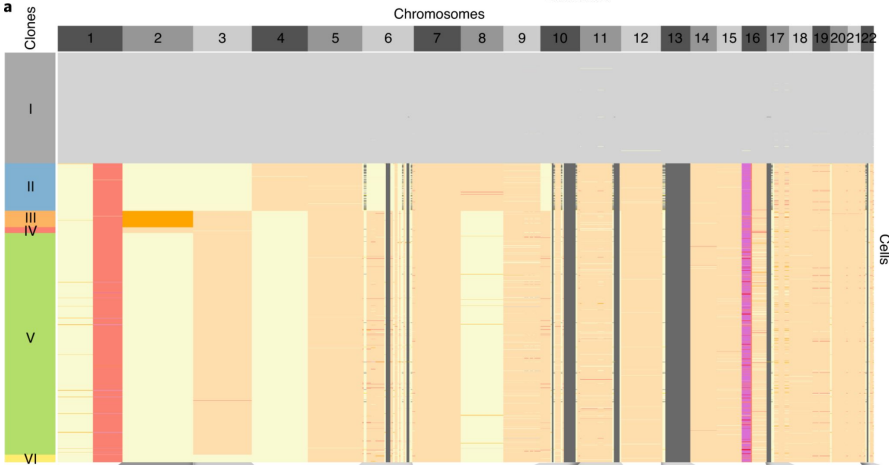


DNA sequencing technologies:

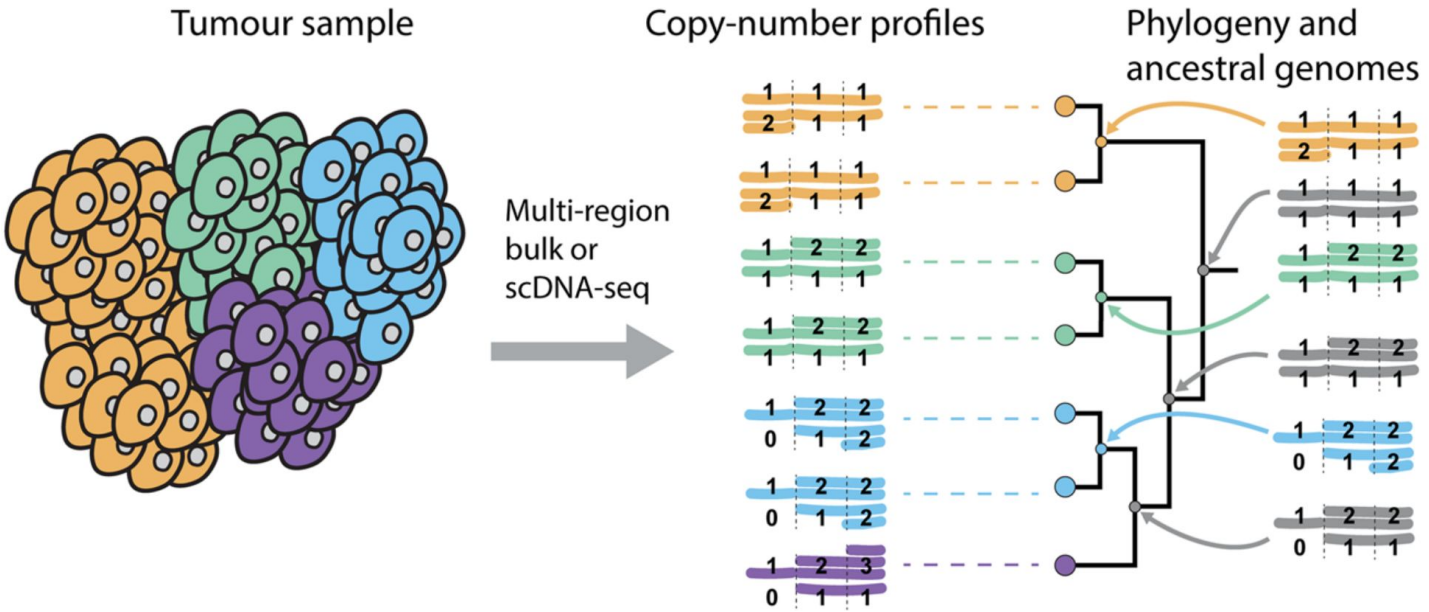
- *10X Genomics* CNV Kit (*Andor et al. 2020)*
- Direct library preparation *(Zahn et al. 2017)*
- DLP+ *(Laks et al. 2019)*
- ACT *(Minussi et al. 2021)*

Copy number inference methods:

- CHISEL *(Zaccaria et al. 2021)*
- Hatchet2 *(Myers et al. TBD)*
- Sccnv *(Dong et al. 2019)*
- Scope *(Wang et al. 2020)*
- SCNV *(Wang et al.* 2018)
- Starch (Elyanow et al. 2021)



The evolutionary history of copy number aberrations



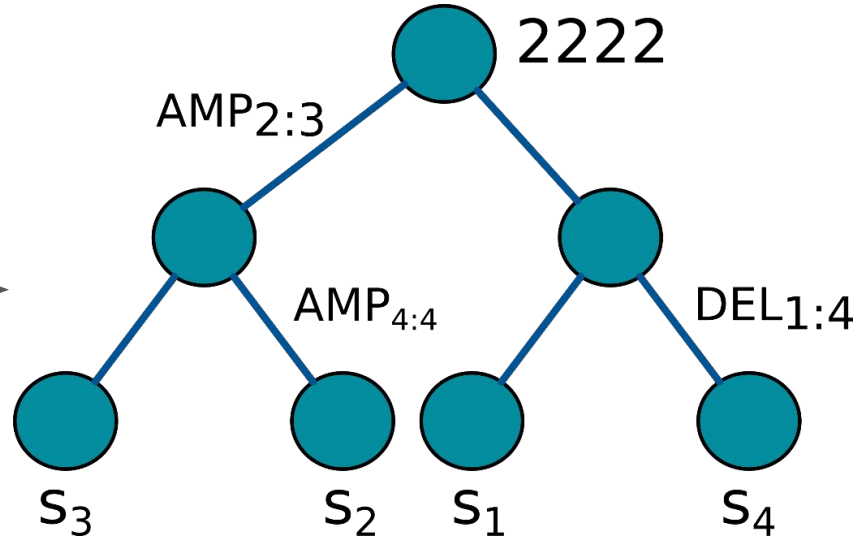
Constructing the evolutionary history of copy number aberrations *(copy number phylogenies)* from copy number profiles.

[2] Kaufmann, Tom L., et al. "MEDICC2: whole-genome doubling aware copy-number phylogenies for cancer evolution." *Genome biology* 23.1 (2022): 241.

Inferring copy number phylogenies

	C_1	C_2	C_3	C_4
S_1	2	2	2	2
S_2	2	3	3	2
S_3	2	3	3	3
S_4	1	1	1	1

Inference



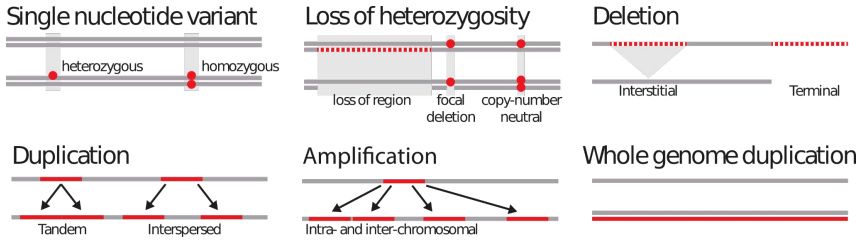
Input: Matrix of copy number profiles. Each copy number profile is a *non-negative integer vector*

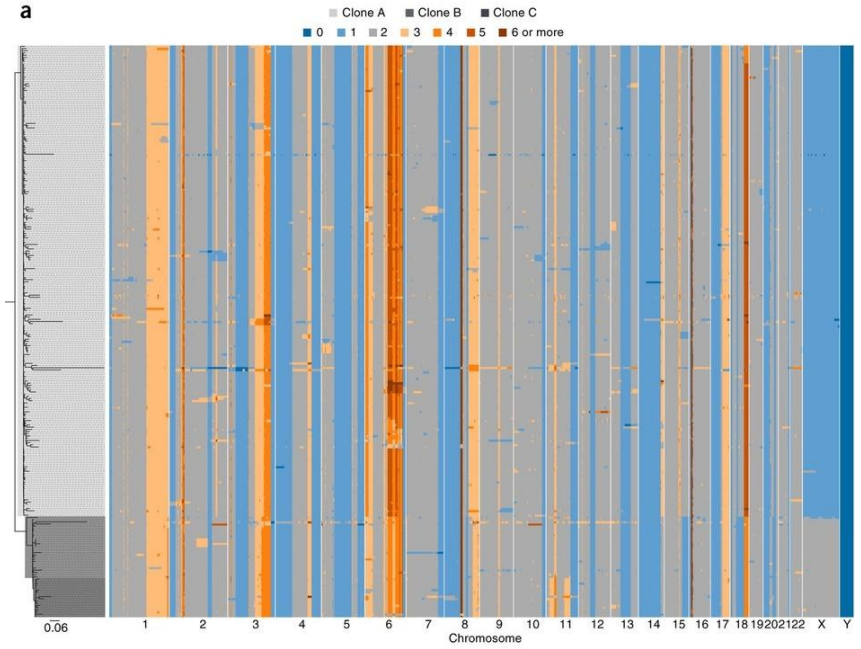
Output: Phylogenetic tree describing the evolutionary relationships between copy number profiles.

Challenges of inferring copy number phylogenies

Copy number aberrations are distinct from other modes of evolution:

- Phylogenetic characters are integers
- Copy number aberrations affect many loci simultaneously
- Large number of cells (100-1000s) and thousands of characters (> 4000)





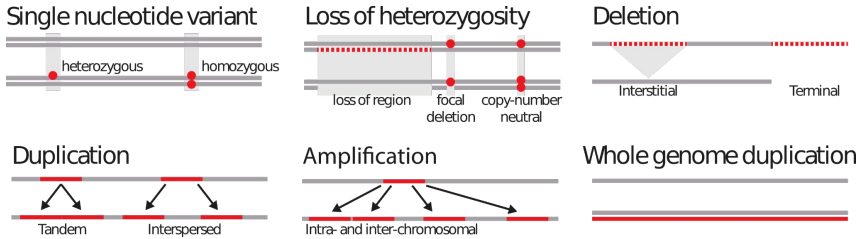
260 cells and 150Kb bins

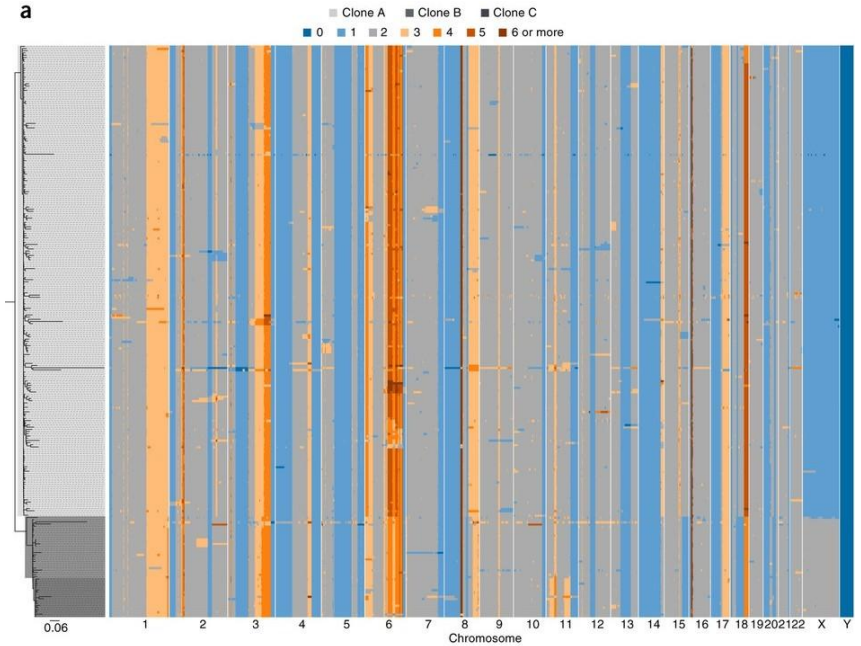
[3] Zahn, Hans, et al. "Scalable whole-genome single-cell library preparation without preamplification." Nature methods 14.2 (2017): 167-173.

Challenges of inferring copy number phylogenies

Copy number aberrations are distinct from other modes of evolution:

- Phylogenetic characters are integers
- Copy number aberrations affect many loci simultaneously
- Large number of cells (100-1000s) and thousands of characters (> 4000)



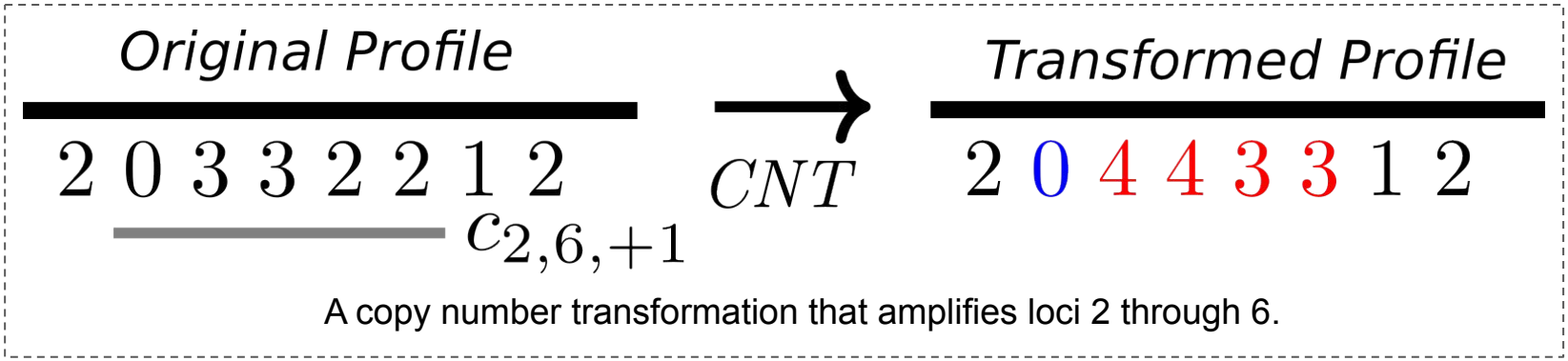


260 cells and 150Kb bins

[3] Zahn, Hans, et al. "Scalable whole-genome single-cell library preparation without preamplification." Nature methods 14.2 (2017): 167-173.

The copy number transformation model [9, 10]

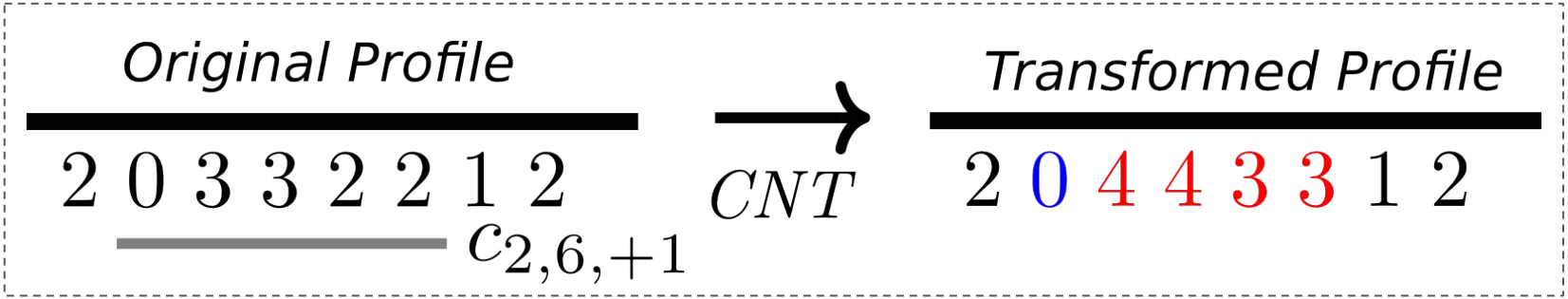
- Represents a copy number profile as a non-negative integer vector $p \in \mathbb{Z}_{+}^{n}$.
- Amplifications and deletions are represented as functions that increase or decrease the number of all loci in the target region:



A copy number transformation that amplifies loci 2 through 6.

[4] Schwarz, Roland F., et al. "Phylogenetic quantification of intra-tumour heterogeneity." *PLoS computational biology* 10.4 (2014): e1003535.
[5] El-Kebir, Mohammed, et al. "Complexity and algorithms for copy-number evolution problems." *Algorithms for Molecular Biology* 12 (2017): 1-11.

The copy number transformation model [9, 10]



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*

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

where $s \le t$ *and* $b \in \{+1, -1\}$.

Many state-of-the-art methods for phylogenetic inference are based on a single model of copy number evolution

In particular, the *copy number transformation (CNT)* [9, 10] model is a widely used model of copy number evolution and is the basis for many of the methods for phylogenetic inference of copy number phylogenies:

[4] Schwarz, Roland F., et al. "Phylogenetic quantification of intra-tumour heterogeneity." *PLoS computational biology* 10.4 (2014): e1003535.
[5] El-Kebir, Mohammed, et al. "Complexity and algorithms for copy-number evolution problems." *Algorithms for Molecular Biology* 12 (2017): 1-11.
[6] Zeira, Ron, and Benjamin J. Raphael. "Copy number evolution with weighted aberrations in cancer." *Bioinformatics* 36.Supplement_1 (2020): i344-i352.
[7] Shamir, Ron, Meirav Zehavi, and Ron Zeira. "A linear-time algorithm for the copy number transformation problem." *27th Annual Symposium on Combinatorial Pattern Matching (CPM 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.
[8] Kaufmann, Tom L., et al. "MEDICC2: whole-genome doubling aware copy-number phylogenies for cancer evolution." *Genome biology* 23.1 (2022): 241.

The copy number transformation model [9, 10]

We can use the copy number transformation model to compute *evolutionary distances* between copy number profiles

Definition 2. The *CNT distance between* two copy number profiles u and v is the **minimum** number of copy number events are needed to transform u to v. This distance is denoted d(u, v).

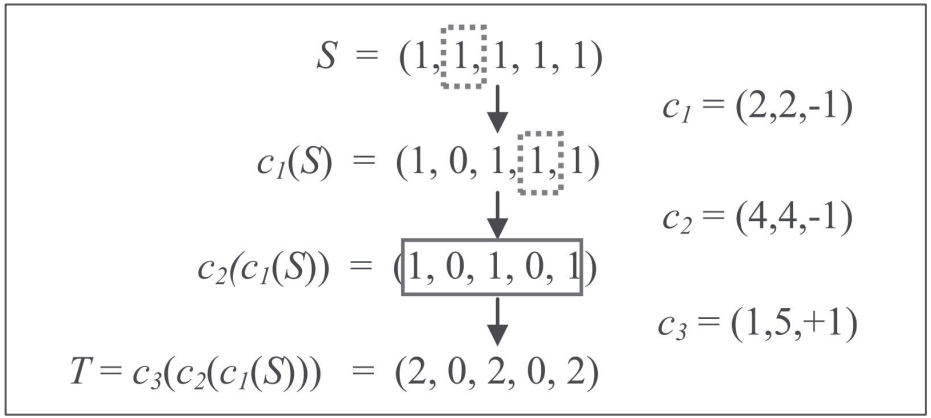
[6] Zeira, Ron, and Benjamin J. Raphael. "Copy number evolution with weighted aberrations in cancer." *Bioinformatics* 36.Supplement_1 (2020): i344-i352.
[7] Shamir, Ron, Meirav Zehavi, and Ron Zeira. "A linear-time algorithm for the copy number transformation problem." *27th Annual Symposium on Combinatorial Pattern Matching (CPM 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

The copy number transformation model [9, 10]

We can use the copy number transformation model to compute *evolutionary distances* between copy number profiles

Definition 2. The *CNT distance between* two copy number profiles u and v is the **minimum** number of copy number events are needed to transform u to v. This distance is denoted d(u, v).

[6] Zeira, Ron, and Benjamin J. Raphael. "Copy number evolution with weighted aberrations in cancer." *Bioinformatics* 36.Supplement_1 (2020): i344-i352.
[7] Shamir, Ron, Meirav Zehavi, and Ron Zeira. "A linear-time algorithm for the copy number transformation problem." *27th Annual Symposium on Combinatorial Pattern Matching (CPM 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.



At most 3 events to transform S into T (i.e. d(S, T) ≤ 3)

The copy number transformation model [9, 10]

We can use the copy number transformation model to compute *evolutionary distances* between copy number profiles

Definition 2. The *CNT distance between* two copy number profiles u and v is the **minimum** number of copy number events are needed to transform u to v. This distance is denoted d(u, v).

[6] Zeira, Ron, and Benjamin J. Raphael. "Copy number evolution with weighted aberrations in cancer." *Bioinformatics* 36.Supplement_1 (2020): i344-i352.
[7] Shamir, Ron, Meirav Zehavi, and Ron Zeira. "A linear-time algorithm for the copy number transformation problem." *27th Annual Symposium on Combinatorial Pattern Matching (CPM 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

Computable in linear time [7]

Definition 3. The *CNT median distance* between two copy number profiles u and v is the **minimum** of d(w, u) + d(w, v) over all copy number profiles w.

[9] Schwarz, Roland F., et al. "Phylogenetic quantification of intra-tumour heterogeneity." *PLoS computational biology* 10.4 (2014): e1003535.
[10] El-Kebir, Mohammed, et al. "Complexity and algorithms for copy-number evolution problems." *Algorithms for Molecular Biology* 12 (2017): 1-11.
[8] Kaufmann, Tom L., et al. "MEDICC2: whole-genome doubling aware copy-number phylogenies for cancer evolution." *Genome biology* 23.1 (2022): 241.

The copy number transformation model [9, 10]

We can use the copy number transformation model to compute *evolutionary distances* between copy number profiles

Definition 2. The *CNT distance between* two copy number profiles u and v is the **minimum** number of copy number events are needed to transform u to v. This distance is denoted d(u, v).

[6] Zeira, Ron, and Benjamin J. Raphael. "Copy number evolution with weighted aberrations in cancer." *Bioinformatics* 36.Supplement_1 (2020): i344-i352.
[7] Shamir, Ron, Meirav Zehavi, and Ron Zeira. "A linear-time algorithm for the copy number transformation problem." *27th Annual Symposium on Combinatorial Pattern Matching (CPM 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

Computable in linear time [7]

Definition 3. The *CNT median distance* between two copy number profiles u and v is the **minimum** of d(w, u) + d(w, v) over all copy number profiles w.

[9] Schwarz, Roland F., et al. "Phylogenetic quantification of intra-tumour heterogeneity." *PLoS computational biology* 10.4 (2014): e1003535.
[10] El-Kebir, Mohammed, et al. "Complexity and algorithms for copy-number evolution problems." *Algorithms for Molecular Biology* 12 (2017): 1-11.
[8] Kaufmann, Tom L., et al. "MEDICC2: whole-genome doubling aware copy-number phylogenies for cancer evolution." *Genome biology* 23.1 (2022): 241.

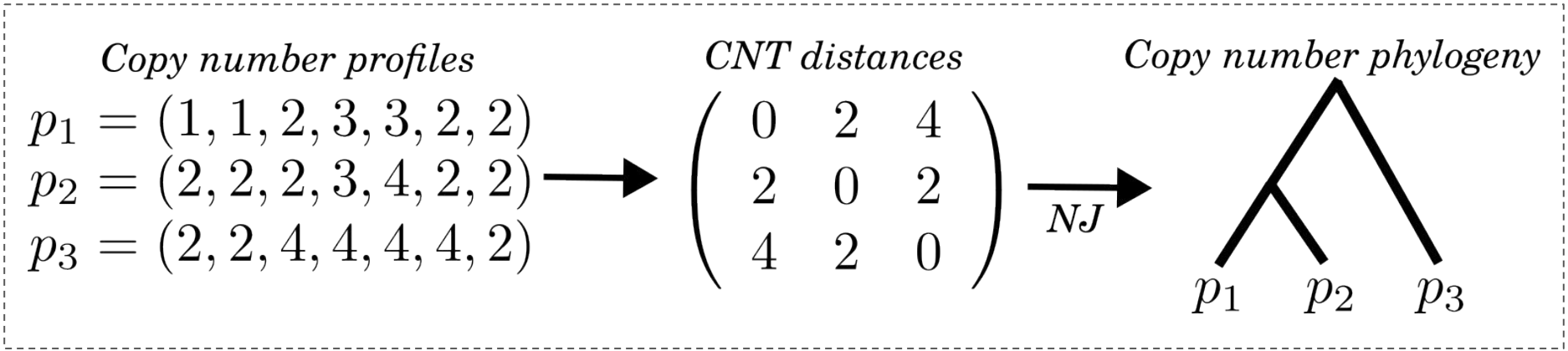
Computable in pseudo-polynomial time [10]

The copy number transformation model [9, 10]

Unfortunately, computing evolutionary distances is *all* we know how to do

Phylogenetic inference using the copy number transformation model

Inference of copy number phylogenies using CNT always proceeds in three steps*:



* [10] (El-Kebir et al. 2017) does solve the CNT large parsimony exactly, but it uses an ILP and only scales to ~20 cells.

Question: Can we employ phylogenetic techniques *beyond distance based methods* using the copy number transformation model?

Question: Can we employ phylogenetic techniques *beyond distance based methods* using the copy number transformation model?

Goal: We will start out simply, by trying to employ the *method of maximum parsimony* for the copy number transformation model.

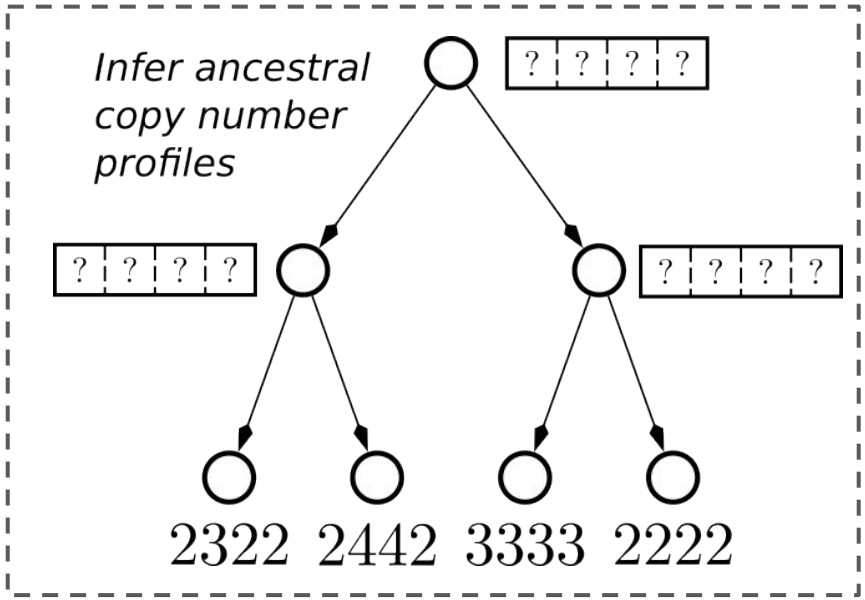
Question: Can we employ phylogenetic techniques *beyond distance based methods* using the copy number transformation model?

Goal: We will start out simply, by trying to employ the *method of maximum parsimony* for the copy number transformation model.

- Even simpler: given candidate copy number phylogenies, which best fits the data?

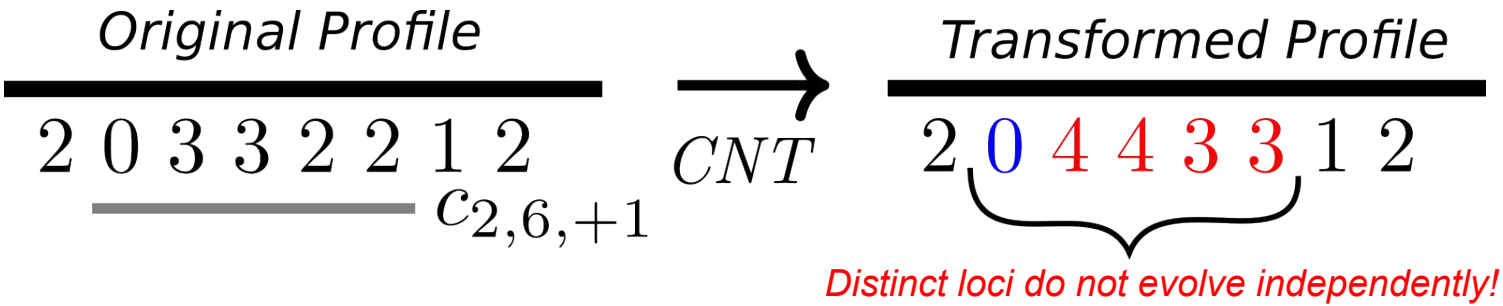
CNT small parsimony

Goal: Given a leaf labeled tree, infer **the ancestral states** on the tree that minimizes the total number of CNT events required to explain the tree:



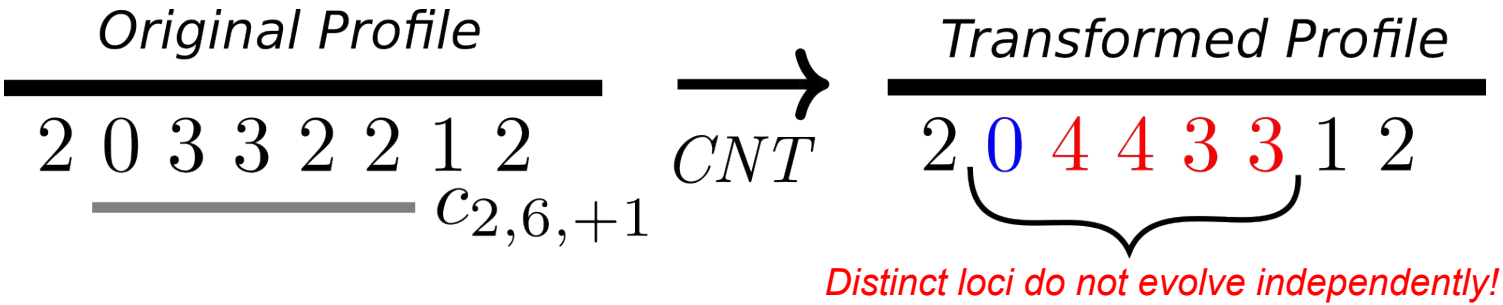
CNT small parsimony

But unlike for standard phylogenetic models, the CNT small parsimony problem is really challenging to solve*:



CNT small parsimony

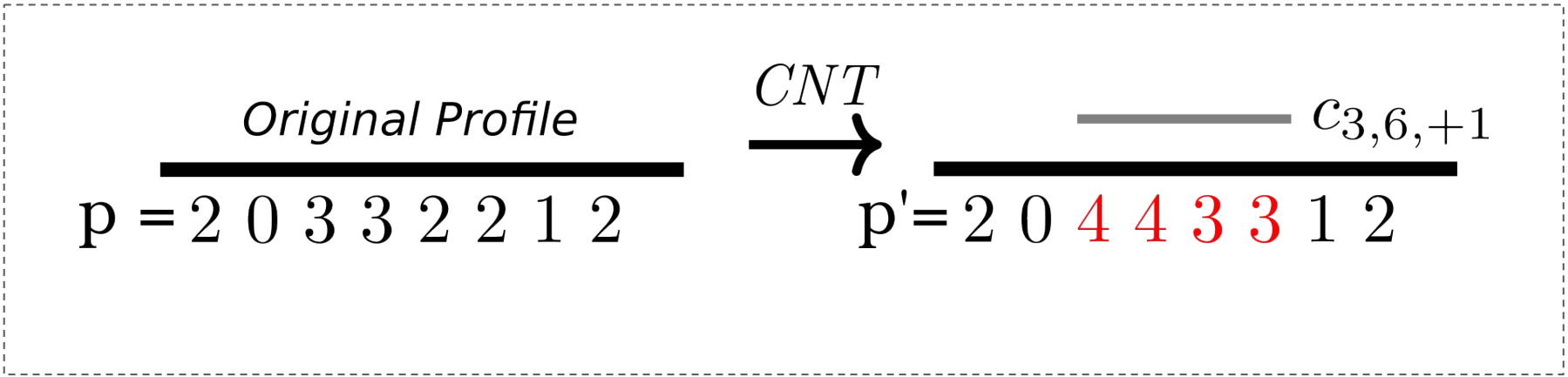
But unlike for standard phylogenetic models, the CNT small parsimony problem is really challenging to solve*:



Applications of bottom-up dynamic programming approaches to the small parsimony problem, such as Sankoff's algorithm, will not result in polynomial time algorithms!

** In fact, I will purchase dinner for anyone who can solve the CNT small parsimony problem in polynomial time, since I do not believe this to be possible.*

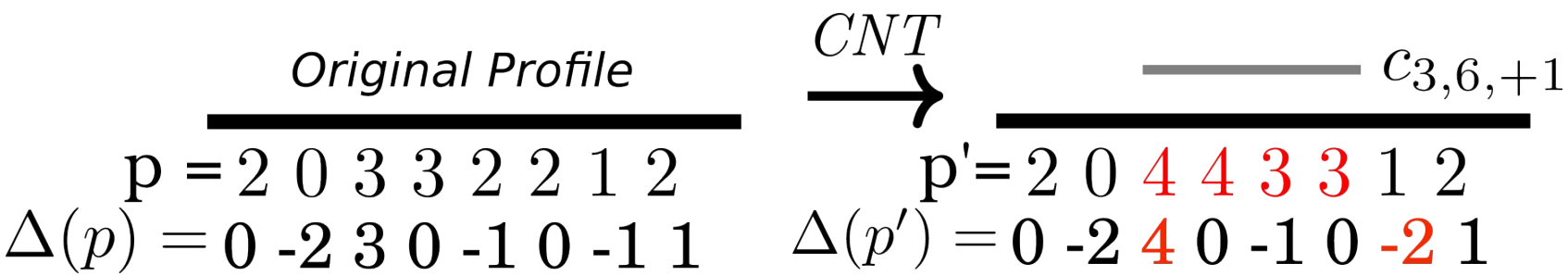
A curious idea: the “derivative” of a copy number event



That is, while a copy number event affects the copy number of the entire region {3, 4, 5, 6}, it only affects two of the *differences* in copy number: $p_3 - p_2$ and $p_7 - p_6$.

A curious idea: the “derivative” of a copy number event

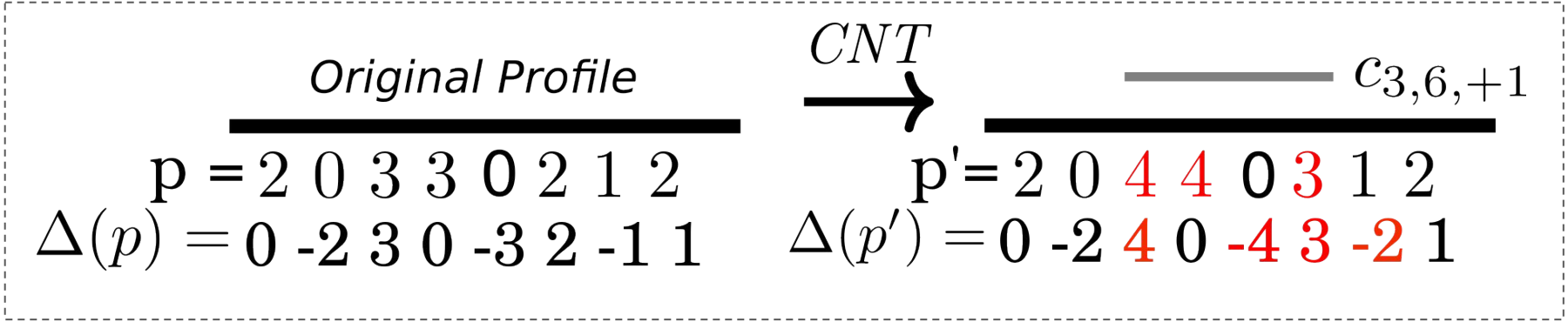
A copy number event only affects the endpoints of the derivative!



That is, while a copy number event affects the copy number of the entire region {3, 4, 5, 6}, it only affects two of the *differences* in copy number: $p_3 - p_2$ and $p_7 - p_6$.

A curious idea: the “derivative” of a copy number event

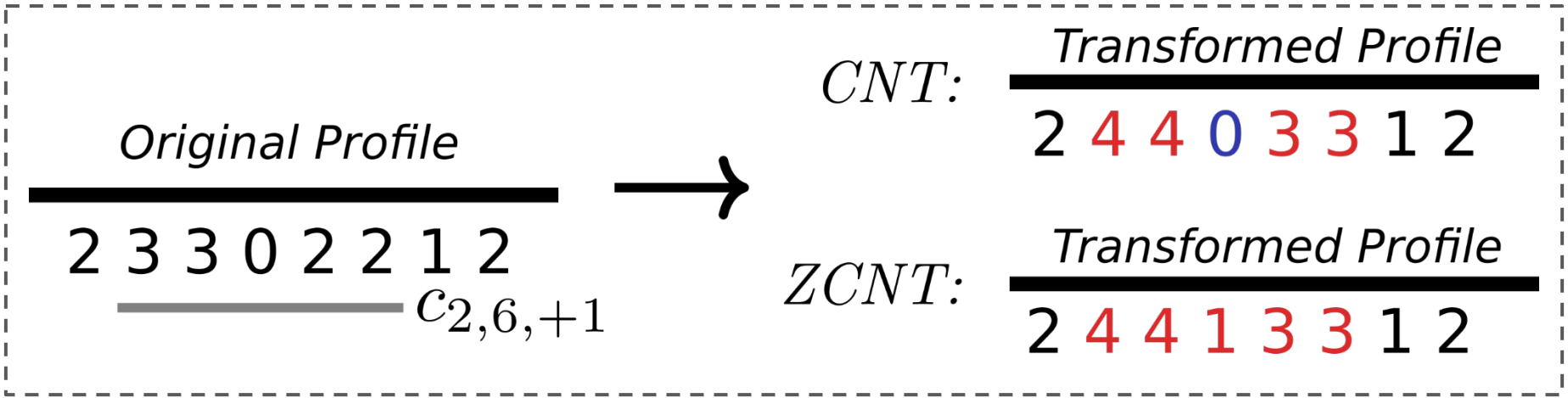
But I lied, this won’t work when the copy number event spans loci with zero copy number!



That is, the copy number event will also affect the differences on both sides of any zero it spans!

Zero-agnostic copy number transformation (ZCNT) model

To make this idea work, we need to slightly tweak the CNT model to ensure that it suffices to analyze the "*derivative*" of copy number events*:



* This *"derivative"* is called the *delta map* Δ in our paper.

Zero-agnostic copy number transformation (ZCNT) model

To make this idea work, we need to slightly tweak the CNT model to ensure that it suffices to analyze the "*derivative*" of copy number events*:

Zero agnostic copy number event

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*

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

where $s \le t$ *and* $b \in \{+1, -1\}$.

* This *"derivative"* is called the *delta map* Δ in our paper.

ZCNT small parsimony

In fact, this basic tweak to the model and our idea of the *derivative* of a copy number event allows us to solve two natural relaxations of the small parsimony problem in polynomial time!

> **Theorem 3.** *If the balancing condition 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.*

To our knowledge, this makes our work the first attempt at solving the small parsimony problem for a segment-based (i.e. non-independent) model of evolution.

ZCNT small parsimony conjecture

Further, we conjecture that the ZCNT small parsimony problem is exactly solvable in polynomial time:

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

ZCNT small parsimony conjecture

Further, we conjecture that the ZCNT small parsimony problem is exactly solvable in polynomial time:

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

- Confirmed by a simulation study where copy number profiles were drawn from real, cancer data

ZCNT small parsimony conjecture

Further, we conjecture that the ZCNT small parsimony problem is exactly solvable in polynomial time:

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

- Confirmed by a simulation study where copy number profiles were drawn from real, cancer data
- We have drafted a proof of this conjecture, which is not included in the current manuscript

Lazac (Large analysis of zero agnostic copy number): An algorithm for ZCNT large parsimony

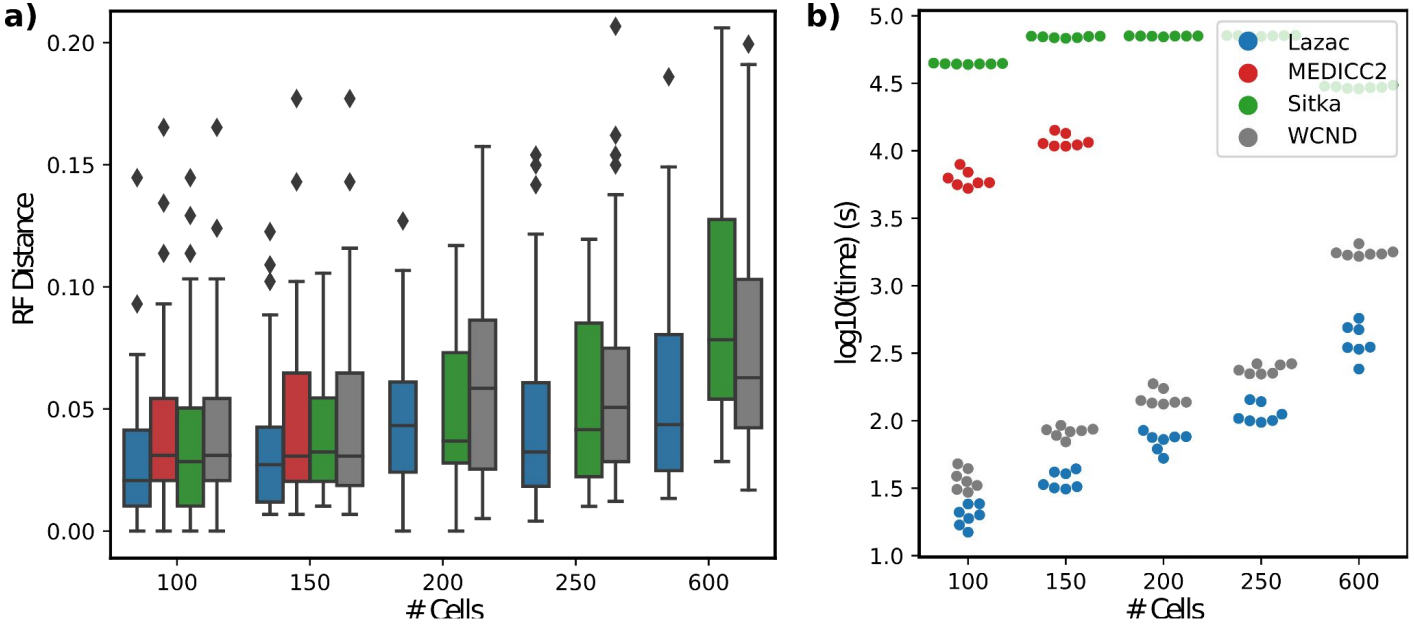
Our efficient solution to the ZCNT small parsimony problem enables us to derive a *stochastic algorithm* based on NNIs to infer copy number phylogenies [14]:

\+ ZCNT =

Lazac

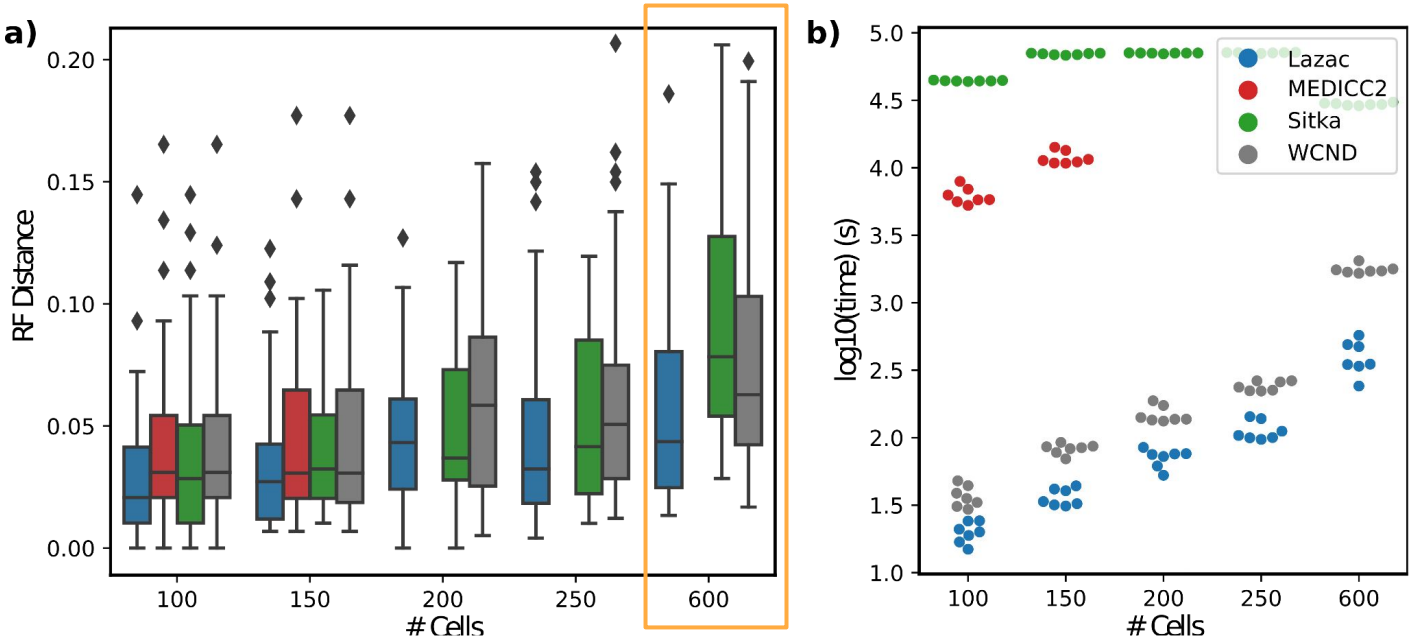
[14] Nguyen, Lam-Tung, et al. "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies." *Molecular biology and evolution* 32.1 (2015): 268-274.

Lazac infers more accurate phylogenies than other methods, much more quickly



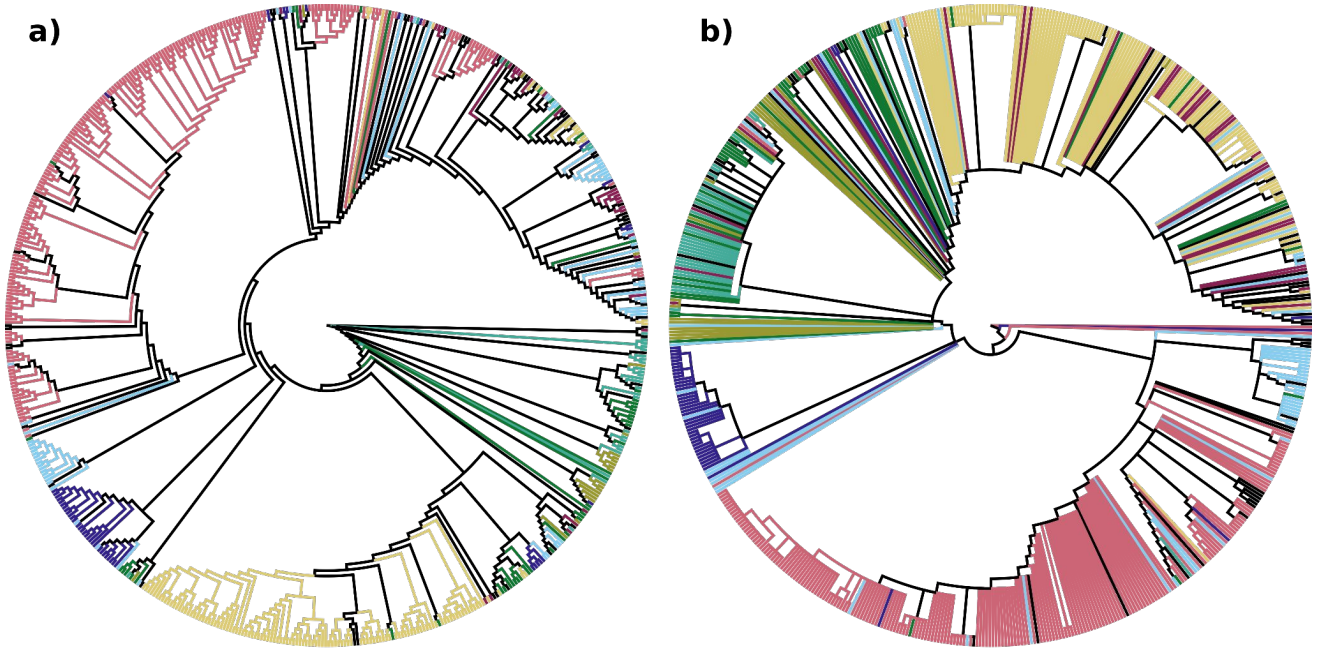
Comparison of reconstruction accuracy (RF distance) on CONET simulated data for several state-of-the art methods for copy number tree reconstruction with varying number of cells.

Lazac infers more accurate phylogenies than other methods, much more quickly



Comparison of reconstruction accuracy (RF distance) on CONET simulated data for several state-of-the art methods for copy number tree reconstruction with varying number of cells.

Single-cell DNA sequencing data from human ovarian and breast tumor samples



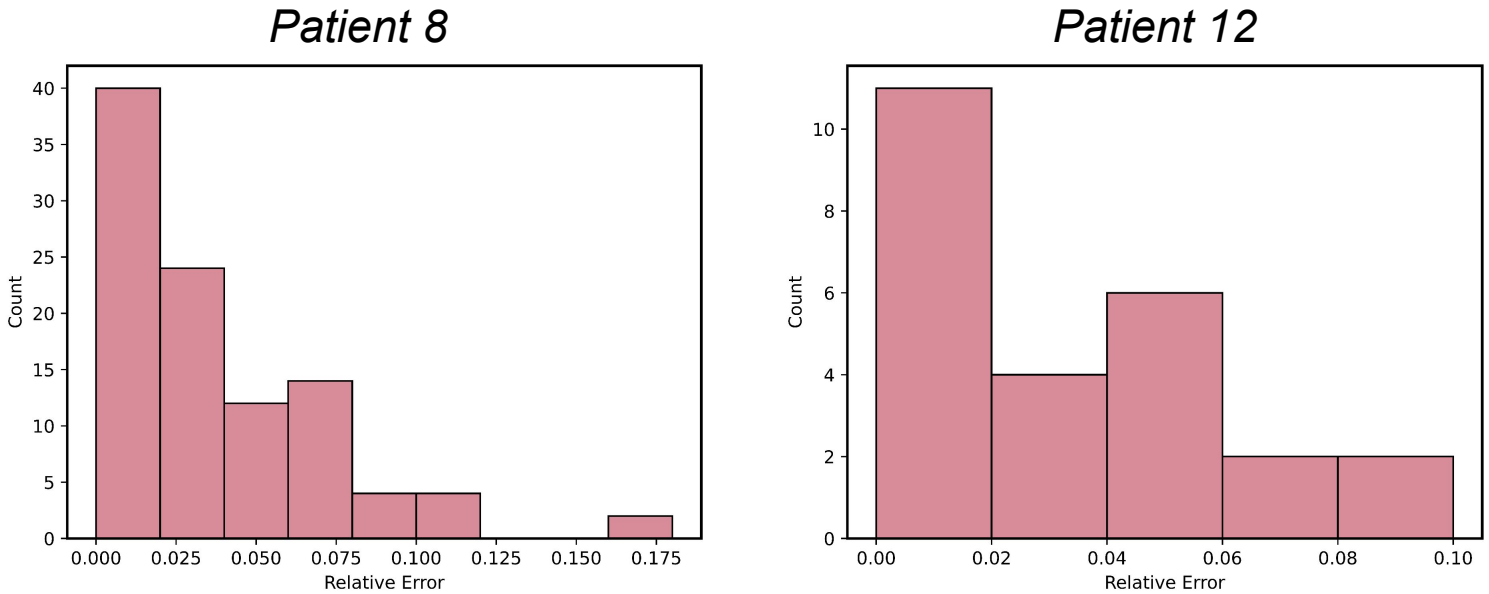
Lazac Copy Number Phylogeny
Clonal Discordance: 95

Sitka Copy Number Phylogeny
Clonal Discordance: 117

The copy number phylogeny inferred by *Lazac* **(Left)** and Sitka **(Right)** on sample SA1184 [12] with the leaves colored by the corresponding clone labels. The normalized RF distance between the two trees is 0.9869.

[15] Funnell, Tyler, et al. "Single-cell genomic variation induced by mutational processes in cancer." Nature (2022): 1-10.

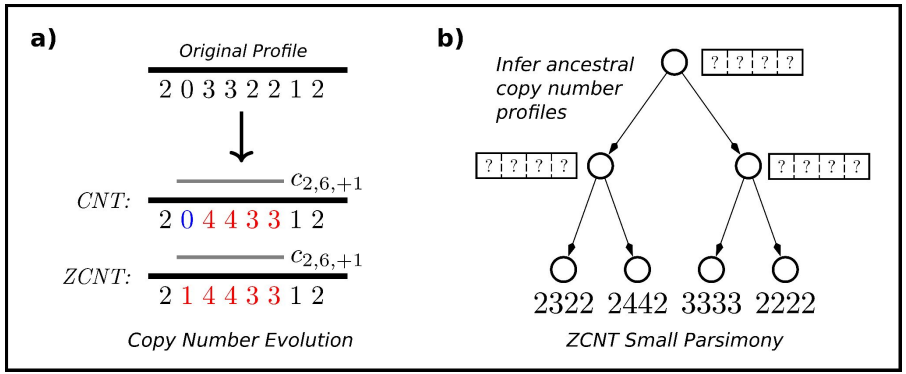
ZCNT distances approximate CNT distances



The relative error between d_{CNT} and d_{ZCNT} for all pairs of distances from bulk whole-genome sequencing data from two metastatic prostate cancer patients [13].

[16] Gundem, Gunes, et al. "The evolutionary history of lethal metastatic prostate cancer." Nature 520.7547 (2015): 353-357.

ZCNT Summary



- We introduce a new model of copy number evolution, the *zero-agnostic copy number transformation model*, which enables us to analyze the "*derivative"* of copy number events
- With our new model and analysis technique, we derive *polynomial time* algorithms for two relaxations of the small parsimony problem for copy number transformations
- We derive an efficient method, *Lazac*, for solving the large parsimony problem that scales to *thousands of single cells* and *recovers more accurate phylogenies* than existing methods

Thank You

Group Members

Ben Raphael
Metin Balaban
Cong Ma
Uyen Mai
Palash Sashittal
Uthsav Chitra
Sereno Lopez-Darwin
Hirak Sarkar
Brian Arnold
Peter Halmos

Gillian Chu
Maya Gupta
Xinhao Liu
Henri Schmidt
Ahmed Shuaibi
Alexander Strzalkowski
Akhil Jakatdar
Gary Hu
Clover Zheng

The Raphael Lab

Lazac is implemented in C++17 and available on GitHub

bioRχiv
THE PREPRINT SERVER FOR BIOLOGY