Sc-TUSV-ext: Single-cell clonal lineage inference from single nucleotide variants (SNV), copy number alterations (CNA) and structural variants (SV)

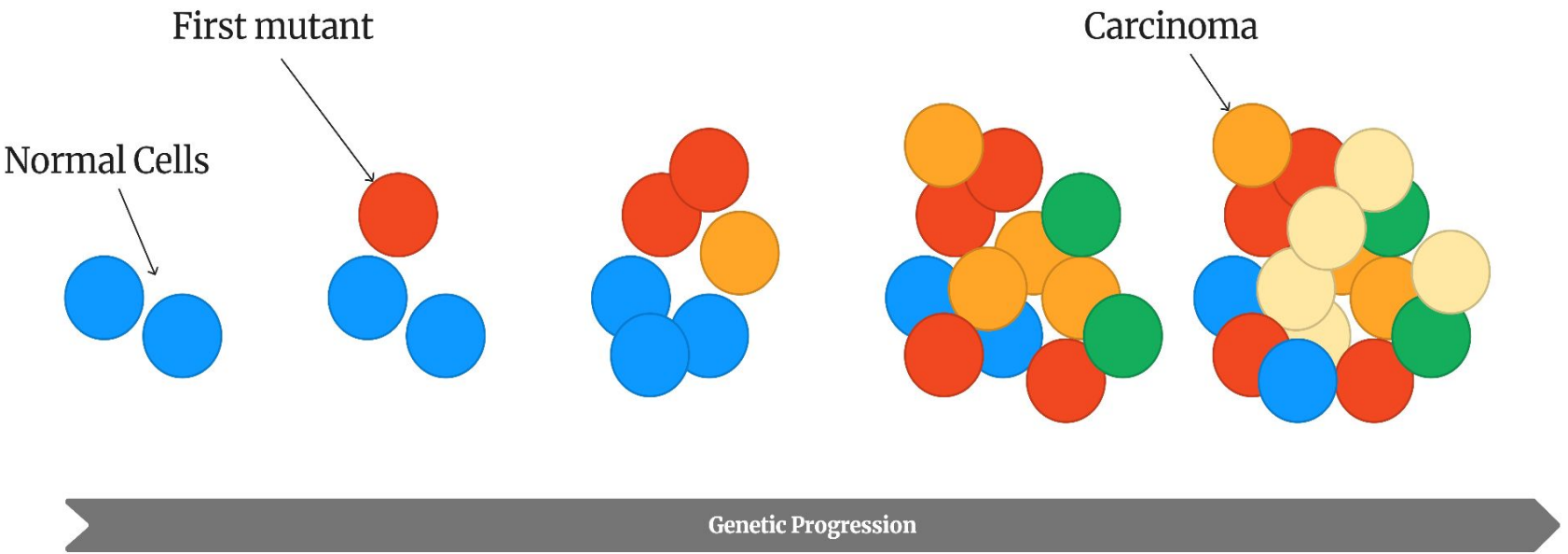
Nishat Anjum Bristy, Xuecong Fu, Russell Schwartz

Computational Biology Department

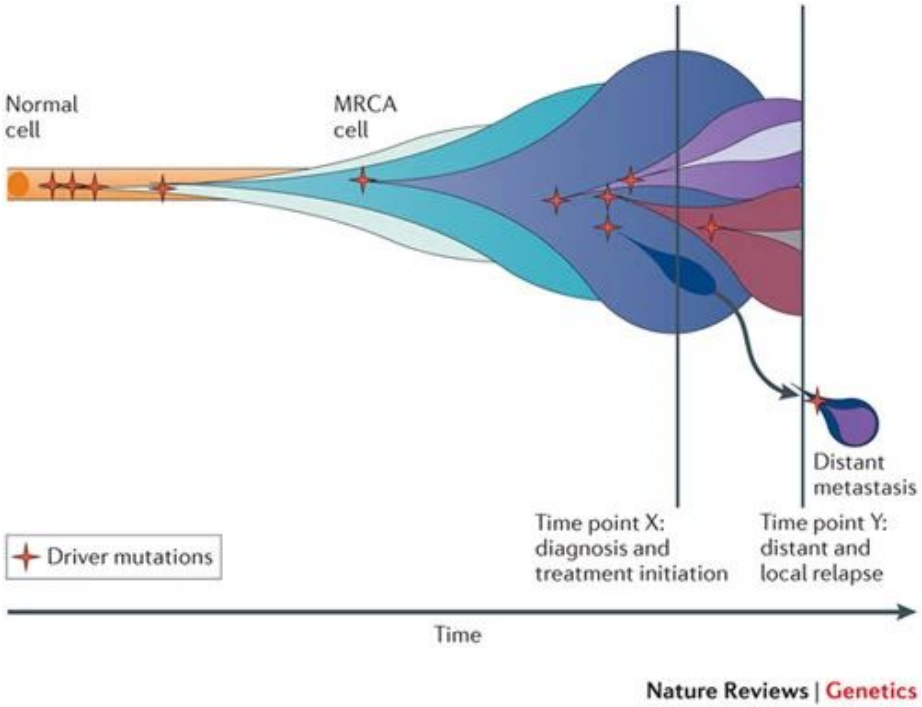
Carnegie Mellon University

Cancer is a Disease of Evolution

(Nowell. Science, 194:23-28, 1976)

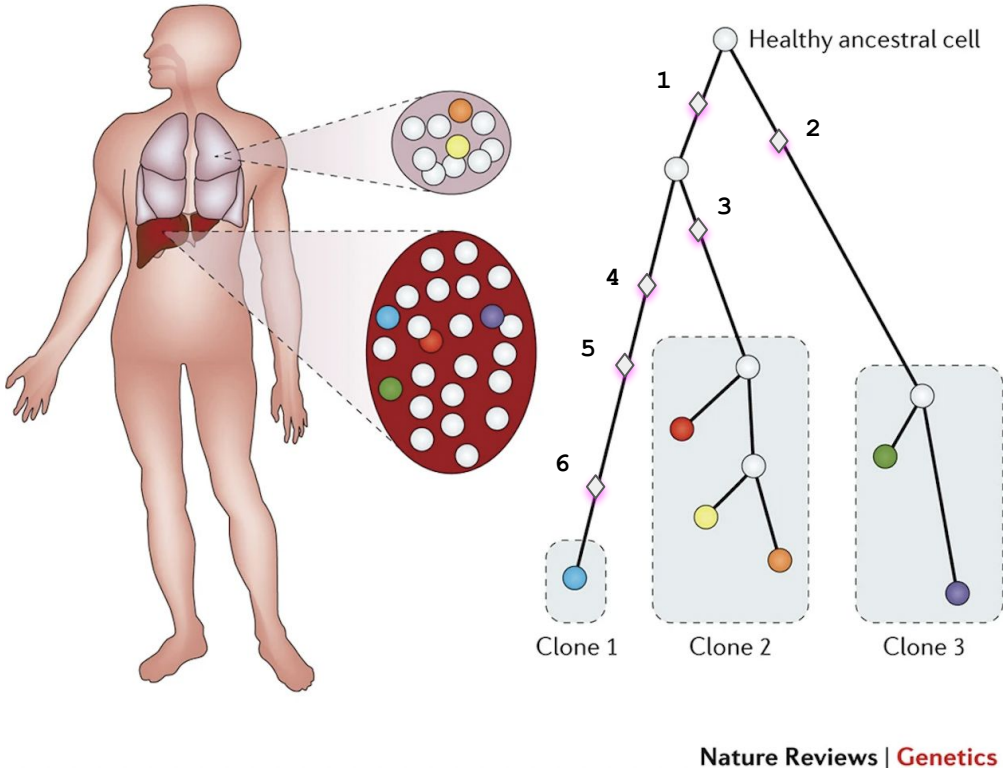


Natural Selection Creates Distinct Cell Subpopulations (Clones) within the Tumor



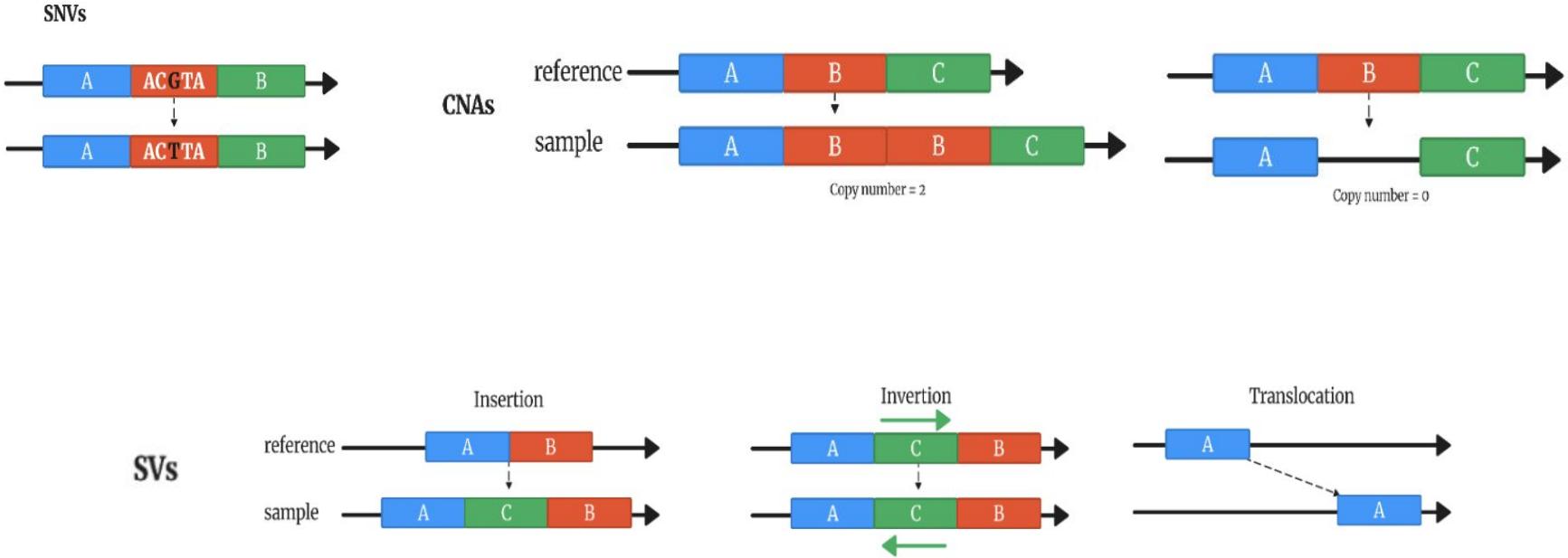
Yates and Campbell, Nature Reviews Genetics, 2012

Mutation Order Matters and Evolutionary Trees Help Quantify Them



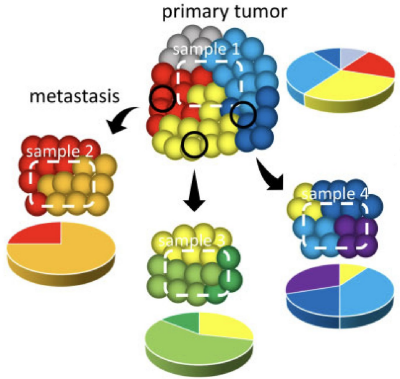
Schwartz and Schäffer, Nature Reviews Genetics, 2017

Different Types of Mutations Can Drive Different Tumor Clones

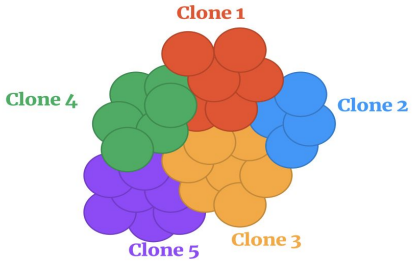


Different types data and methods are used in the research area

Bulk sequencing



Single-cell sequencing



From these datasets, clonal/cell trees are built with numerous approaches -

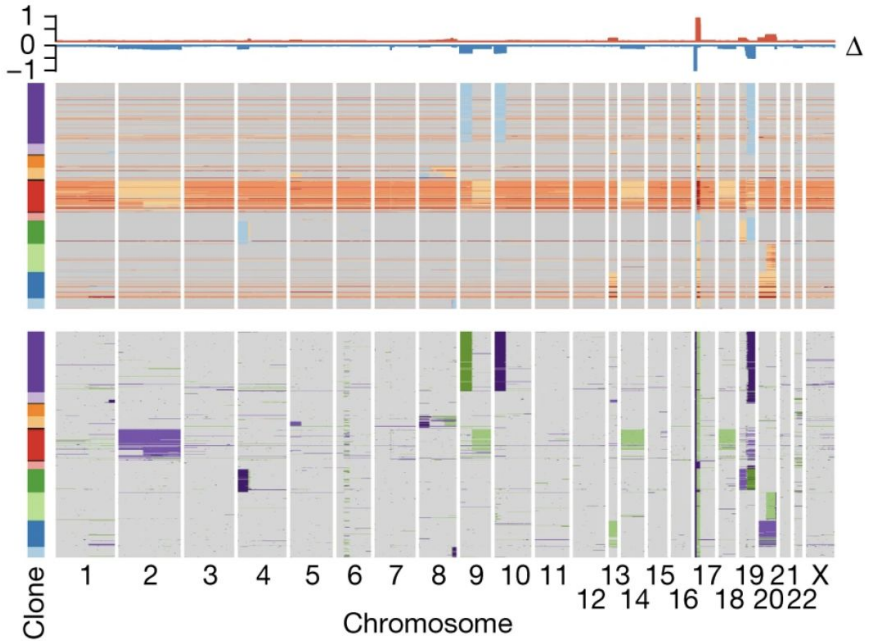
- Probabilistic approaches
- Combinatorial approaches
- Integer linear programming or constraint satisfaction problems
- Distance based approaches: finite state transducers.

Single-Cell DNA sequencing datasets

- DLP (Zahn et al., 2017)
- DLP+ (Laks et al. 2019)
- 10X Genomics

Variant calling technologies

- SNV: MutationSeq, Strelka
- CNA: HMMCopy, SIGNALS
- SV: LUMPY, deStruct



Funnell et al., Nature, 2022

Problem

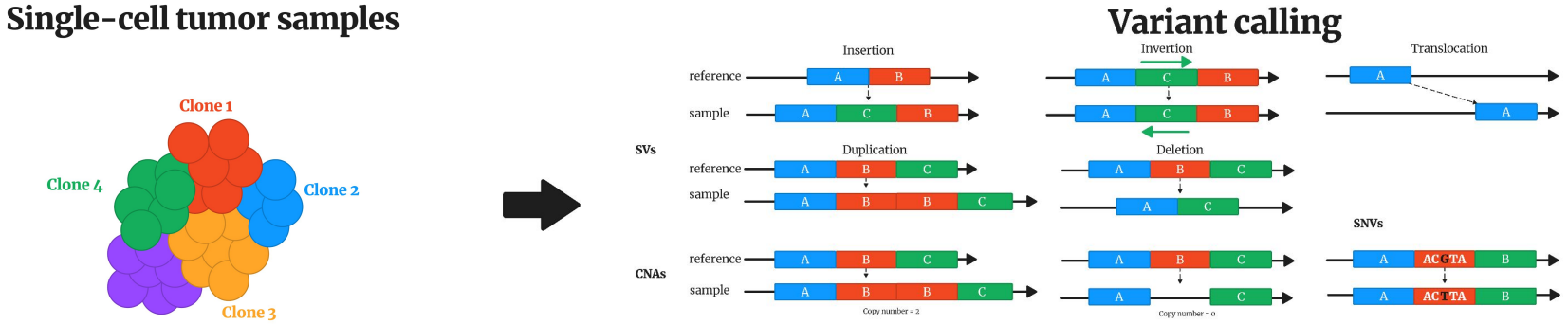
Tumor Phylogeny Reconstruction from erroneous single-cell data

We divide this problem into two sub-problems.

Sub-Problem 1

Given single-cell whole genome sequences, identify tumor clones.

Problem Statement: Identifying tumor clones from single-cell WGS.

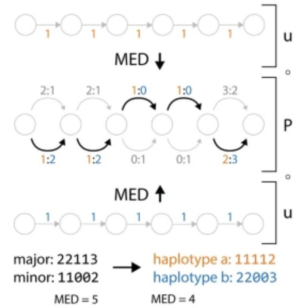


Input: Single-cell variant calls

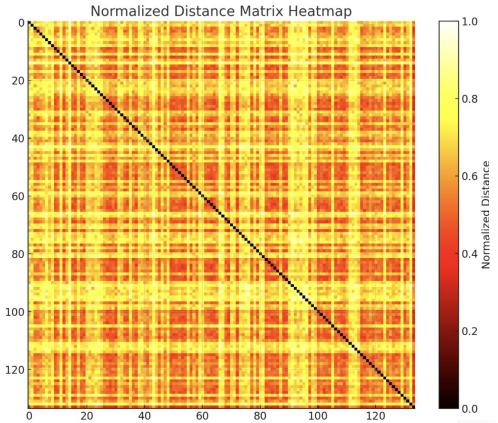
Output: Tumor clones and clonal mutation matrix.

MEDICC2 Distances are Used for Copy Number Clustering

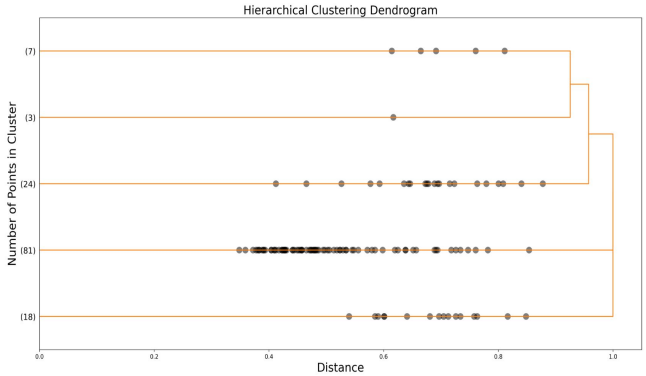
MEDICC2 algorithm



MEDICC2 distances



Hierarchical clustering



Kaufmann et al., Genome Biology, 2022

Sub-Problem 2

Given the single-cell clonal mutation matrix with SNVs, CNAs and SVs, reconstruct the tumor phylogeny as well as correct single-cell sequencing errors.

Problem Statement: Finding the Optimal Evolutionary Tree

Mutations

Clones

C_{obs}	sv 1	...	sv l	snv 1	...	snv g	cna 1p	...	cna rp	cna 1m	...	cna rm
Clone 1	0	..	1	0	..	0	1	..	2	1	..	1
Clone 2	1	..	1	1	..	0	1	..	1	1	..	2
Clone 3	..	..	..	..	..	..	..	..	..	..	..	..
Clone 4	..	..	..	..	..	..	..	..	..	..	..	..
clone 5	0	..	0	0	..	1	1	..	1	1	..	1

l : Number of SV breakpoints
r: Number of allele specific segments
g: Number of SNV positions
n: Number of clones

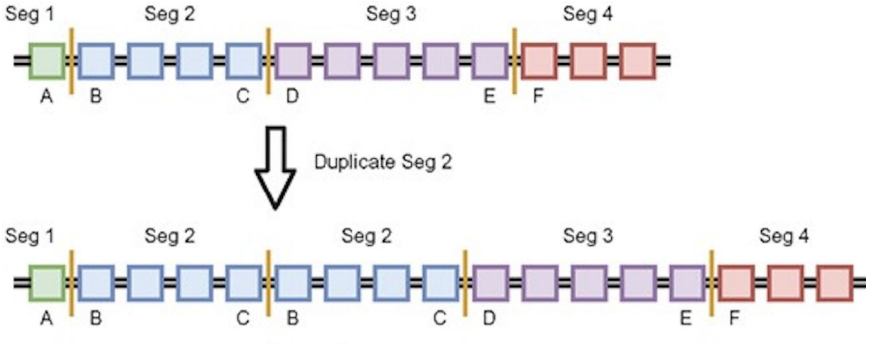
Input: Mutation matrix of size (# Clones $\times$ # Mutations)

Output: Joint inference of the **clonal lineage tree** as well as an **estimated mutation matrix** that minimizes the evolutionary distance as well as the difference between true and estimated mutations.

Question: How do we incorporate SVs and SNVs with CNAs in an ILP framework?

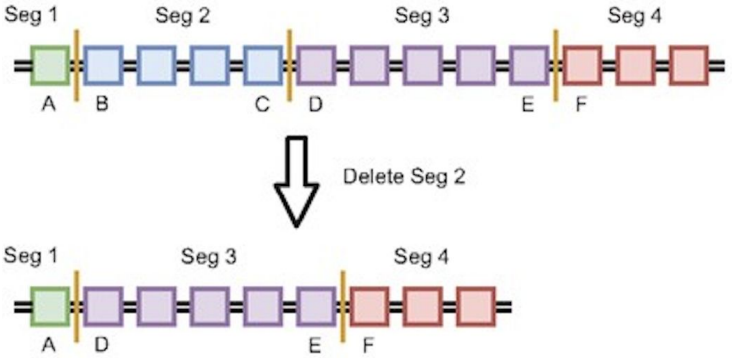
SVs are represented as a pair of breakpoints

Duplication

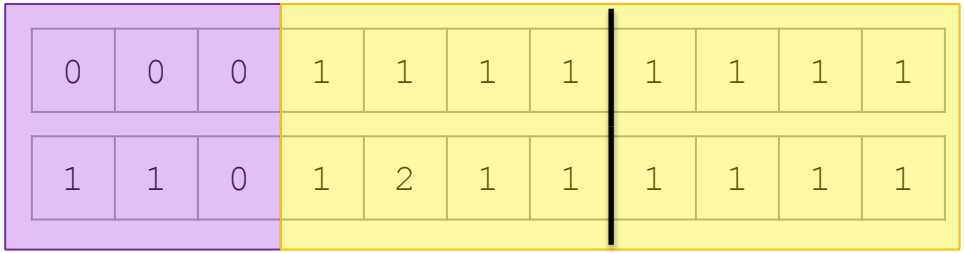


]C][B[

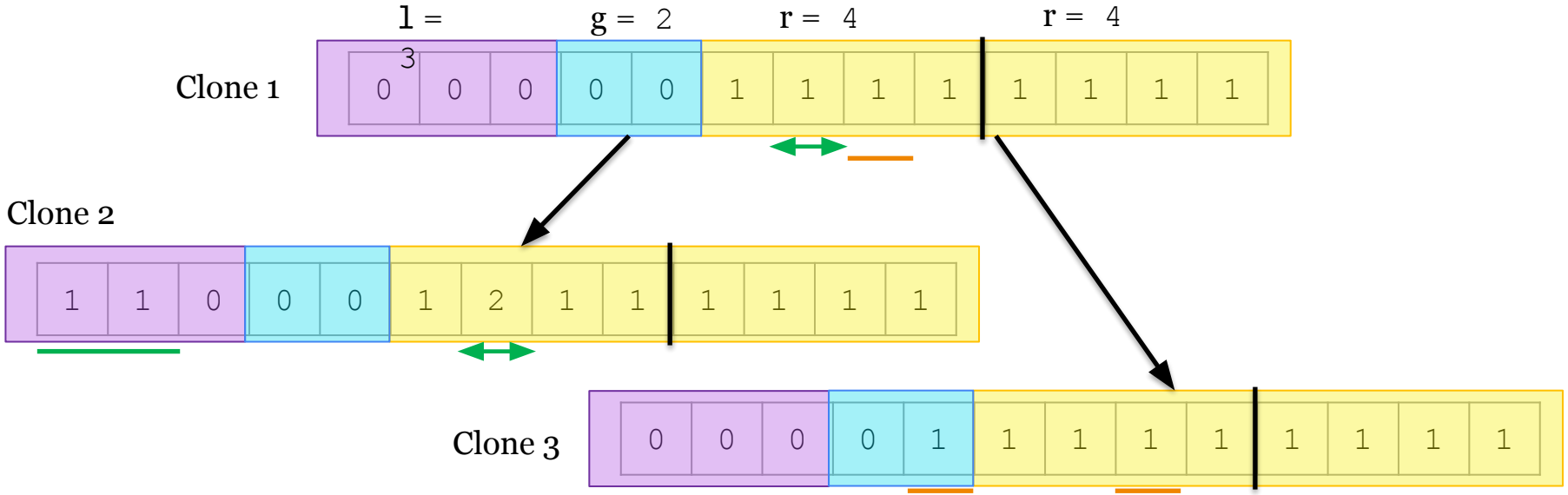
Deletion



]A][D[

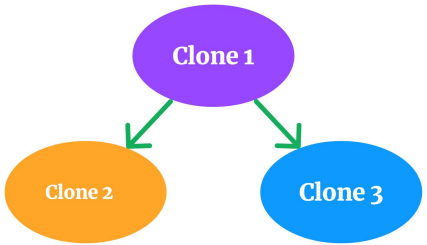


SNVs are Represented in Binary and the Input Matrix is a Combination of these SVs, SNVs and CNAs



$$C_{obs} \rightarrow C_{est}$$

The Phylogeny is Represented by an Edge Matrix



$$E = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}_{n \times n}$$ ← Root

$$Z = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}_{n \times n}$$ ← Root

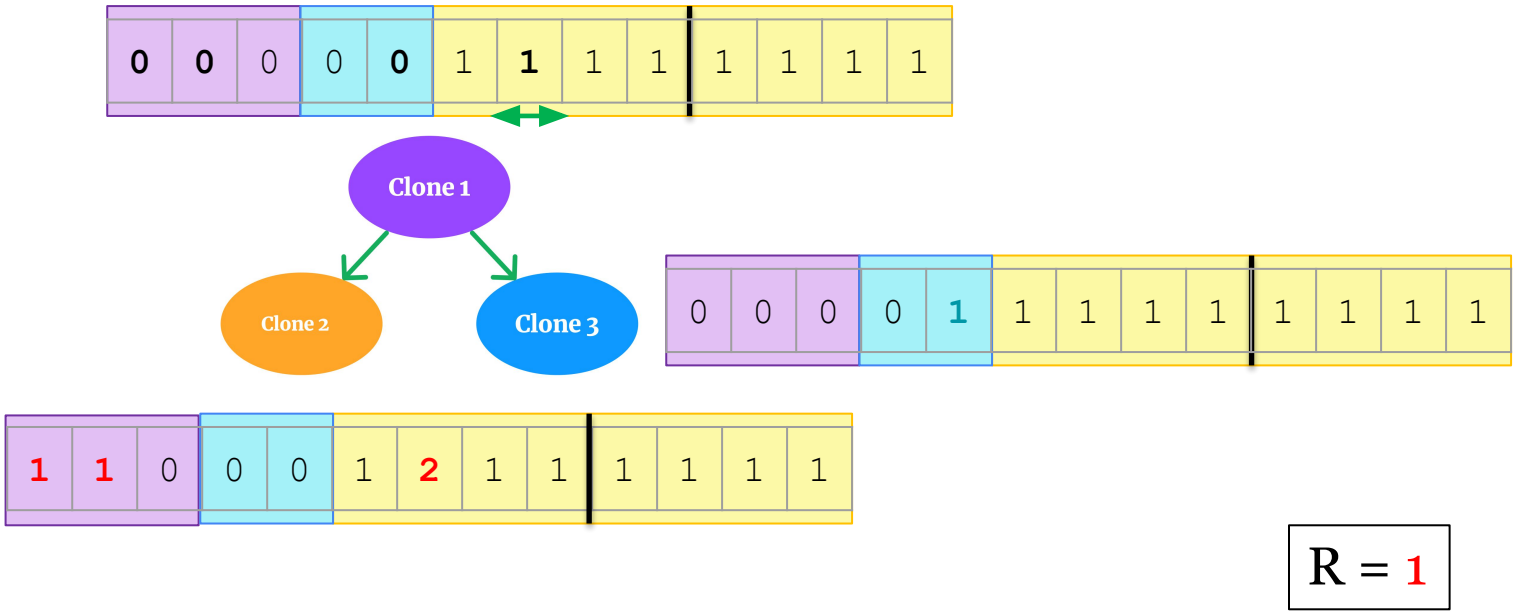
Estimating Z and C_{obs} to Minimize the Cost of Objective

Objective:

$$\min_{Z, C_{est}} |Z.C_{est} - C_{obs}| + \underbrace{\lambda R}$$

Regularization term
Phylogenetic cost

Phylogenetic Cost Depends on the Difference of Copy Number Changes Along the Edges

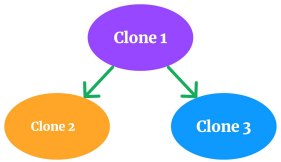


R = 1

Dollo Phylogeny on Breakpoints and SNVs Ensure a Less Restrictive Model

A mutation can be gained only once but lost multiple times in the evolutionary history

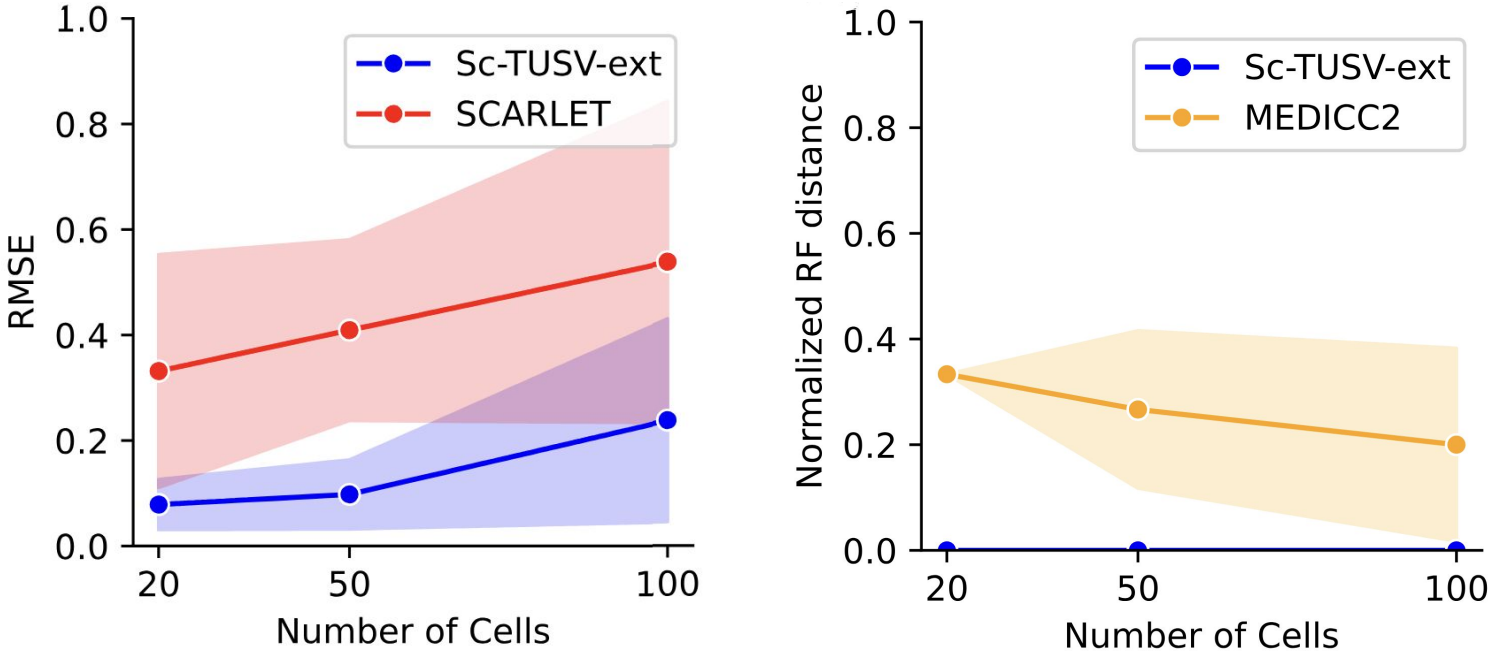
$$w_{i,j,b} = \begin{cases} 1 & \text{if mutation } b \text{ occurs along the edge from node } i \text{ to node } j \\ 0 & \text{otherwise} \end{cases}$$



Losses are allowed due to concurrent segmental copy number losses.

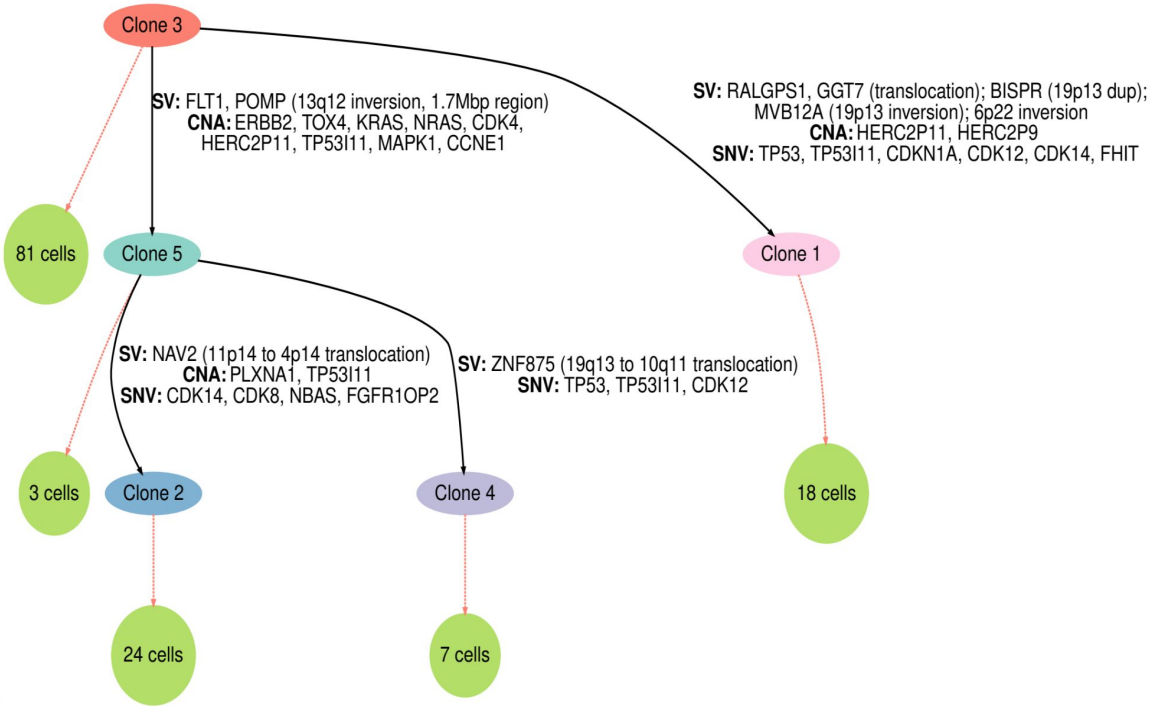
Rogozin et al., Parsimony, phylogeny, and genomics 190, 2005

Sc-TUSV-ext Estimates Single-Cell Copy Numbers and Phylogeny More Accurately Than Other Methods



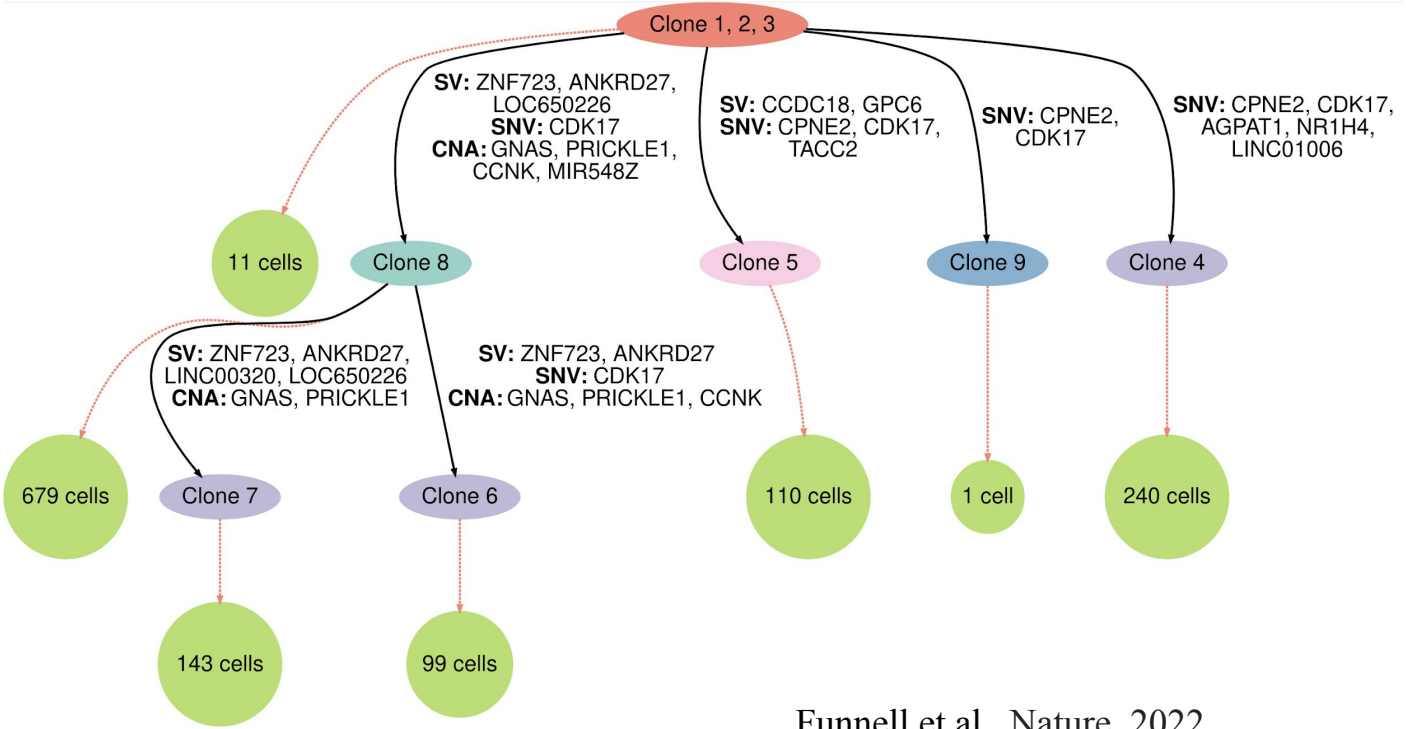
X-axis shows different number of single-cells. Y-axis shows variant reconstruction accuracy (left) and tree reconstruction accuracy (right).

Evol. History of a High Grade Serous Ovarian Cancer Patient's Dataset from Funnell et al. with 133 Single-Cells



Funnell et al., Nature, 2022

Evol. History of another High Grade Serous Ovarian Cancer Patient's Dataset from Funnell et al. with 1283 Single-Cells



Funnell et al., Nature, 2022

Summary of Sc-TUSV-ext

- We introduce the first phylogenetic tree reconstruction method that incorporates SNVs, CNAs and SVs from single-cell whole-genome sequences.
- Though simulations and real datsets, we saw that different tumor clones are driven by different types of somatic variants.
- One future direction is to extend our method to deal with complex structural variants like chromothripsis events.
- Another future direction is to use a better clustering method for the single-cells using all the variants.

Acknowledgements

Schwartz Lab

Prof. Russell Schwartz

Xuecong Fu

Thomas Rachman

Arjun Srivatsa

Lanting Li

Yueqian Deng

Kefan Cao

Alan Luo

Minhang Xu

Github

bioRxiv