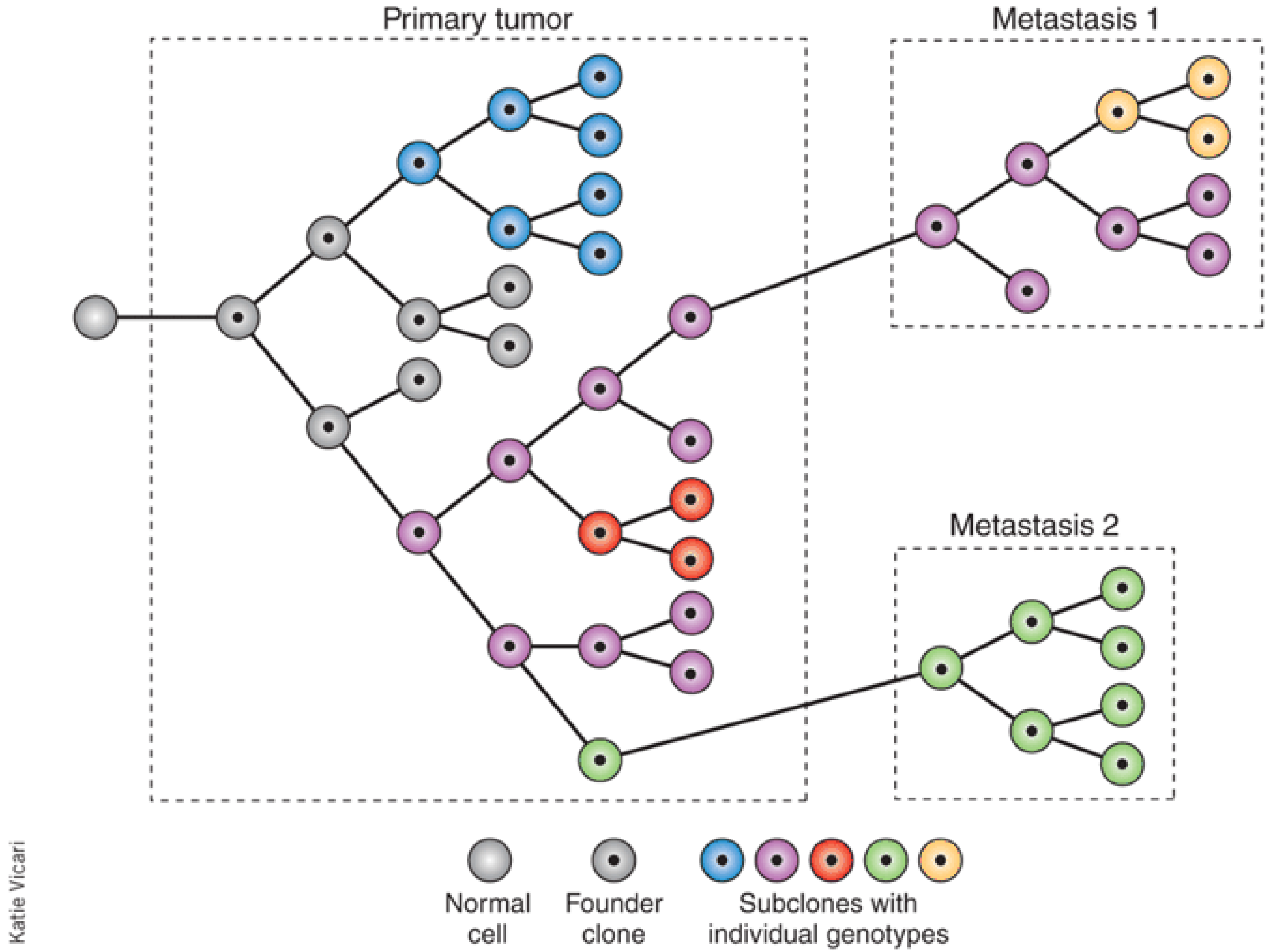
ALGORITHMIC CHALLENGES IN RECONSTRUCTING COPY-NUMBER EVOLUTION

Manuel Lafond

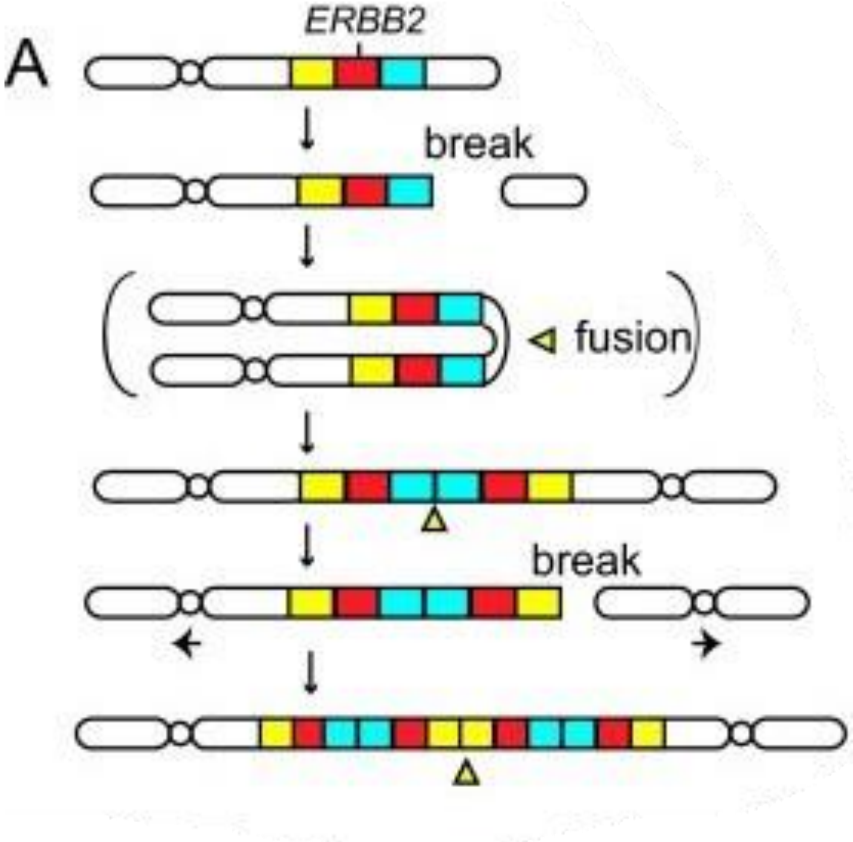
UNIVERSITÉ DE SHERBROOKE

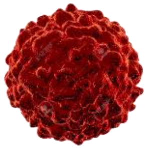
Tumours and evolution

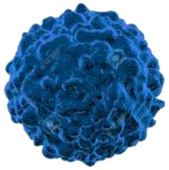


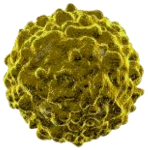
- Tumor cells undergo rapid genomic evolutionary changes
  - Rearrangements
  - Amplifications
  - Deletions

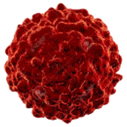
  - e.g. : breakage-fusion-bridge cycle

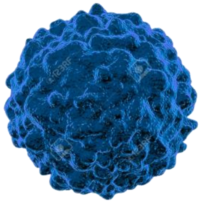


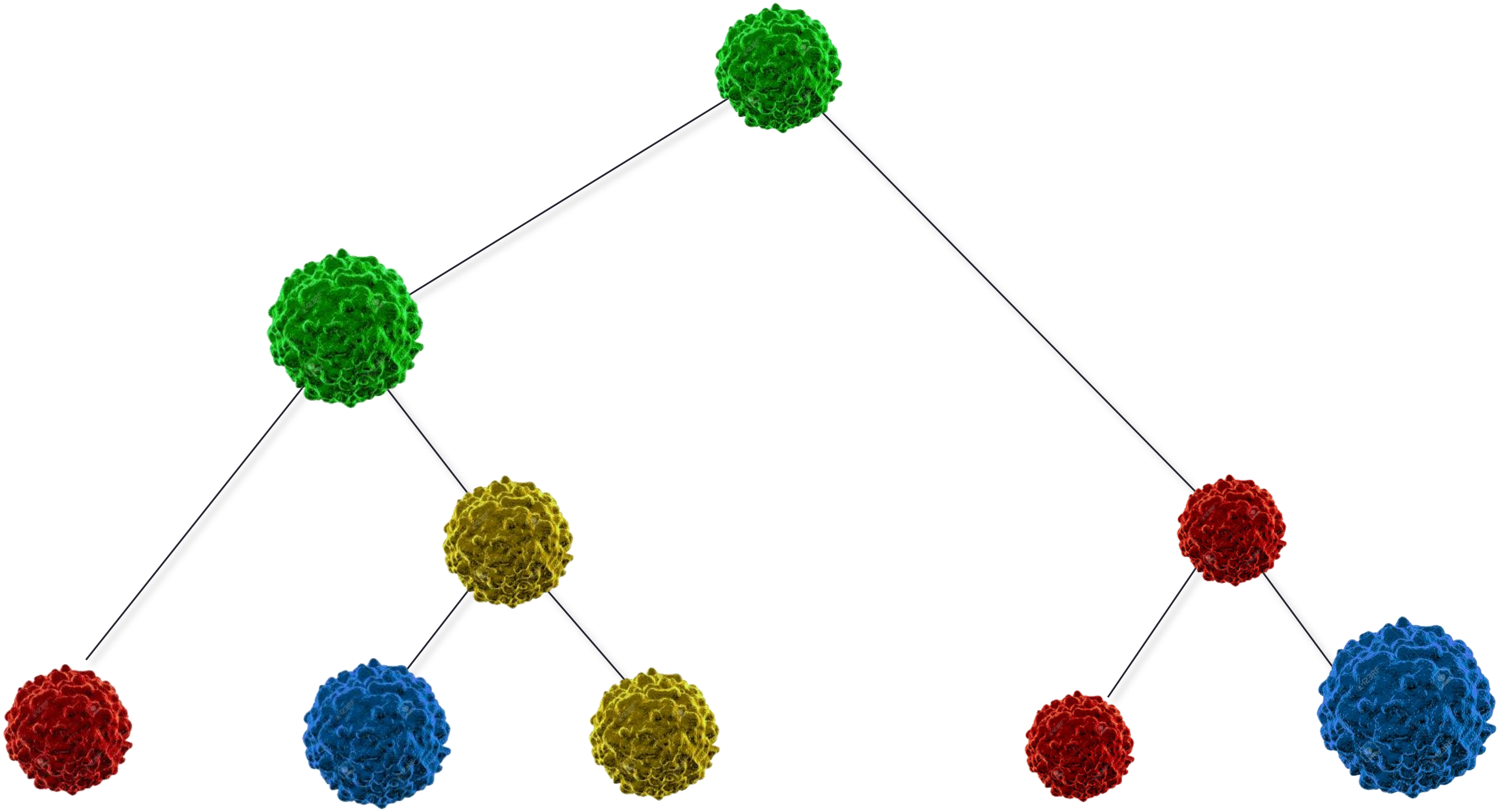












Single cell sequencing

- Assessment of tumor heterogeneity
- Cells within a tumor undergo somatic mutations

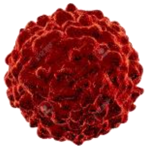
  => Different rearrangements, duplications, deletions

  => Difficult to obtain complete genome of every cell

- Copy-numbers are easier to obtain.
  - High coverage = expensive $$
  - Accuracy at low coverage now possible

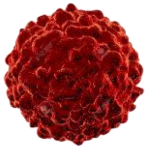
    (Direct libray preparation, 10X Genomics, ...)

Copy-Number Profile (CNP)

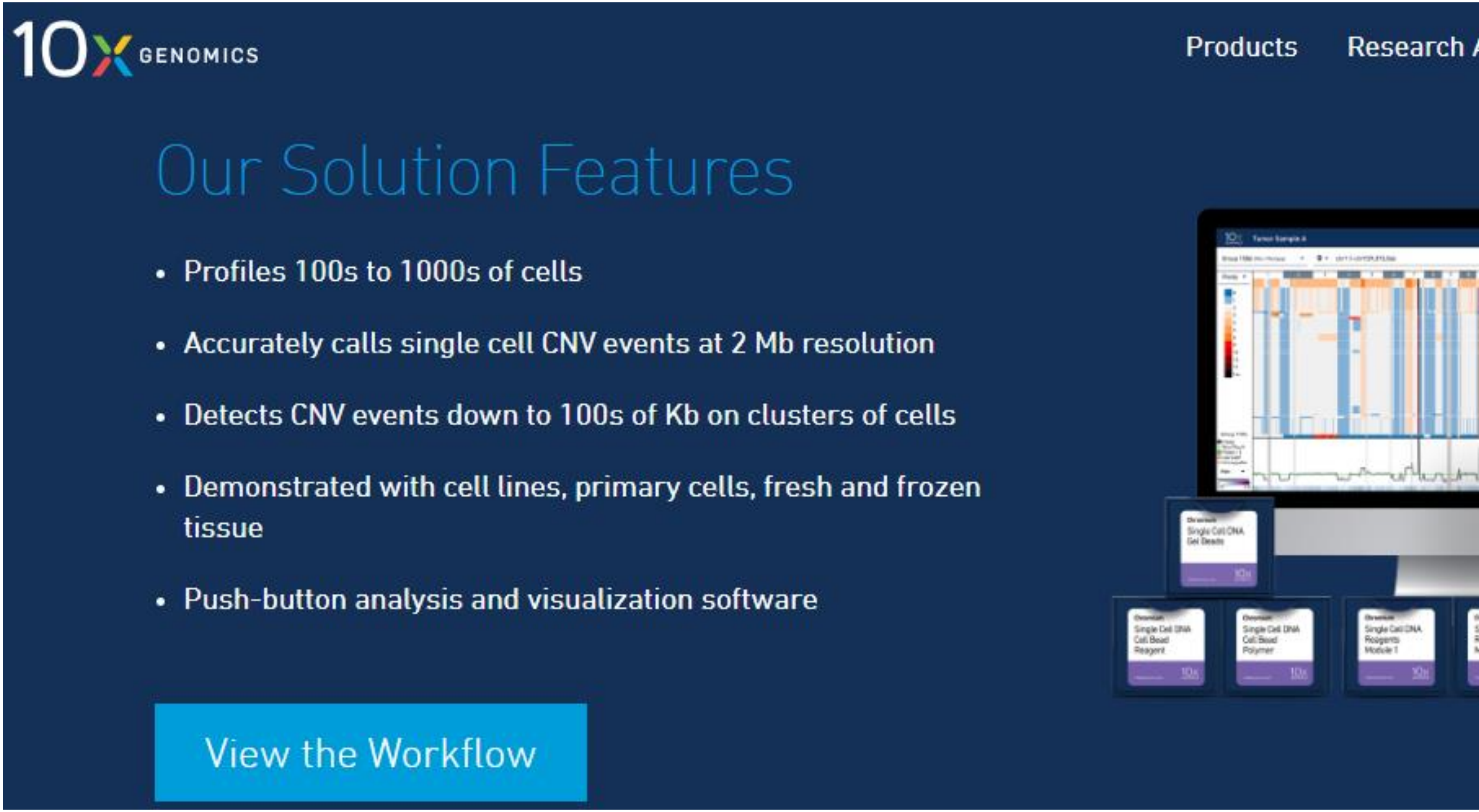


10 copies of segment 1
5 copies of segment 2
0 copies of segment 3
4 copies of segment 4
10 copies of segment 5

Copy-Number Profile (CNP)



(10, 5, 0, 4, 10)


10x GENOMICS
Products
Research
Single Cell DNA Gel Beads
Single Cell DNA Cell Bead Reagent
Single Cell DNA Cell Bead Polymer
Single Cell DNA Reagents Module 1

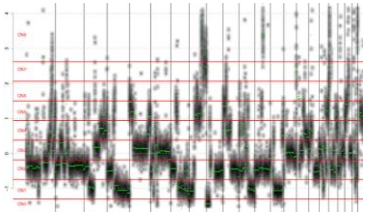
Our Solution Features

- Profiles 100s to 1000s of cells
- Accurately calls single cell CNV events at 2 Mb resolution
- Detects CNV events down to 100s of Kb on clusters of cells
- Demonstrated with cell lines, primary cells, fresh and frozen tissue
- Push-button analysis and visualization software

View the Workflow

Sequence cells from same tumor (single cell sequencing)

Infer copy-number for each segment of interest (for each allele)



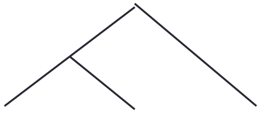
Phasing: assign copy-number to allele, get chromosome Copy-Number Profile (CNP)

Major: (4, 4, 12, 0, 2)
Minor: (4, 3, 10, 0, 1)

Compare CNPs of each pair of cells => distance matrix

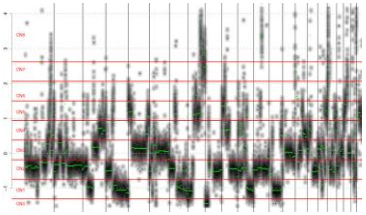
0	16	47	72	77	79
16	0	37	57	65	66
47	37	0	40	30	35
72	57	40	0	31	23
77	65	30	31	0	10
79	66	35	23	10	0

Reconstruct phylogeny



Sequence cells from same tumor (single cell sequencing)

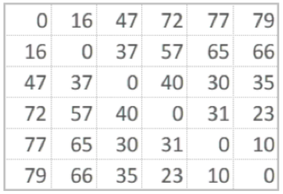
Infer copy-number for each segment of interest (for each allele)



Phasing: assign copy-number to allele, get chromosome Copy-Number Profile (CNP)

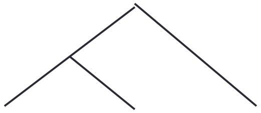
Major: (4, 4, 12, 0, 2)
Minor: (4, 3, 10, 0, 1)

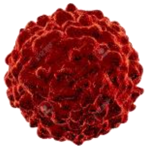
Compare CNPs of each pair of cells => distance matrix



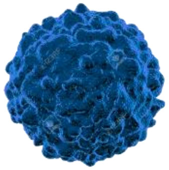
0	16	47	72	77	79
16	0	37	57	65	66
47	37	0	40	30	35
72	57	40	0	31	23
77	65	30	31	0	10
79	66	35	23	10	0

Reconstruct phylogeny

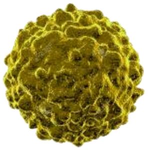




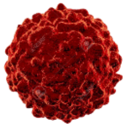
(10, 5, 0, 4, 10)



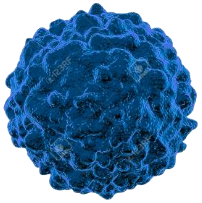
(8, 6, 2, 0, 8)



(1, 4, 4, 4, 8)



(6, 6, 2, 0, 0)



(7, 7, 3, 0, 3)

If we know the Copy-Number Profile (CNP) of each segment of interest in several tumor cells, what can we say about their evolution?

In this talk

- Comparing integer vectors.

- Comparing genomes an CNPs.

- Comparing integer vectors (with rearrangements).

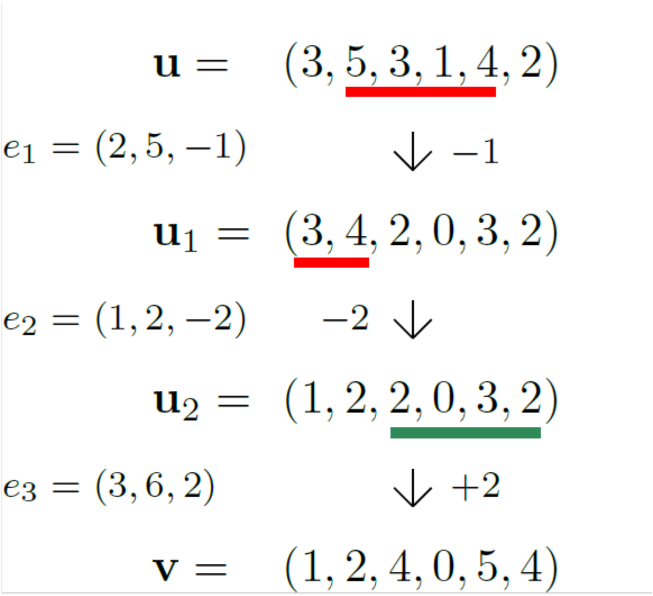
In this talk

- Comparing integer vectors.
  - We have no idea how!

- Comparing genomes an CNPs.
  - We have no idea how!

- Comparing integer vectors (with rearrangements).
  - We have no idea how!

Comparing Copy-Number Profiles



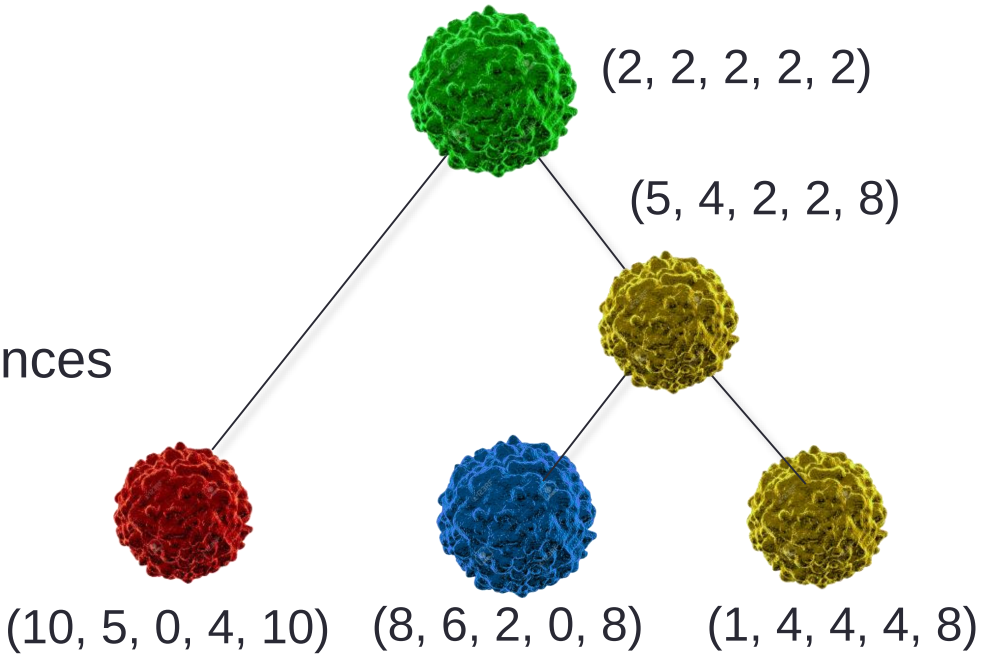
$$\begin{aligned}
& \mathbf{u} = && (3, \underline{5, 3, 1, 4}, 2) \\
e_1 = (2, 5, -1) & && \quad\quad \downarrow -1 \\
& \mathbf{u}_1 = && (\underline{3, 4}, 2, 0, 3, 2) \\
e_2 = (1, 2, -2) & && \quad -2 \downarrow \\
& \mathbf{u}_2 = && (1, 2, \underline{2, 0, 3, 2}) \\
e_3 = (3, 6, 2) & && \quad\quad \downarrow +2 \\
& \mathbf{v} = && (1, 2, 4, 0, 5, 4)
\end{aligned}$$

Why compute distances

- Classic approach 1
  - compute $dist(u, v)$ for each pair u, v
  - get a distance matrix
  - use phylogenetic distance method (e.g. NJ)

	x	y	z
x	5	8	6
y		5	7
z			2

- Classic approach 2
  - infer ancestral CNP states
  - minimize sum of branch distances



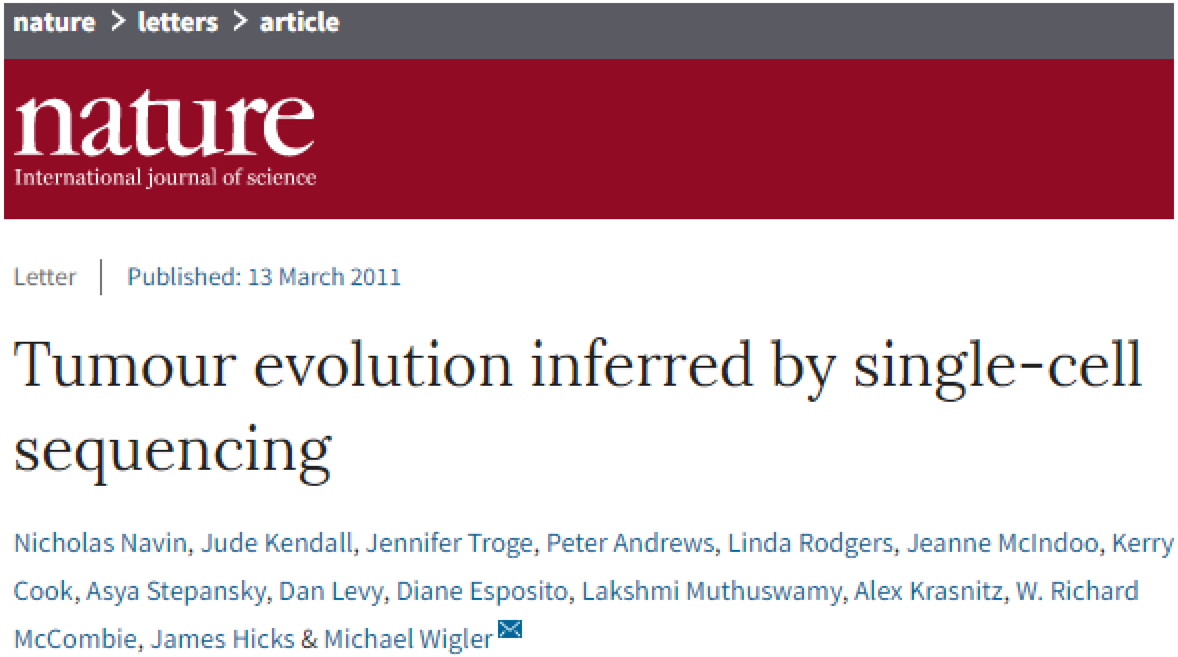
The story so far

1. Let's use the Euclidean distance (2011)

2. Let's model segmental events on integer vectors (2014)
   - Even if minimizing events takes exponential time...
   - No actually it takes polynomial time (2017)

3. Let's weight events by their amplitude (2019)

4. Let's weight events by their length/location (upcoming...)



nature > letters > article

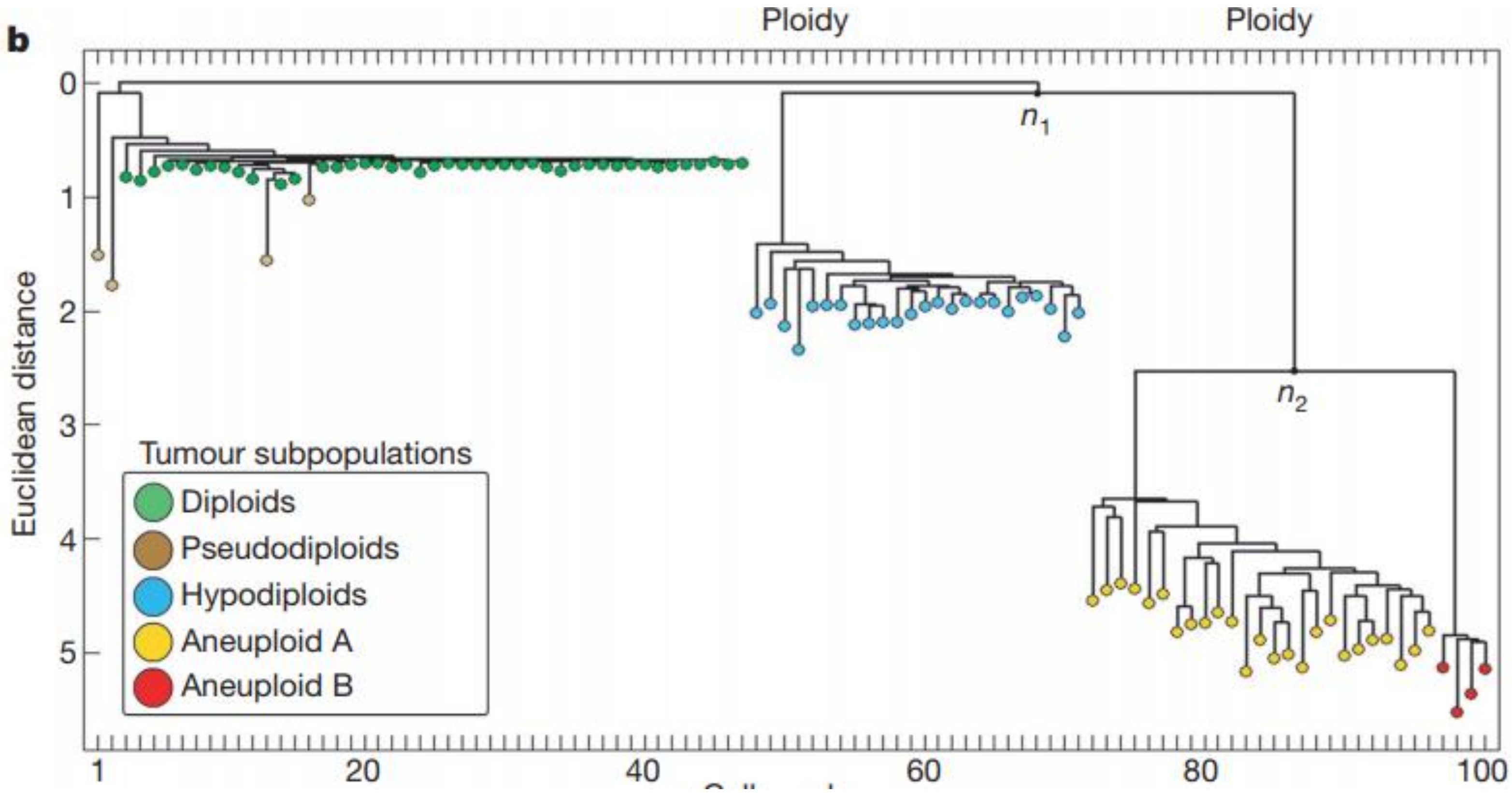
nature
International journal of science

Letter | Published: 13 March 2011

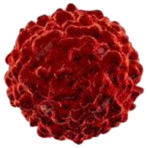
Tumour evolution inferred by single-cell sequencing

Nicholas Navin, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, Lakshmi Muthuswamy, Alex Krasnitz, W. Richard McCombie, James Hicks & Michael Wigler ✉

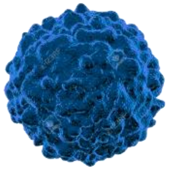
- Sequenced 100 cells from a tumor, reconstructed NJ phylogeny from CNP data.


b
Ploidy
Ploidy
0
1
2
3
4
5
Euclidean distance
n_1
n_2
Tumour subpopulations
Diploids
Pseudodiploids
Hypodiploids
Aneuploid A
Aneuploid B
1
20
40
60
80
100

- In Navin et al. [Nature11] : Euclidean distance
- $dist(\boldsymbol{u},\boldsymbol{v}) = \sqrt{\sum(u_i \ \_ \ vi)^2}$



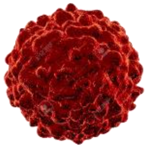
(2, 5)



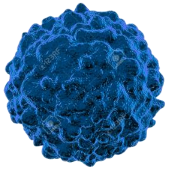
(5, 1)

dist = $\sqrt{9 + 16} = 5$

- In Navin et al. [Nature11] : Euclidean distance
- $dist(\boldsymbol{u}, \boldsymbol{v}) = \sqrt{\sum(u_i \ \_ \ vi)^2}$



(2, 5)



(5, 1)

dist = $\sqrt{9 + 16} = 5$

- Implicit assumption: positions are independent.

In Schwarz et al. [PlosCB14]: MEDICC model

- Positions should NOT be independent!
- Events can affect segments of genomes

CNP-2-CNP problem – MEDICC model

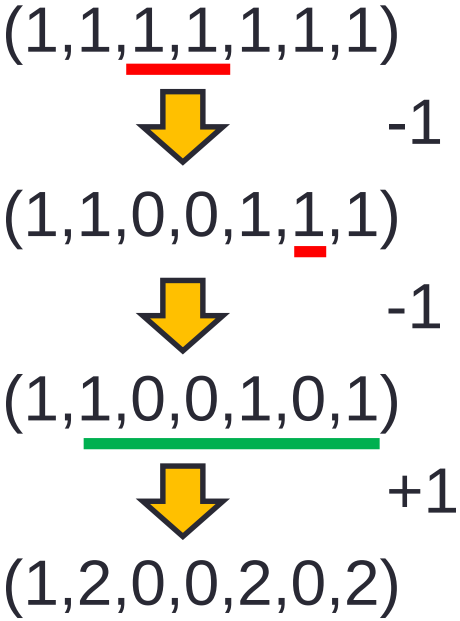
- **Given**: two CNPs $\boldsymbol{u}$ and $\boldsymbol{v}$ (integer vectors)
- **Move**: alter an interval of $\boldsymbol{u}$ by +1/-1 (a 0 stays a 0).
- **Find**: min # of moves to turn $\boldsymbol{u}$ into $\boldsymbol{v}$

(1,1,1,1,1,1,1)

(1,2,0,0,2,0,2)

CNP-2-CNP problem – MEDICC model

- **Given**: two CNPs $\boldsymbol{u}$ and $\boldsymbol{v}$ (integer vectors)
- **Move**: alter an interval of $\boldsymbol{u}$ by +1/-1 (a 0 stays a 0).
- **Find**: min # of moves to turn $\boldsymbol{u}$ into $\boldsymbol{v}$



(1,1,1,1,1,1,1)

-1

(1,1,0,0,1,1,1)

-1

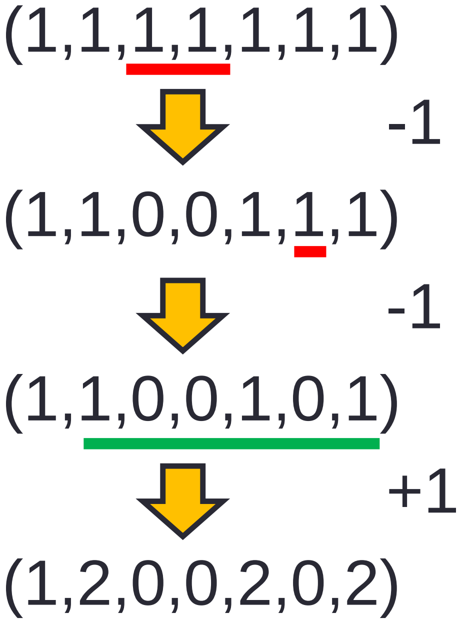
(1,1,0,0,1,0,1)

+1

(1,2,0,0,2,0,2)

CNP-2-CNP problem – MEDICC model

- In Schwarz & al [PlosCB14]
  - Compute $d(u,v)$ in time $\Omega(3^N)$,N = max copy-number



$(1,1,\underline{1,1},1,1,1)$

-1

$(1,1,0,0,1,\underline{1},1)$

-1

$(1,\underline{1,0,0,1,0,1})$

+1

$(1,2,0,0,2,0,2)$

ZZS algorithm

- In Zeira, Zehavi, Shamir [JCB17]:
  - Better algorithms to compute min # of +1/-1 moves
  - Simple DP pseudo-polynomial time algorithm $O(nN^2)$
  - More involved $O(n)$ time algorithm.

ZZS algorithm

- In Zeira, Zehavi, Shamir [JCB17]:
  - Better algorithms to compute min # of +1/-1 moves
  - Simple DP pseudo-polynomial time algorithm O(nN^2)
  - More involved O(n) time algorithm.

- General idea:
  - Show that some optimal solution does all deletions before amplifications.
  - Dynamic programming, optimal for every prefix from left to right.
  - $M[i,d]$ = optimal for i if i-th value is d.

$$\mathrm{M}[i,d] \leftarrow \min_{0\leq d'\leq N} \{\mathrm{M}[\mathrm{prev}(i),d'] + \max\{d-d',0\} + \max\{a(i,d) - a(\mathrm{prev}(i),d'),0\}$$

$$+ \max\{Q_i - \max\{d,d'\},0\}\}.$$

At Recomb-CG 2019

Same problem, but not restricted to +1/-1.

At Recomb-CG 2019

Same problem, but not restricted to +1/-1.

A single event could double copy numbers (e.g. WGD).

CNP-2-CNP problem – extended MEDICC

- **Given**: two CNPs $\boldsymbol{u}$ and $\boldsymbol{v}$ (integer vectors), cost function f
- **Move**: alter a contiguous interval of $\boldsymbol{u}$ by **any** amount.
- **Find**: min # of moves to turn $\boldsymbol{u}$ into $\boldsymbol{v}$

CNP-2-CNP problem – extended MEDICC

- Difference vector $\boldsymbol{w} = \boldsymbol{u} - \boldsymbol{v}$
- Intuition: "squish" values of $\boldsymbol{w}$ to 0.

$$\mathbf{u} = (3, \underline{5, 3, 1, 4}, 2)$$

$$e_1 = (2, 5, -1) \qquad \downarrow -1$$

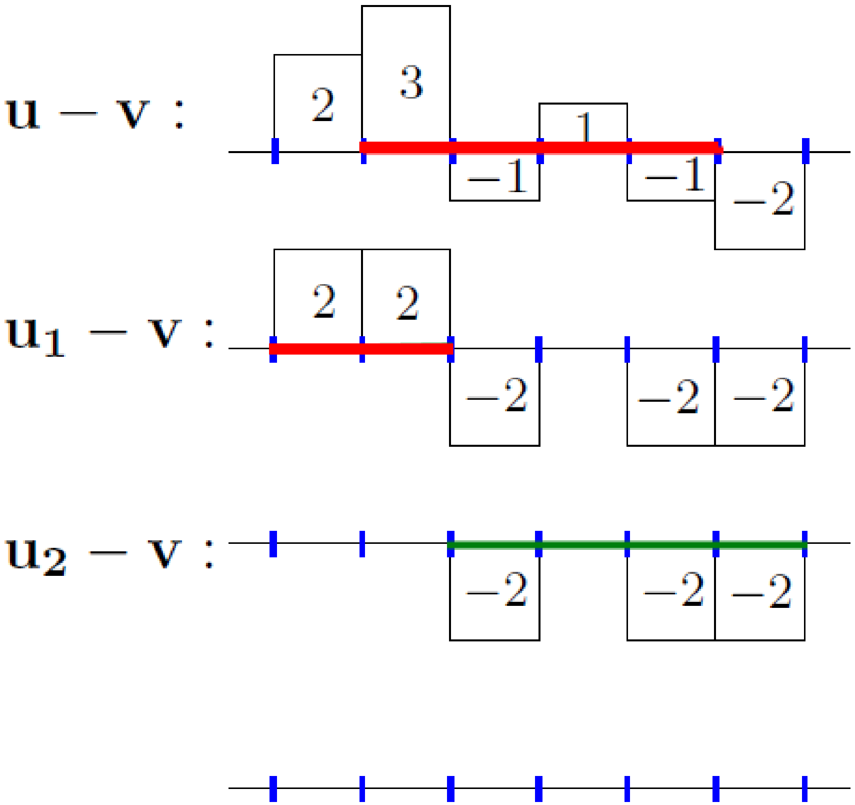
$$\mathbf{u}_1 = (\underline{3, 4}, 2, 0, 3, 2)$$

$$e_2 = (1, 2, -2) \qquad -2 \downarrow$$

$$\mathbf{u}_2 = (1, 2, \underline{2, 0, 3, 2})$$

$$e_3 = (3, 6, 2) \qquad \downarrow +2$$

$$\mathbf{v} = (1, 2, 4, 0, 5, 4)$$



$\mathbf{u} - \mathbf{v}$:
2
3
1
−1
−1
−2
$\mathbf{u}_1 - \mathbf{v}$:
2
2
−2
−2
−2
$\mathbf{u}_2 - \mathbf{v}$:
−2
−2
−2

Theorem

In the extended MEDICC model, the CNP-2-CNP problem is **strongly** NP-hard.

(strongly => hard even if the numbers are polynomial in n)

Positive results

Theorem

If the CNP's have no 0-positions, there is a linear time factor 2 approximation algorithm for the extended CNP-2-CNP problem.

The algorithm

Return the number of flat intervals in the difference vector.

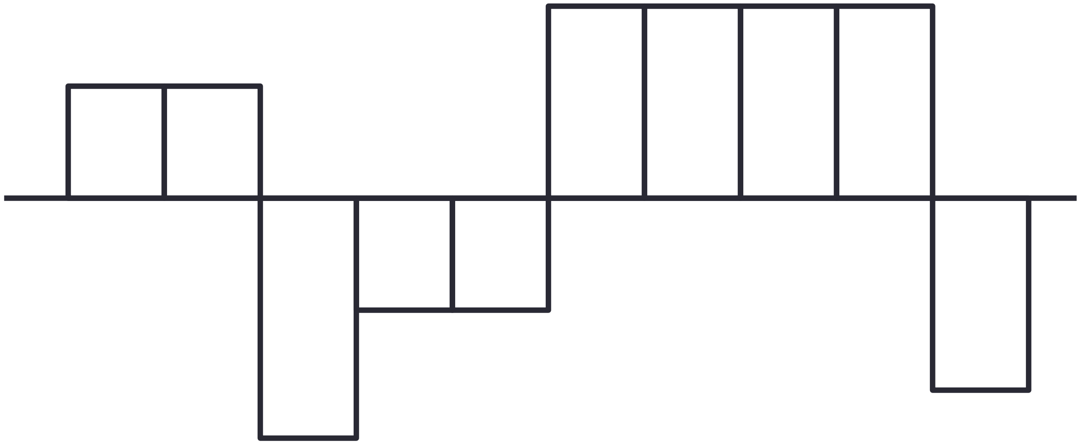
Flat interval = contiguous positions in which difference vector has same value.

Below: 5 flat intervals

Lemma

One moves reduces number of flat intervals by at most 2

=> $dist_{any}(\boldsymbol{u}, \boldsymbol{v})$ is at least ½ the number of flat intervals.



Flat interval = contiguous positions in which difference vector has same value.

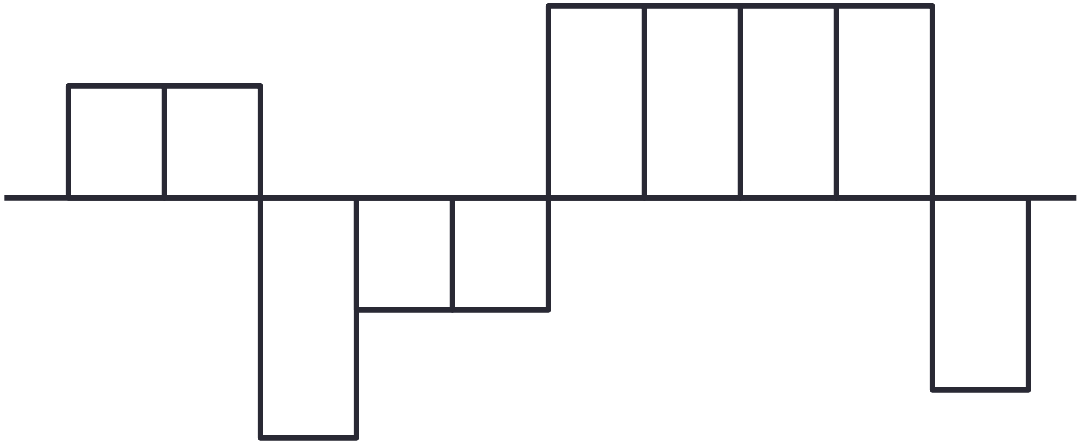
Below: 5 flat intervals

Lemma

One moves reduces number of flat intervals by at most 2

=> $dist_{any}(\boldsymbol{u}, \boldsymbol{v})$ is at least ½ the number of flat intervals.

Trivial 2-approx: remove each flat interval one by one.



Experiments

- If we simulate amplifications and deletions on genomes (and not CNPs), can we reconstruct phylogenies?

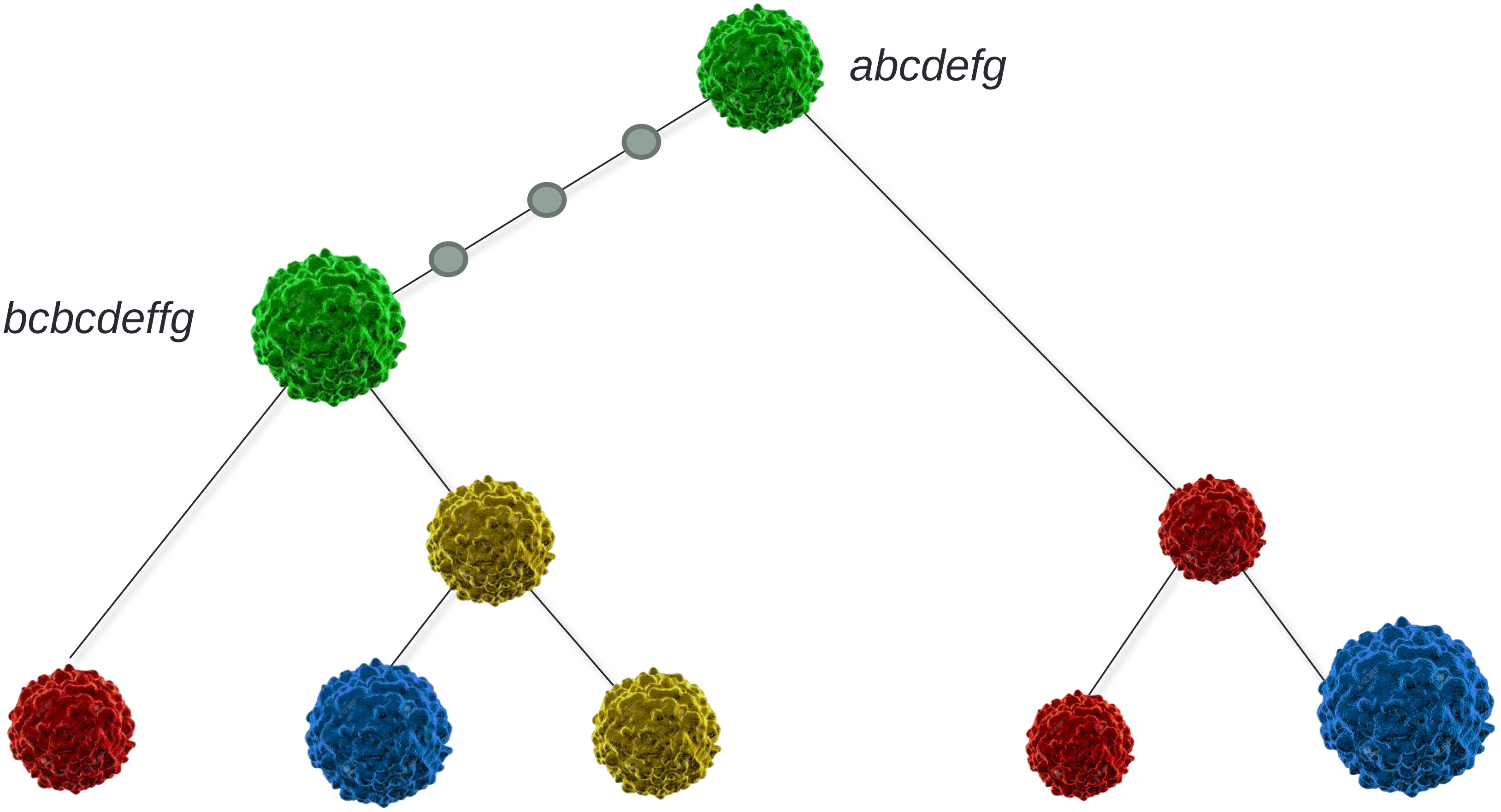
n = number of leaves in random tree (def. n = 100)

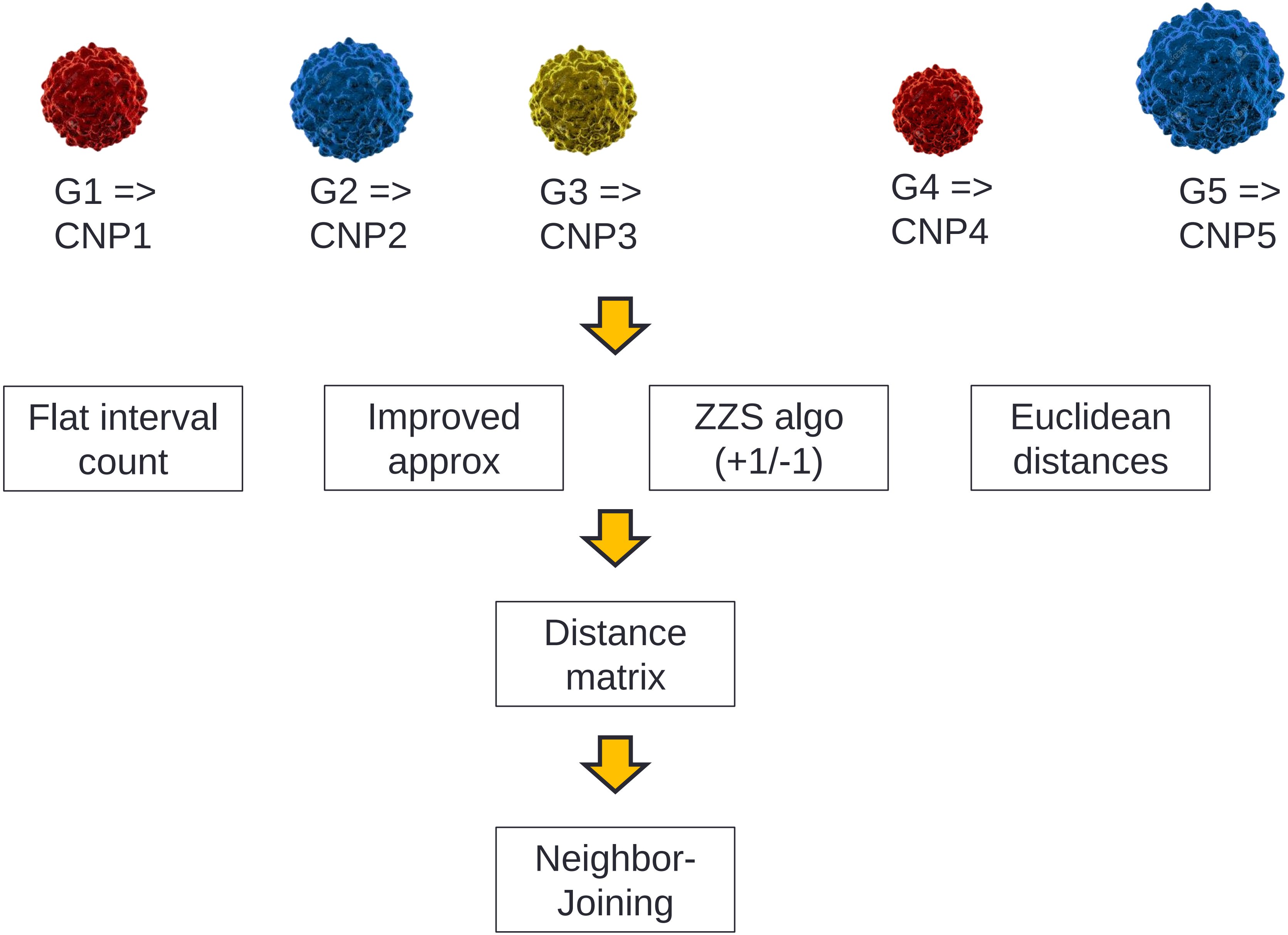
l = number of genes per chromosome (def. l = 100)

∂ = prob. of duplication (def. ∂ = 0.5)

(e_1, e_2) = range of # of events per branch (def. [5..10])

(p, q) = control length of events





G1 => CNP1

G2 => CNP2

G3 => CNP3

G4 => CNP4

G5 => CNP5

Flat interval count

Improved approx

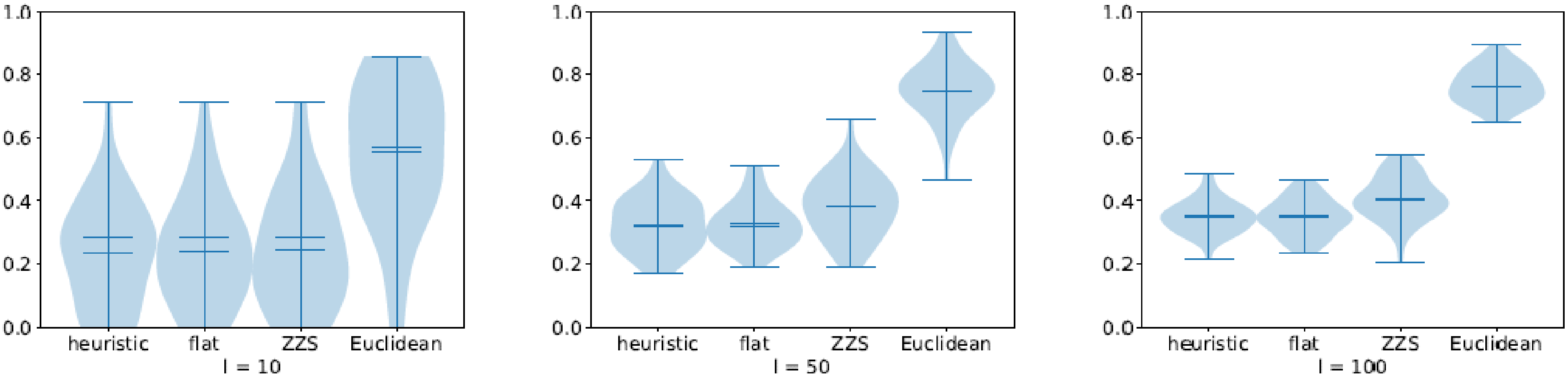
ZZS algo (+1/-1)

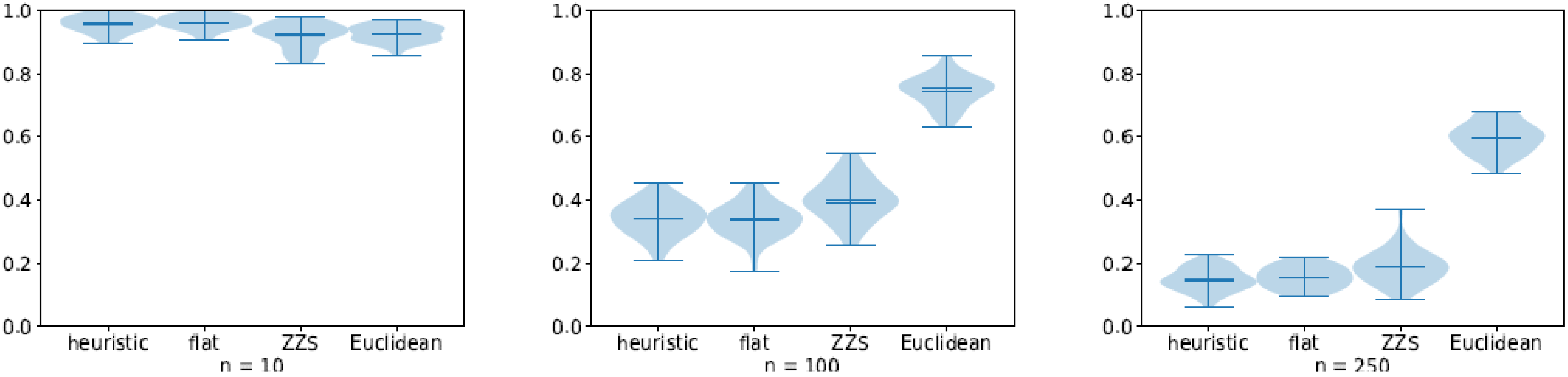
Euclidean distances

Distance matrix

Neighbor-Joining

More leaves = easier to predict





More genes = easier to predict

Extended-extended MEDICC model

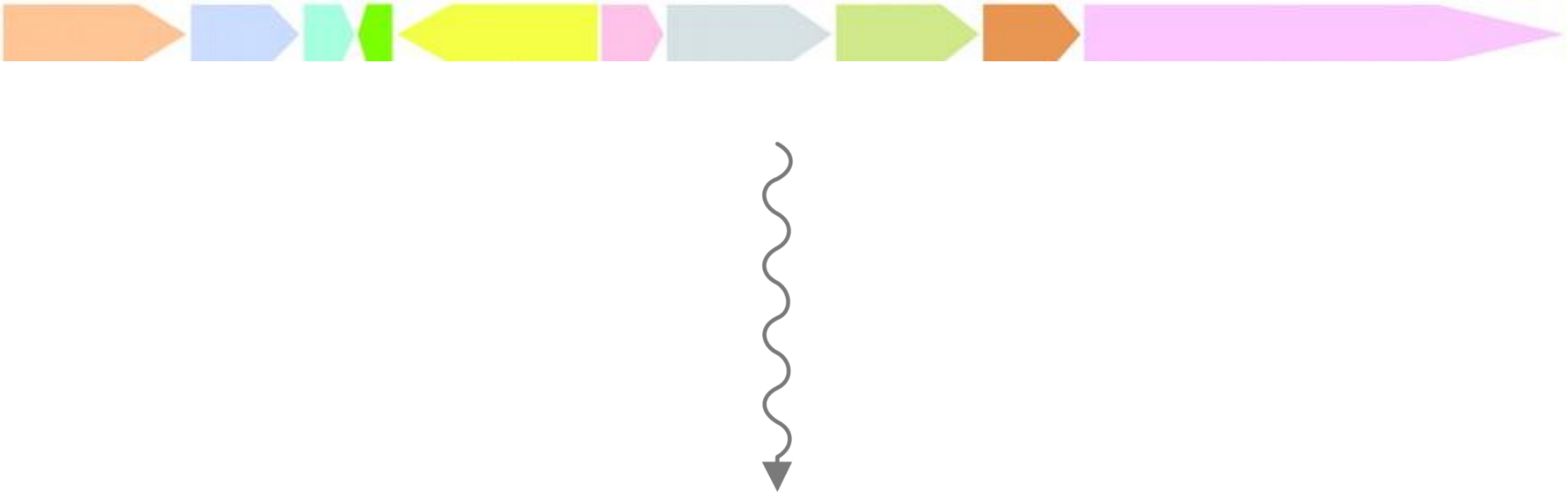
- Some events are more likely to affect certain regions of the genome.
  - e.g. arm duplications => ends of CNP vector more susceptible to amplification

- Extended model : each interval $[i \ldots j]$ has its own weight.

- Weights can be inferred from cancer patient data.

- (not published yet, and not my work)

Comparing Genomes with Copy-Number Profiles



(2, 5, 0, 4, 3)

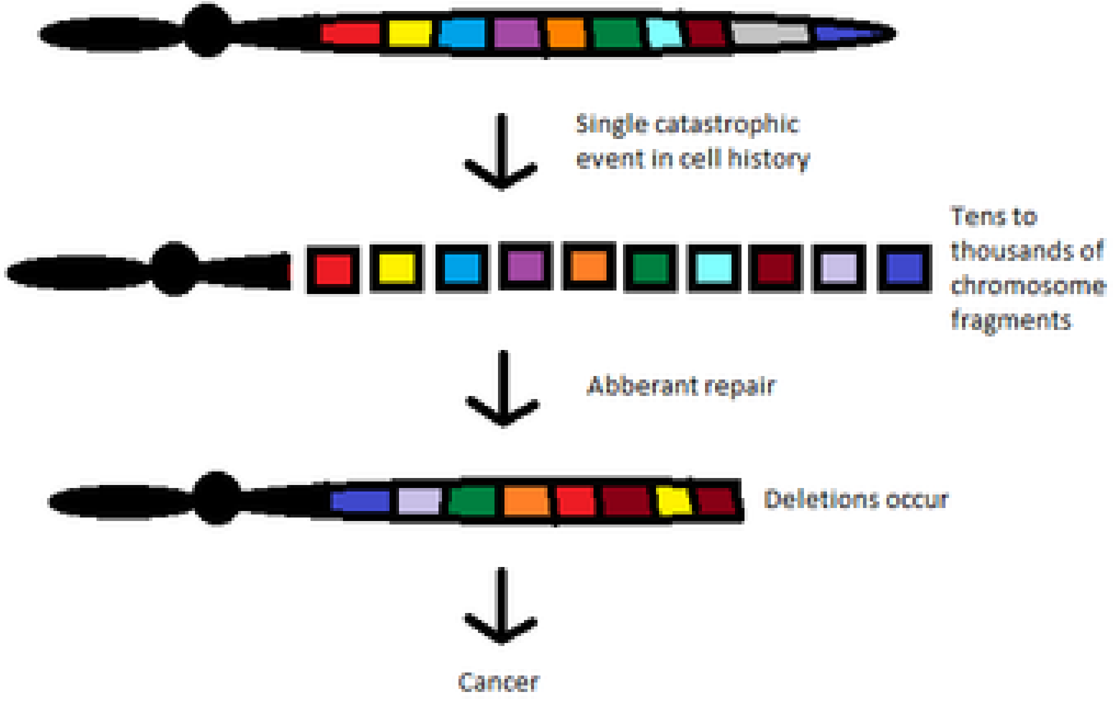
Problems with segmental events on CNPs

- Assumes that order of segments remains fixed.
- Rearrangements change the order.
  - Some, drastically.

Problems with segmental events on CNPs

- Assumes that order of segments remains fixed.
- Rearrangements change the order.
  - Some, drastically.

Chromothripsis



Single catastrophic event in cell history
Tens to thousands of chromosome fragments
Abberant repair
Deletions occur
Cancer

Problems with segmental events on CNPs

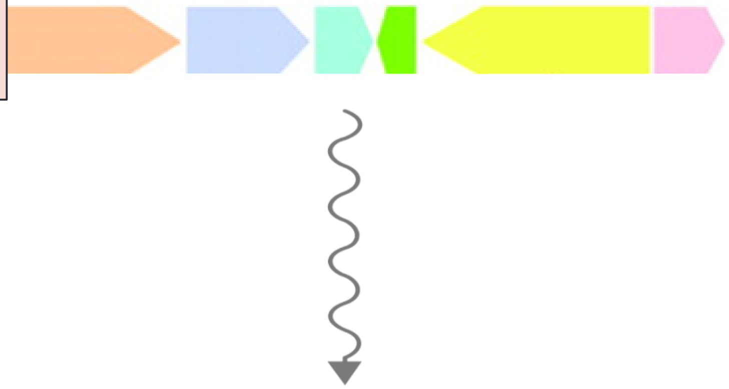
- Assumes that order of segments remains fixed.
- Rearrangements change the order.
  - Some, drastically.

- Introduce actual rearrangements into the model.

Genome-to-CNP

- In [Zhu & al, ACM-BCB 2018]:
- Given the CNP of of a single cell C, infer the rearrangements that occurred from a healthy genome to C.

Normal human genome

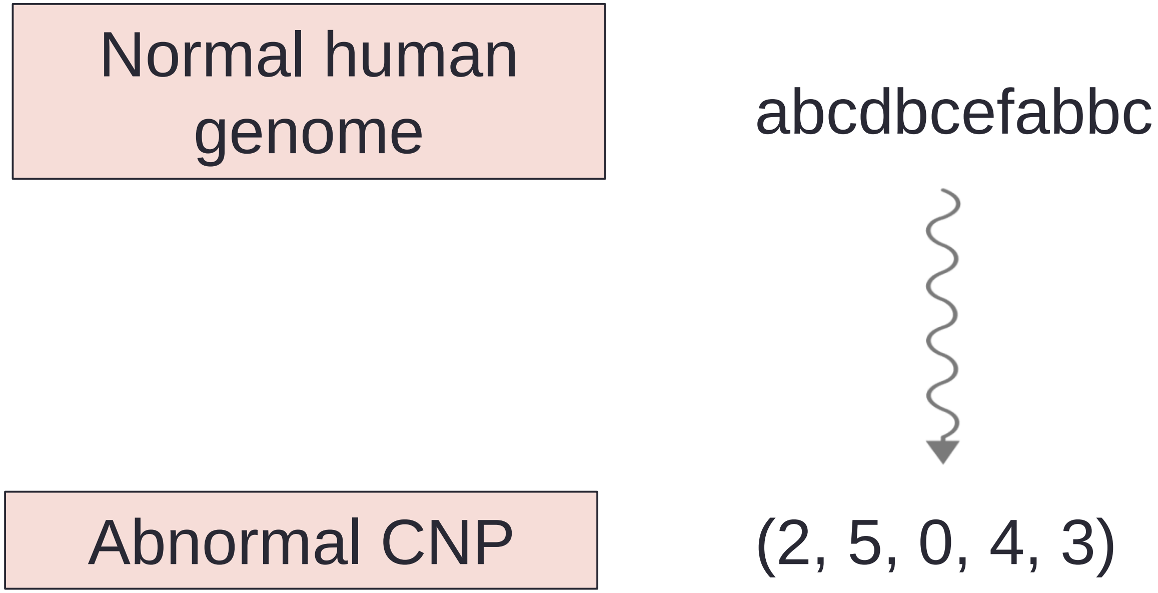


Abnormal CNP

(2, 5, 0, 4, 3)

Genome-to-CNP

- In [Zhu & al, ACM-BCB 2018]:
- Given the CNP of of a single cell C, infer the rearrangements that occurred from a healthy genome to C.



Normal human genome

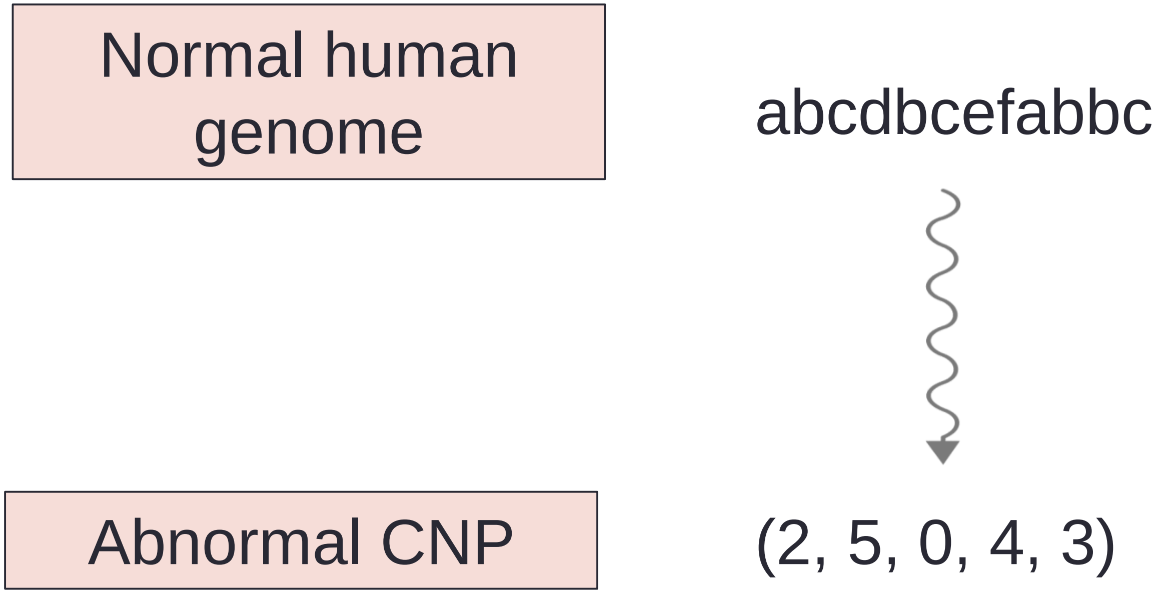
abcdbcefabbc

Abnormal CNP

(2, 5, 0, 4, 3)

Genome-to-CNP

- In [Zhu & al, ACM-BCB 2018]:
- Given the CNP of of a single cell C, infer the rearrangements that occurred from a healthy genome to C.
- Allowed: segmental duplications & deletions.



Normal human genome

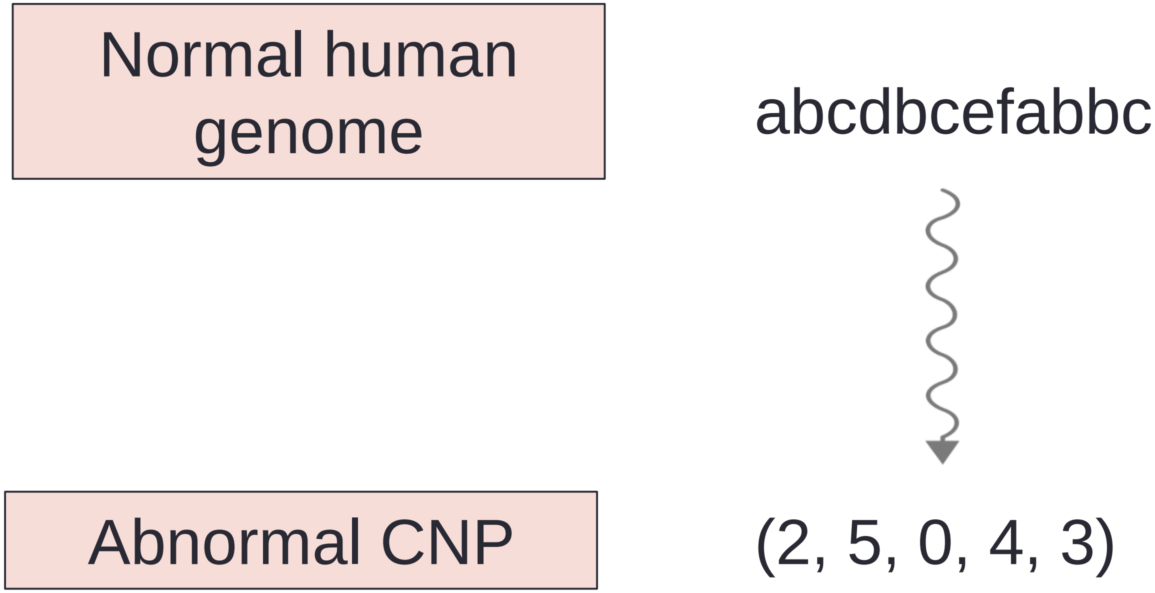
abcdbcefabbc

Abnormal CNP

(2, 5, 0, 4, 3)

Genome-to-CNP

- **Given**: string S and copy-number vector C
- **Move**: segmental duplications and deletions.
- **Find**: min # of moves to turn S into any T whose CNP is C



Normal human genome

abcdbcefabbc

Abnormal CNP

(2, 5, 0, 4, 3)

Example

abcd

(2, 4, 1, 3)

Need
2 x a
4 x b
1 x c
3 x d

Example

abcd

abcdbcd

Segmental duplication
(this one is tandem)

(2, 4, 1, 3)

Need
2 x a
4 x b
1 x c
3 x d

Example

abcd

ab~~c~~dbcd

abdbcd

Deletion (this one is not segmental)

(2, 4, 1, 3)

Need
2 x a
4 x b
1 x c
3 x d

Example

abcd

abcdbcd

abdbcd

abdbabdbcd

(2, 4, 1, 3)

Need
2 x a
4 x b
1 x c
3 x d

Example

abcd

abcdbcd

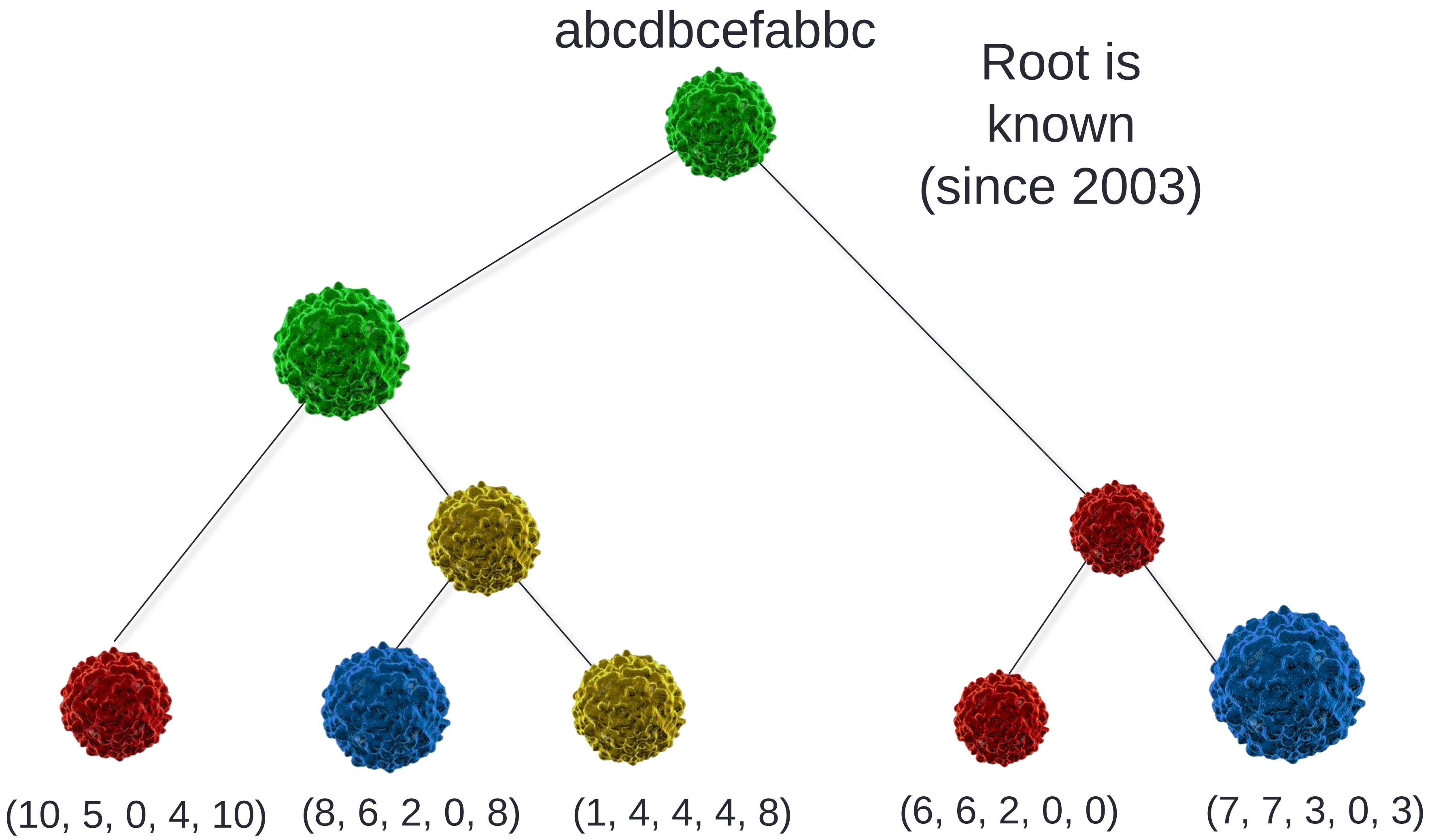
abdbcd

abdbabdbcd

(2, 4, 1, 3)

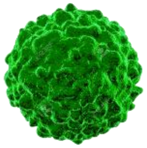
Need
2 x a
4 x b
1 x c
3 x d

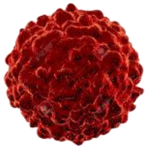
But why?


abcdbcefabbc
Root is
known
(since 2003)
(10, 5, 0, 4, 10)
(8, 6, 2, 0, 8)
(1, 4, 4, 4, 8)
(6, 6, 2, 0, 0)
(7, 7, 3, 0, 3)

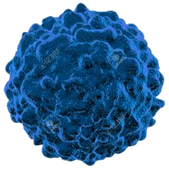
But why?

abcdbcefabbc

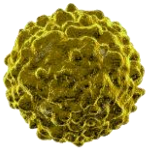




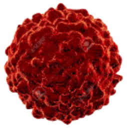
(10, 5, 0, 4, 10)



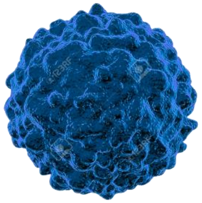
(8, 6, 2, 0, 8)



(1, 4, 4, 4, 8)

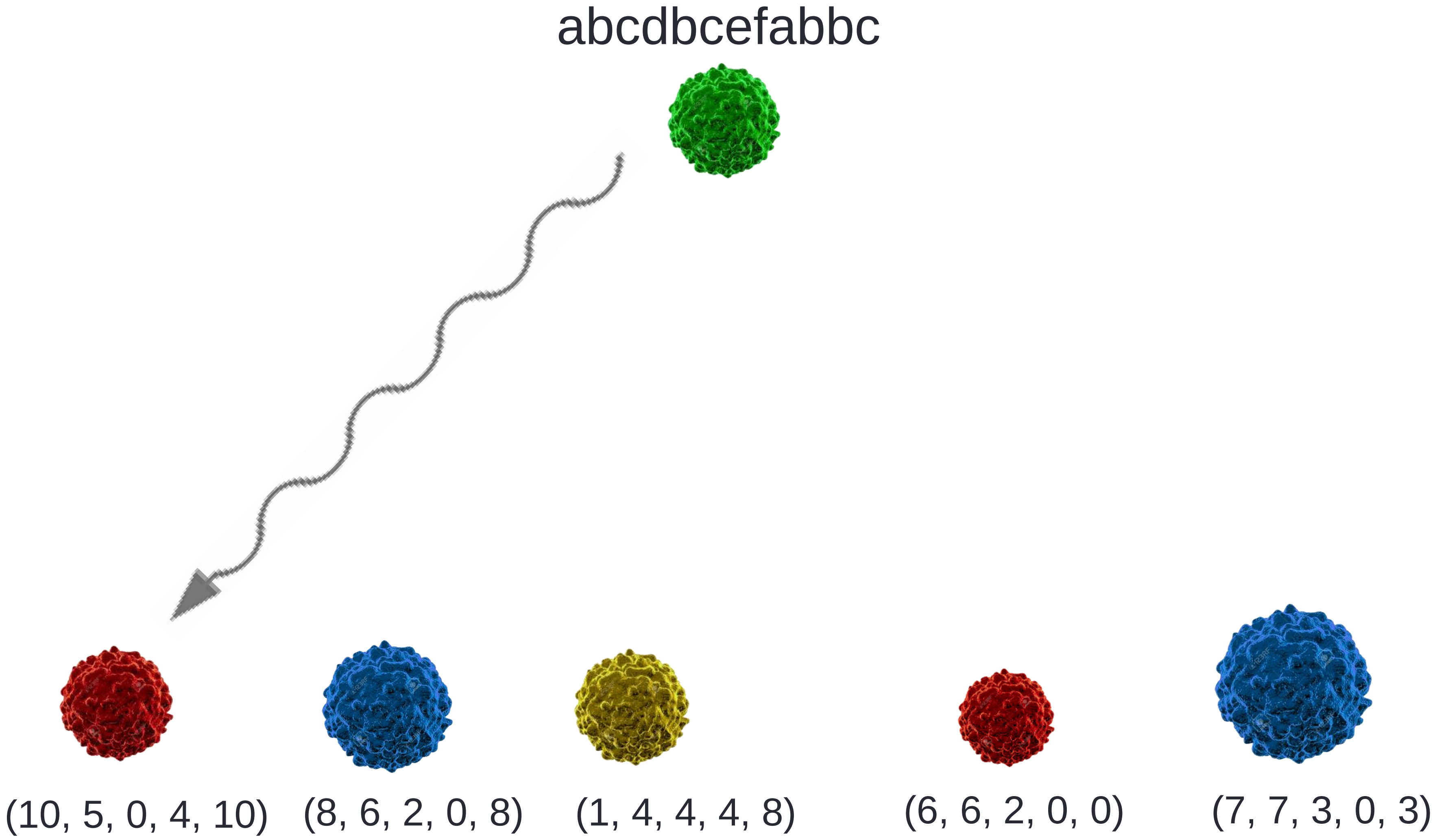


(6, 6, 2, 0, 0)



(7, 7, 3, 0, 3)

But why?

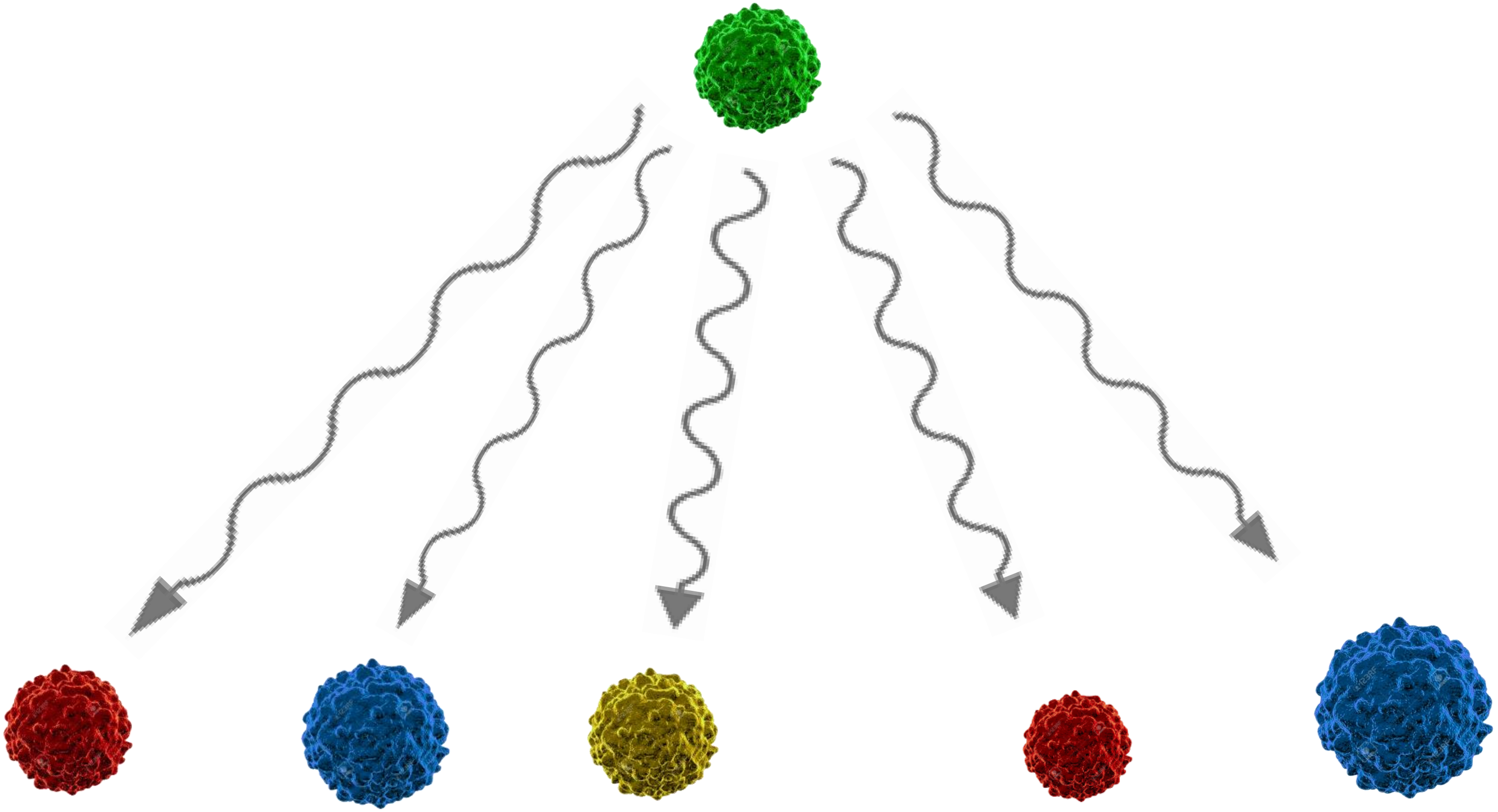


abcdbcefabbc

(10, 5, 0, 4, 10) (8, 6, 2, 0, 8) (1, 4, 4, 4, 8) (6, 6, 2, 0, 0) (7, 7, 3, 0, 3)

But why?

abcdbcefabbc

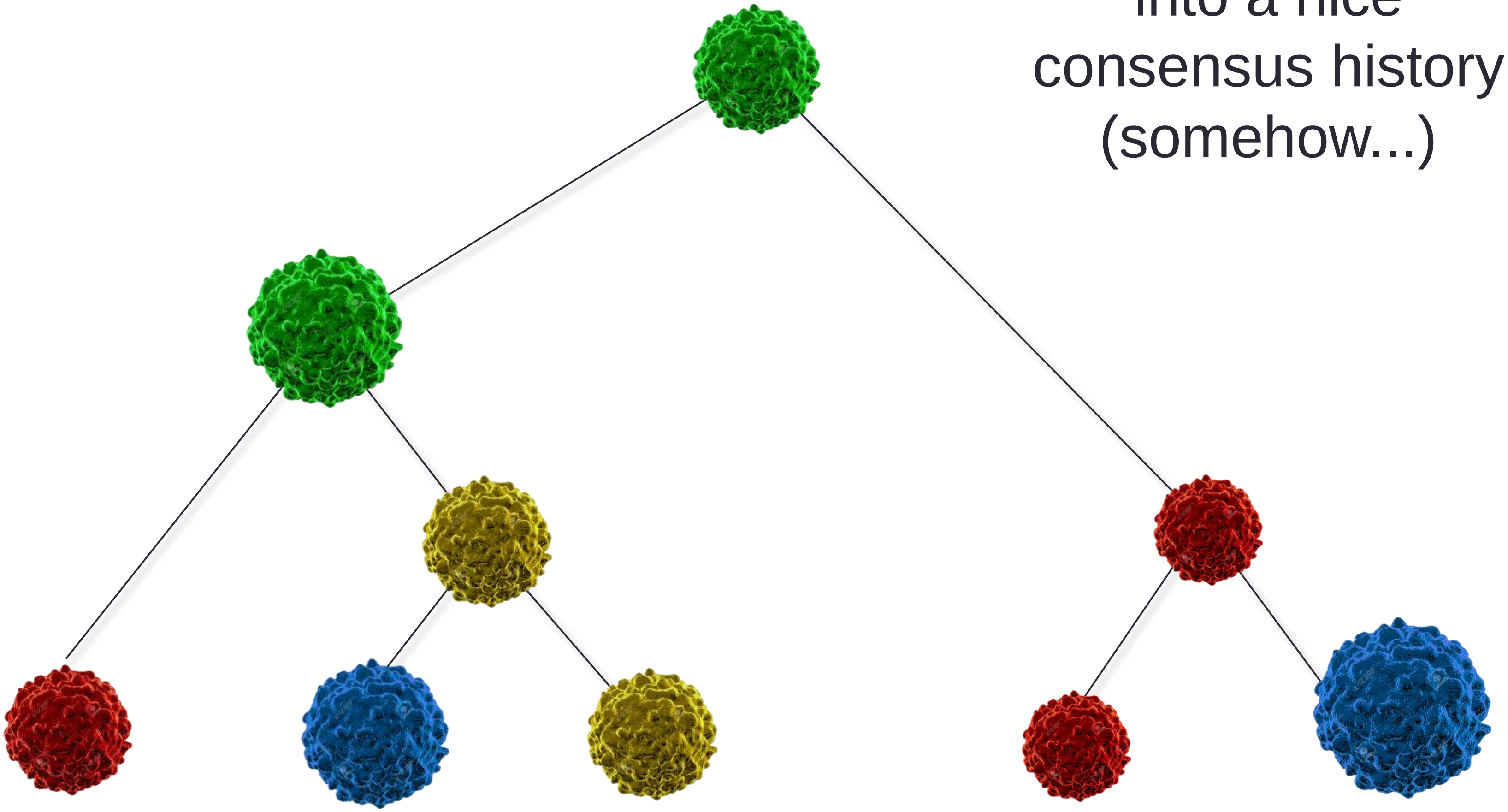


(10, 5, 0, 4, 10) (8, 6, 2, 0, 8) (1, 4, 4, 4, 8) (6, 6, 2, 0, 0) (7, 7, 3, 0, 3)

But why?

Merge all histories into a nice consensus history (somehow...)

abcdbcefabbc



(10, 5, 0, 4, 10) (8, 6, 2, 0, 8) (1, 4, 4, 4, 8) (6, 6, 2, 0, 0) (7, 7, 3, 0, 3)

Genome-to-CNP

- The problem is NP-hard [Zhu & al., 2018]
- Reduction from set-cover: design S and CNP C so that
  - each elements = 1 character
  - choosing a set = deleting elements
  - must delete one occurrence of each element

Genome-to-CNP

In [Lafond, Zhu & Zou, CPM, submitted]

Theorem

The Genome-to-CNP problem (probably) does not admit a constant factor approximation and (probably) is not FPT.

☹

Genome-to-CNP

Open problem

Find any practical approach!

Genome-to-CNP

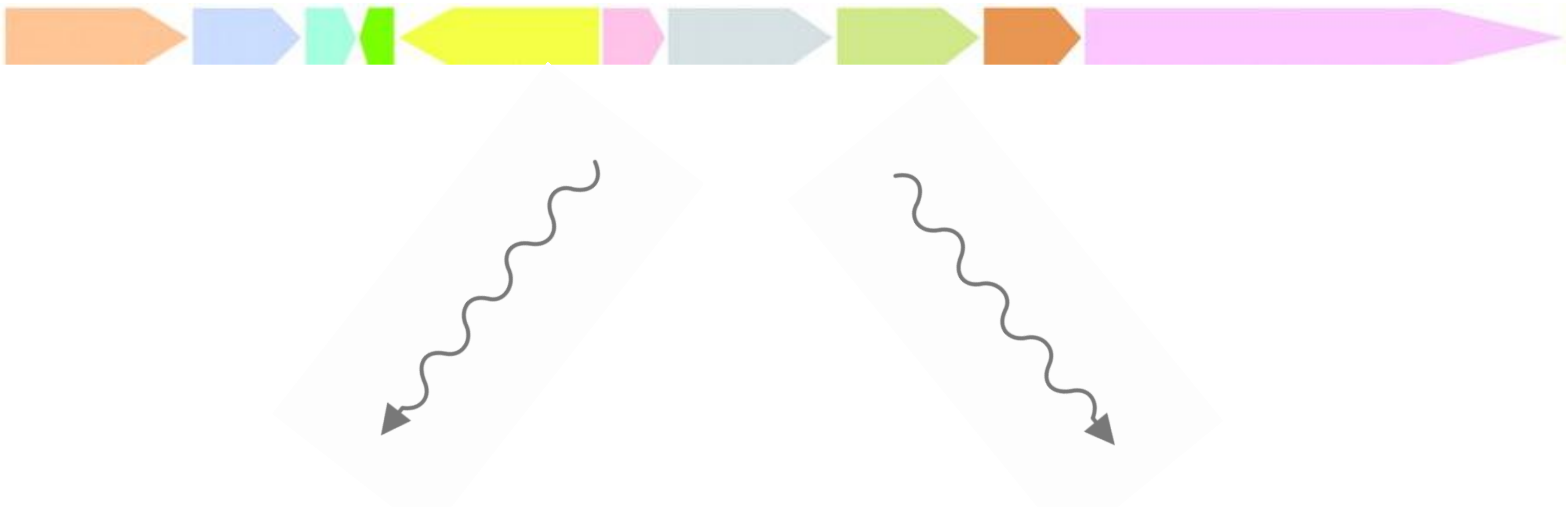
Open problem

If initial string S is exemplar, is Genome-to-CNP in P?

Exemplar = no characer occurs more than once.

Could be useful: we may model each chromosome of the healthy genome as exemplar.

Comparing Integer Vectors
(with rearrangements)



(2, 5, 0, 4, 3)

(3, 6, 2, 1, 3)

- Interval events may give rise to impossible scenarios.

$$\begin{array}{cl}
(1,1,\underline{1,1},1,1,1) & \\
\Downarrow & -1 \\
(1,1,0,0,1,\underline{1},1) & \\
\Downarrow & -1 \\
(1,\underline{1,0,0,1,0,1}) & \\
\Downarrow & +1 \\
(1,2,0,0,2,0,2) &
\end{array}$$

- Interval events may give rise to impossible scenarios.
- Compare CNPs, but require the existence of actual genomes + rearrangements.

$$(1,1,\underline{1,1},1,1,1)$$

$$\Big\downarrow \quad -1$$

$$(1,1,0,0,1,\underline{1},1)$$

$$\Big\downarrow \quad -1$$

$$(1,\underline{1,0,0,1,0,1})$$

$$\Big\downarrow \quad +1$$

$$(1,2,0,0,2,0,2)$$

Consistent CNP-2-CNP problem

- **Given**: two CNPs $\boldsymbol{u}$ and $\boldsymbol{v}$
- **Move**: segmental duplications and deletions (on genomes).
- **Find**:
  - a genome G_1 whose CNP is $\boldsymbol{u}$;
  - a genome G_2 whose CNP is $\boldsymbol{v}$;
  - such that # of dups/deletions from G_1 to G_2 is minimum.

Consistent CNP-2-CNP problem

(1, 2, 2, 2)

(3, 0, 3, 6)

Go from **any** genome with
1 x a
2 x b
2 x c
2 x d

to **any** genome with
3 x a
0 x b
3 x c
6 x d

Consistent CNP-2-CNP problem

(1, 2, 2, 2)

a d d c c b b

(3, 0, 3, 6)

Go from **any** genome with
1 x a
2 x b
2 x c
2 x d

to **any** genome with
3 x a
0 x b
3 x c
6 x d

Consistent CNP-2-CNP problem

(1, 2, 2, 2)

a d d c c ~~b b~~

a d d c c

(1, 0, 2, 2)

(3, 0, 3, 6)

Go from **any** genome with
1 x a
2 x b
2 x c
2 x d

to **any** genome with
3 x a
0 x b
3 x c
6 x d

Consistent CNP-2-CNP problem

(1, 2, 2, 2)

a d d c c ~~b b~~

a d d c c (1, 0, 2, 2)

a d d c a d d c c (2, 0, 3, 4)

(3, 0, 3, 6)

Go from **any** genome with
1 x a
2 x b
2 x c
2 x d

to **any** genome with
3 x a
0 x b
3 x c
6 x d

Consistent CNP-2-CNP problem

(1, 2, 2, 2)

a d d c c ~~b b~~

a d d c c (1, 0, 2, 2)

a d d c a d d c c (2, 0, 3, 4)

a d d a d d c a d d c c

(3, 0, 3, 6)

Go from **any** genome with
1 x a
2 x b
2 x c
2 x d

to **any** genome with
3 x a
0 x b
3 x c
6 x d

Consistent CNP-2-CNP problem

Open problem

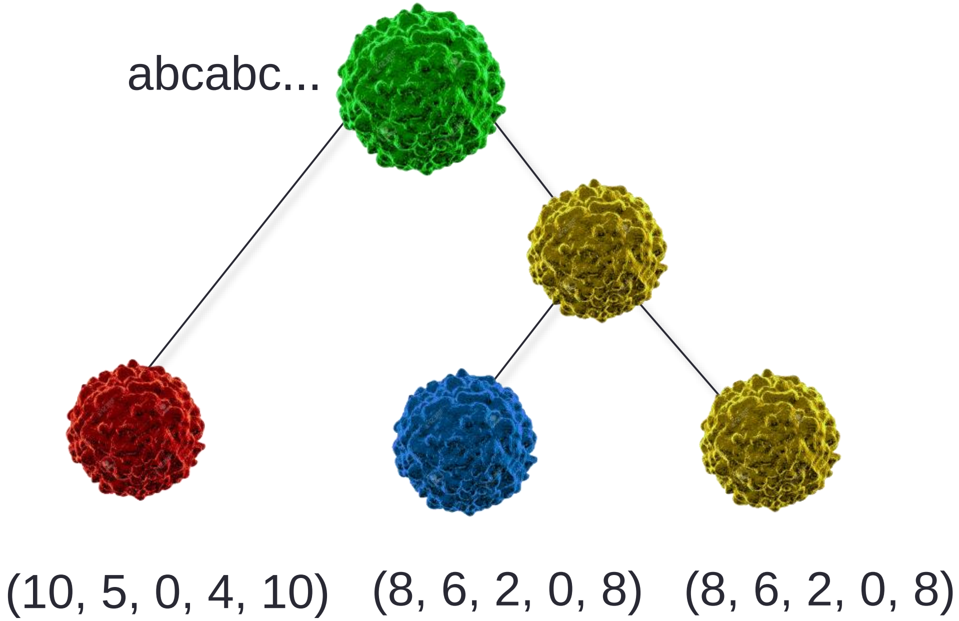
Any question you can think of about this problem!

- Very interesting theoretical problem.
- In practice...
  - Some optimal solution always has 0 or 1 deletion (we think)
  - Gives rise to ridiculous genomes
  - e.g. aaaaaabbbbbbccccdddd

- More useful formulation: global inference of genomes on a phylogeny

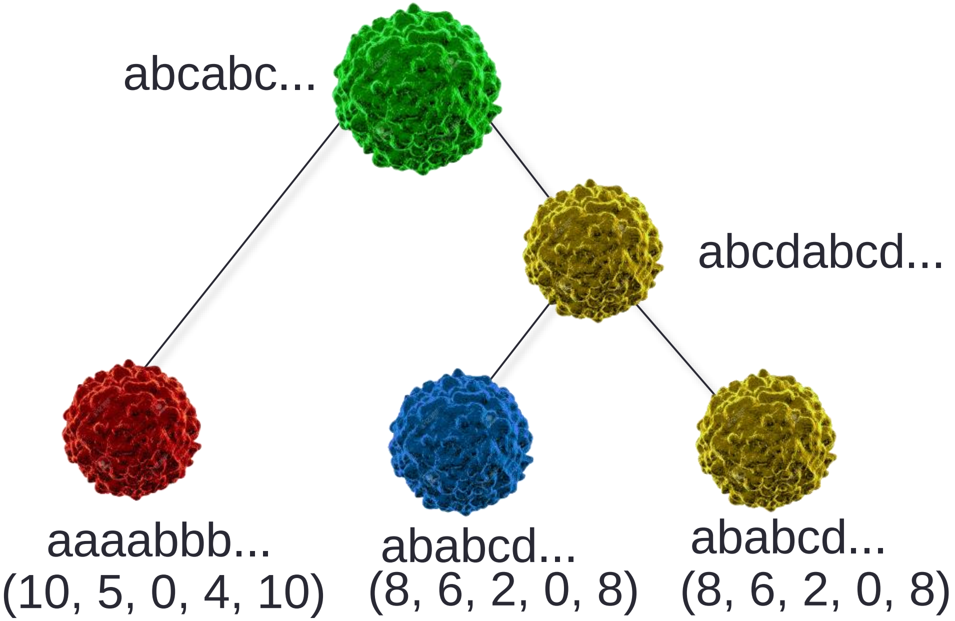
Phylogenetic CNP problem

- **Given**: phylogeny T with CNPs at leaves, human genome at root
- **Find**: a genome assignment at each node of T such that:
  - each genome at a leaf has correct CNP;
  - sum of rearrangements at branches is minimum.



Phylogenetic CNP problem

- **Given**: phylogeny T with CNPs at leaves, human genome at root
- **Find**: a genome assignment at each node of T such that:
  - each genome at a leaf has correct CNP;
  - sum of rearrangements at branches is minimum.



Conclusion

- Copy-number profiles carry useful information on tumor heterogeneity.
- Easier to obtain than whole genomes.

Conclusion

- Copy-number profiles carry useful information on tumor heterogeneity.
- Easier to obtain than whole genomes.

- We are not exploiting this information at its full potential!

Conclusion

- What this line of research needs:
  - Better models
  - Better problem formulations
  - Better algorithms
  - Better access to real data!!!