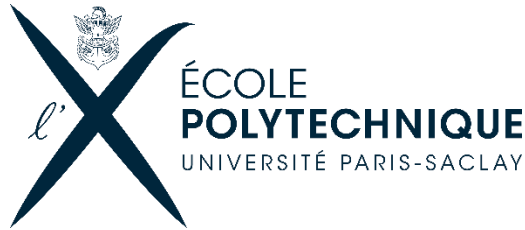


COMPARING COPY-NUMBER PROFILES UNDER MULTI-COPY AMPLIFICATIONS AND DELETIONS

Garance Cordonnier, Manuel Lafond



If you know the number of copies of each segment of interest in several tumor cells, can you reconstruct the cells phylogeny?

If you know the number of copies of each segment of interest in several tumor cells, can you reconstruct the cells phylogeny?

Need to compare copy-number profiles (CNP).

In this talk

- CNP-2-CNP problem
- Turn a copy-number vector into another with min # of moves
- **Allowed move:** take any contiguous interval of the vector, alter values by the same amount. A zero stays a zero forever.

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

(1,3,0,0,3,0,3)

In this talk

- CNP-2-CNP problem
- Turn a copy-number vector into another with min # of moves
- **Allowed move:** take any contiguous interval of the vector, alter values by the same amount. A zero stays a zero forever.

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



-1

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



-1

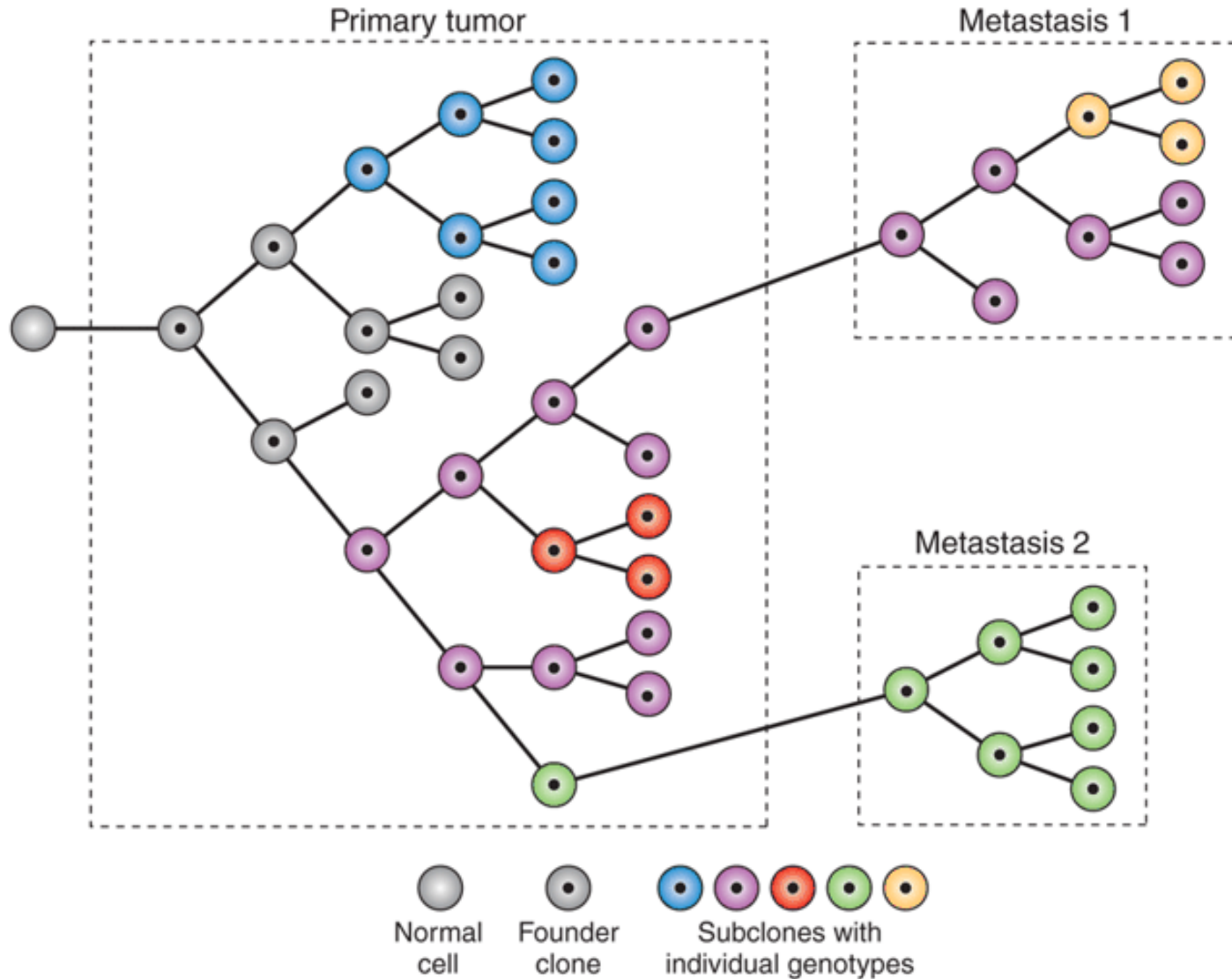
(1,1,0,0,1,0,1)



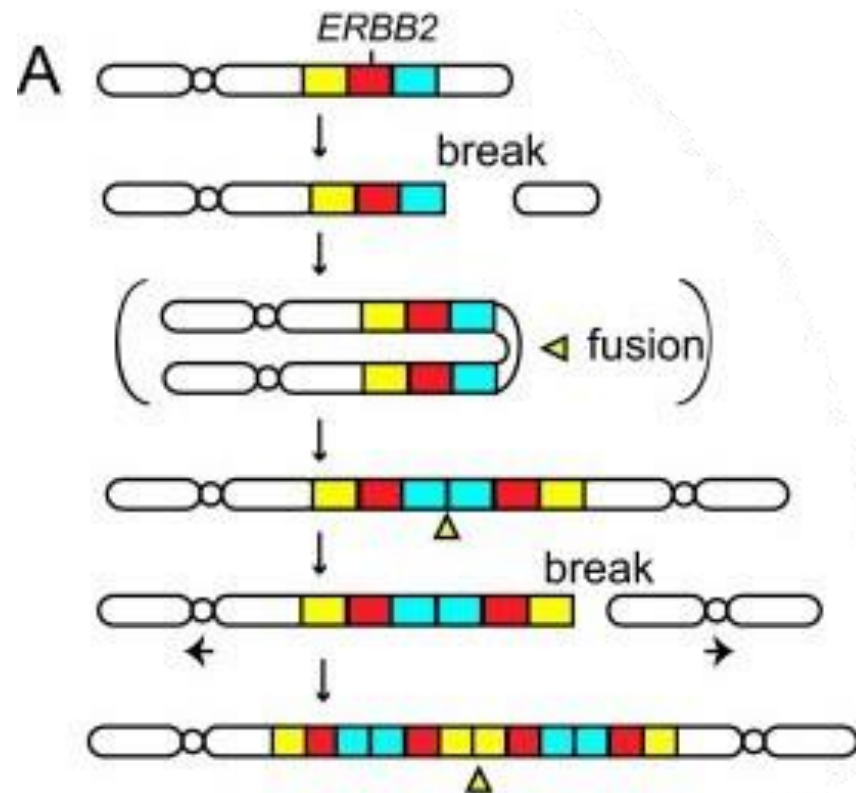
+2

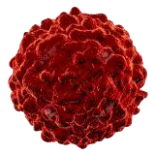
(1,3,0,0,3,0,3)

Tumours and evolution

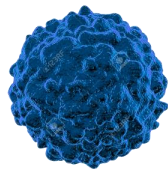


- Tumor cells undergo rapid genomic evolutionary changes
 - Rearrangements
 - Amplifications
 - Deletions
- e.g. : breakage-fusion-bridge cycle

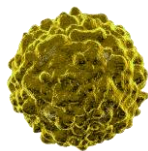




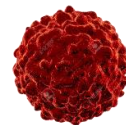
C1



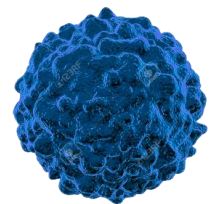
C2



C3

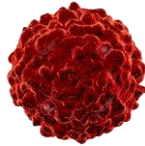


C4



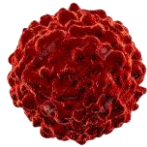
C5

Copy-Number Profile (CNP)

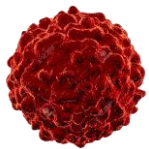


10 copies of segment 1
5 copies of segment 2
0 copies of segment 3
4 copies of segment 4
10 copies of segment 5

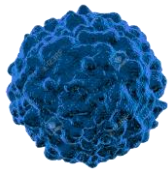
Copy-Number Profile (CNP)



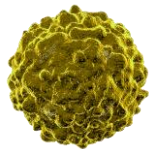
(10, 5, 0, 4, 10)



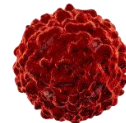
(10, 5, 0, 4, 10)



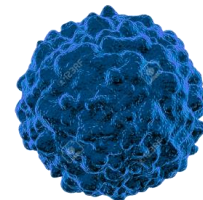
(8, 6, 2, 0, 8)



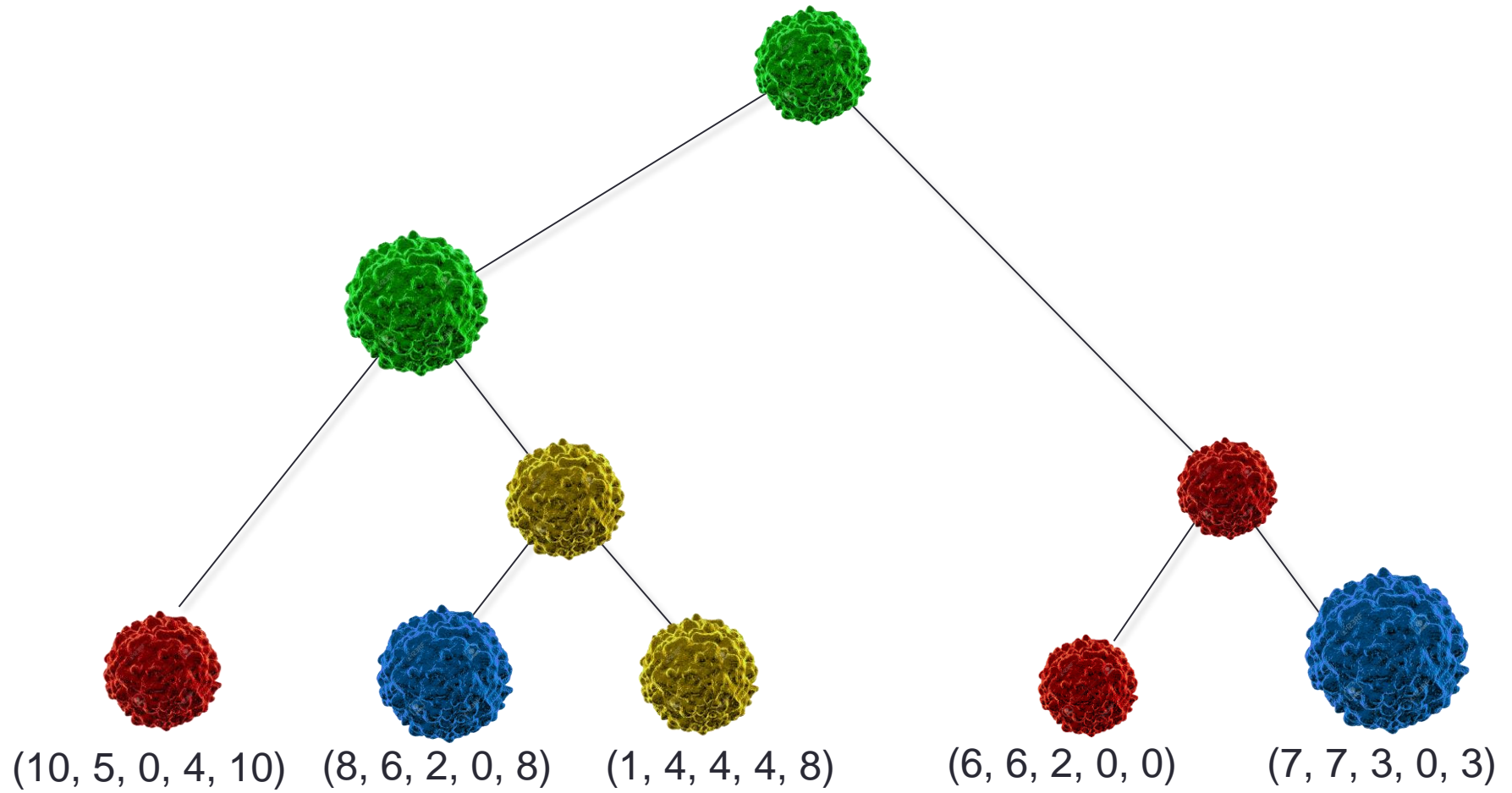
(1, 4, 4, 4, 8)



(6, 6, 2, 0, 0)



(7, 7, 3, 0, 3)



Sequence cells from same tumor (single cell sequencing)



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



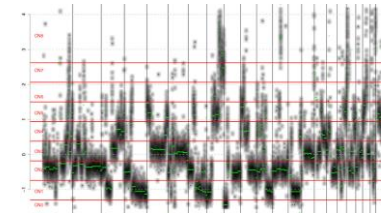
Phasing: assign copy-number to allele, get chromosome Copy-Number Profile (CNP)



Compare CNPs of each pair of cells => distance matrix

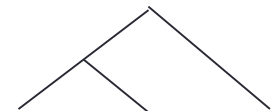


Reconstruct phylogeny



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

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



Sequence cells from same tumor (single cell sequencing)



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



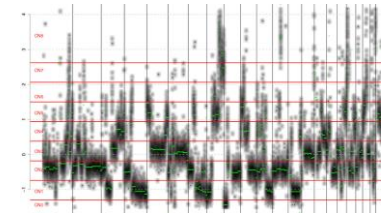
Phasing: assign copy-number to allele, get chromosome Copy-Number Profile (CNP)



Compare CNPs of each pair of cells => distance matrix

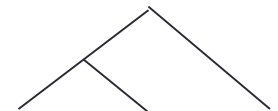


Reconstruct phylogeny



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

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




nature > letters > article

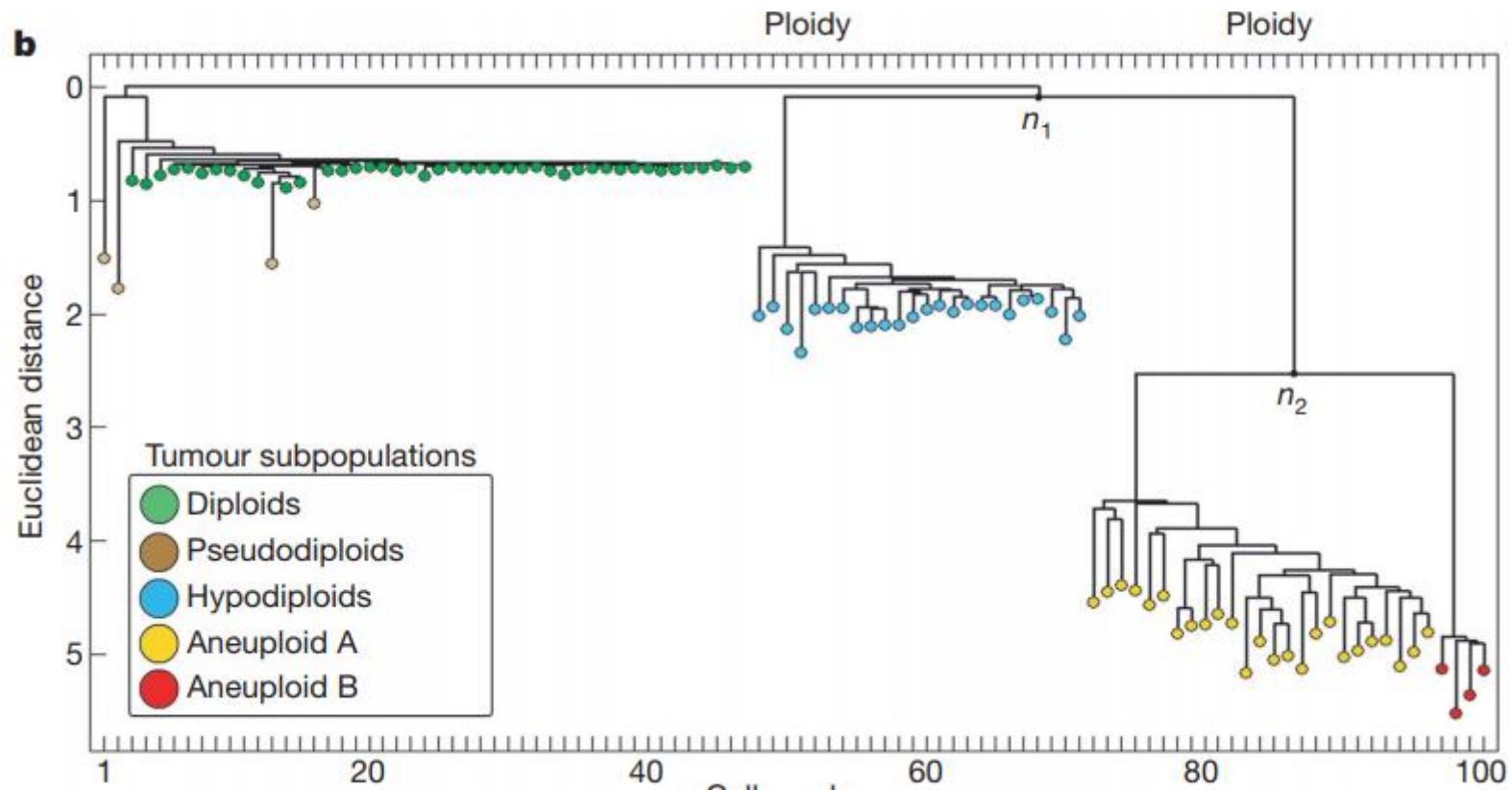
nature
International journal of science

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.



- In Navin et al. [Nature11] : Euclidean distance
- $dist(\mathbf{u}, \mathbf{v}) = \sqrt{\sum (u_i - v_i)^2}$
- Implicit assumption: positions are independent.

In Schwarz et al. [PlosCB14]: MEDICC model

- Allowed move = alter contiguous interval by +1/-1.
- A 0 stays a 0.
- $dist(\mathbf{u}, \mathbf{v}) = \min \#$ of moves to turn \mathbf{u} into \mathbf{v}
 - Exponential-time algorithm $\Omega(3^N)$, $N = \max$ copy-number

(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)

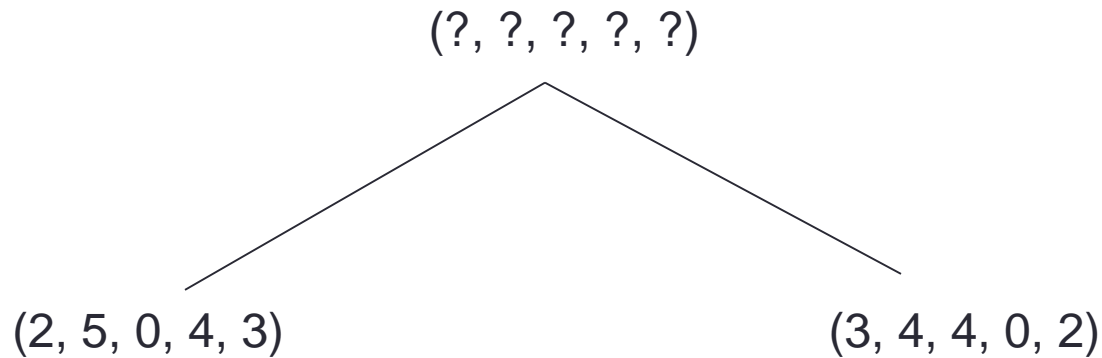


+1

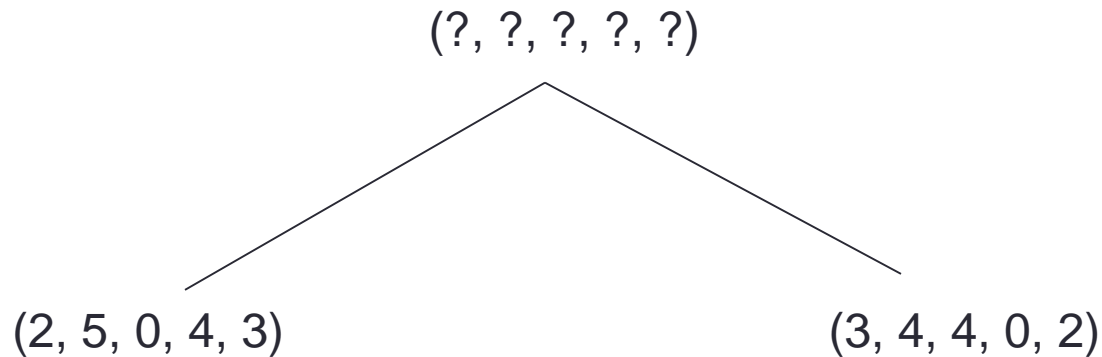
(1,3,0,0,3,0,3)

- 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.

- In El-Kebir et al. [AMB17]:
 - Copy-number triplet: under same +1/-1 distance, given CNPs u and v , find w that minimizes $d(w, u) + d(w, v)$.



- In El-Kebir et al. [AMB17]:
 - Copy-number triplet: under same +1/-1 distance, given CNPs u and v , find w that minimizes $d(w, u) + d(w, v)$.
 - Solvable in pseudo-polynomial time.
 - CNP phylogeny: infer ancestral CNPs to minimize distances on branches.



In this paper

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

In this paper

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

Generalized cost:

We are given a function $f: \mathbb{N} \rightarrow \mathbb{N}$ saying that altering value at u_i by an amount of d costs $f(u_i, d)$.

Altering interval (i, j) by amount of d costs $\max_{i \leq k \leq j} f(u_k, d)$

In this paper

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

Generalized cost:

We are given a function $f: \mathbb{N} \rightarrow \mathbb{N}$ saying that altering value at u_i by an amount of d costs $f(u_i, d)$.

Altering interval (i, j) by amount of d costs $\max_{i \leq k \leq j} f(u_k, d)$

Rationale: a genomic event could amplify/delete two or more copies of the same segment.

MEDICC model:

- $f(i, 1) = f(i, -1) = 1$ for all i
- $f(i, d) = \infty$ for all d other than $1, -1$

"DBL" model

Assume a copy-number cannot more than double in a single event

- $f(i, d) = 1$ if $i + d \leq 2i$
- $f(i, d) = \infty$ otherwise

"ANY" model

Any movement is allowed at unit cost.

- $f(i, d) = 1$ for any d

CNP-2-CNP problem

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

CNP-2-CNP problem ("any" cost)

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

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

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

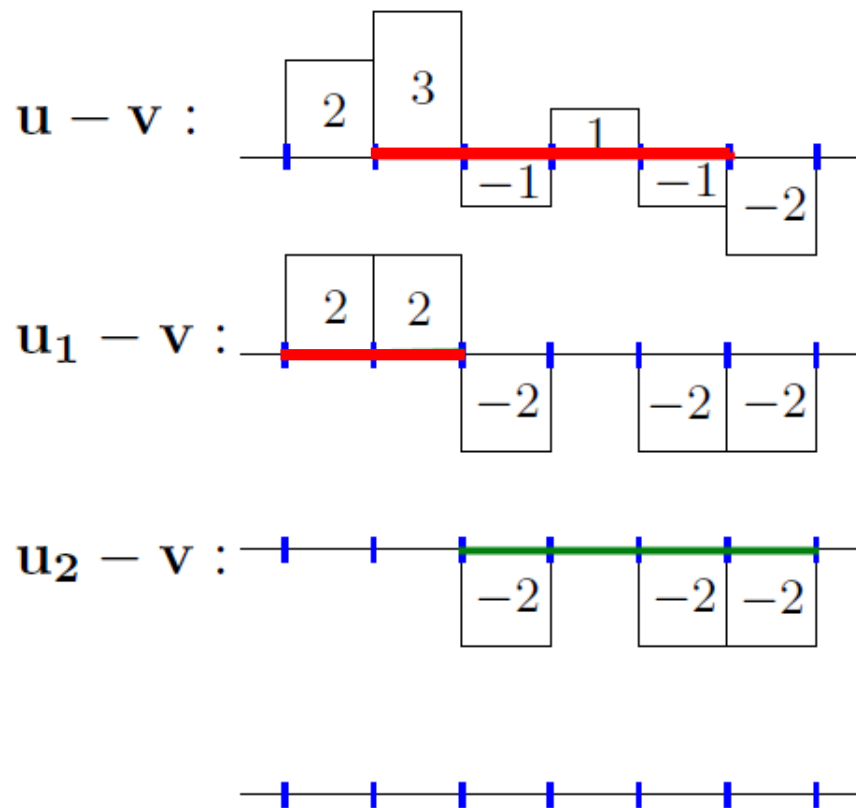
$$u_1 = (\underline{3}, \underline{4}, 2, 0, 3, 2)$$

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

$$u_2 = (1, 2, \underline{2}, \underline{0}, \underline{3}, \underline{2})$$

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

$$v = (1, 2, 4, 0, 5, 4)$$



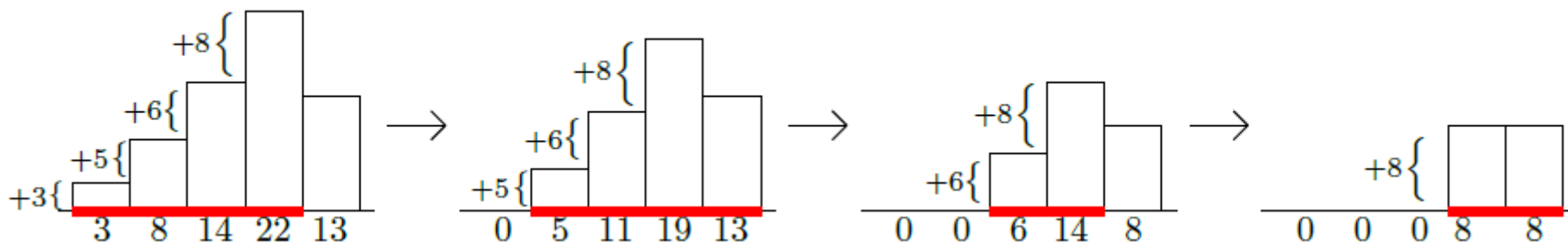
Theorem

For any cost scheme f that allows any number of deletions at unit cost, the CNP-2-CNP problem is **strongly** NP-hard.

(strongly \Rightarrow hard even if the numbers are polynomial in n)

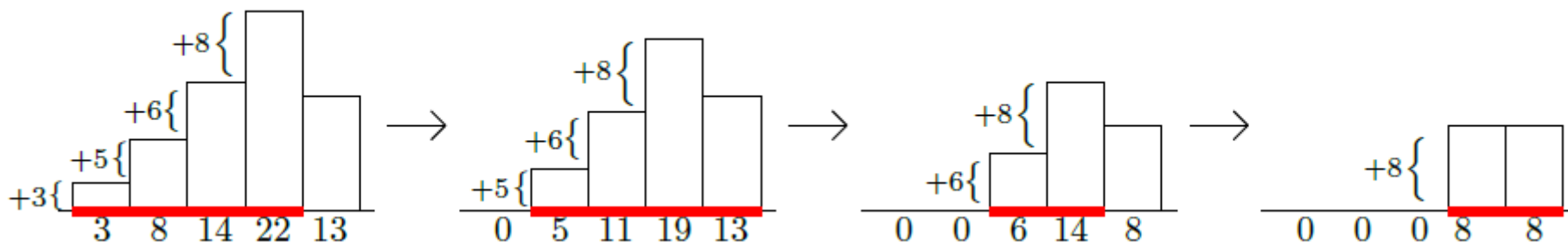
Easy hardness version (weak NP-h)

- Reduction from Subset Sum
- Instance $S = \{3,5,6,8\}$, is there a subset that sums to 13?
- Setup u and v so that difference vector is a staircase, followed by 13. Each step of the stairs = a member of S .



Easy hardness version (weak NP-h)

- Reduction from Subset Sum
- Instance $S = \{3,5,6,8\}$, is there a subset that sums to 13?
- Setup u and v so that difference vector is a staircase, followed by 13. Each step of the stairs = a member of S .
- Goal: squish each step while getting rid of that last 13.
- Strong hardness version uses 3d-matching.



Positive results

Theorem

For the cost scheme "ANY" without 0-positions, there is a linear time factor 2 approximation algorithm for the CNP-2-CNP problem.

The algorithm

Return the number of flat intervals in the difference vector.

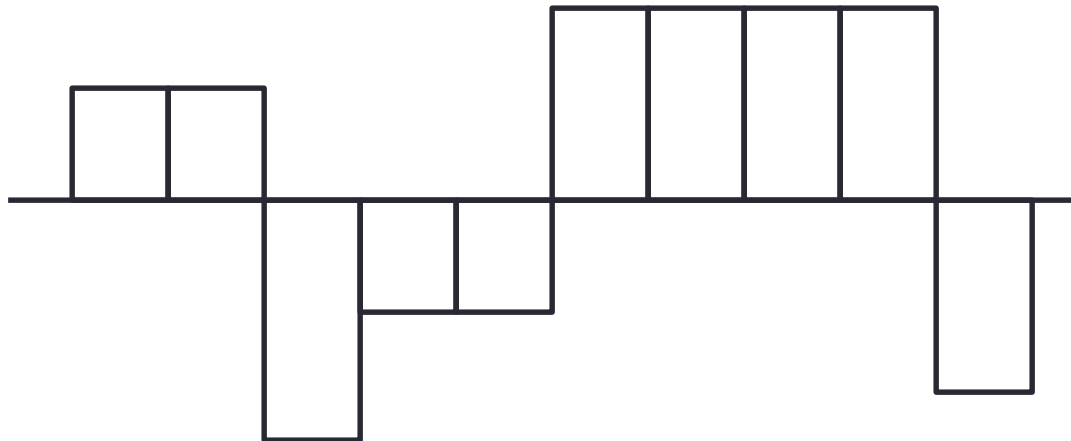
Flat interval = contiguous positions in which difference vector has same value.

Below: 5 flat intervals

Lemma

One move reduces number of flat intervals by at most 2

$\Rightarrow dist_{any}(\mathbf{u}, \mathbf{v})$ is at least $\frac{1}{2}$ the number of flat intervals.



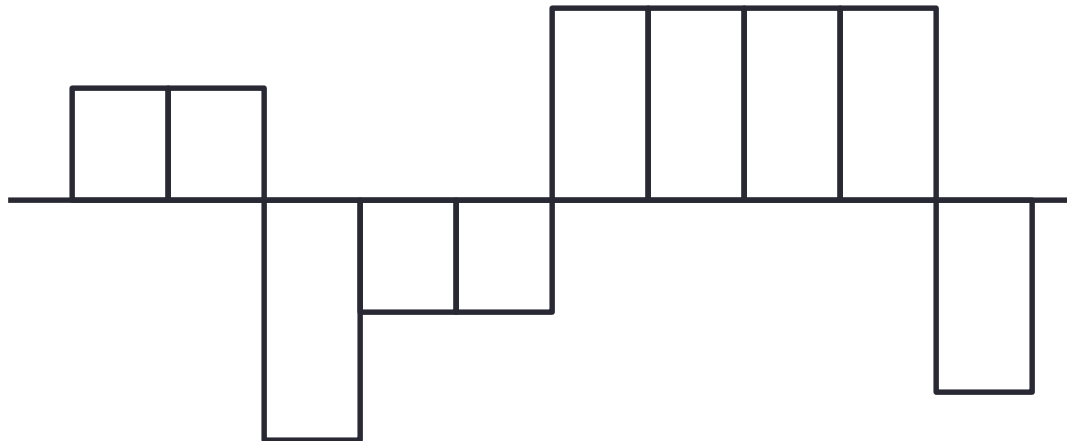
Flat interval = contiguous positions in which difference vector has same value.

Below: 5 flat intervals

Lemma

One move reduces number of flat intervals by at most 2
 $\Rightarrow dist_{any}(\mathbf{u}, \mathbf{v})$ is at least $\frac{1}{2}$ the number of flat intervals.

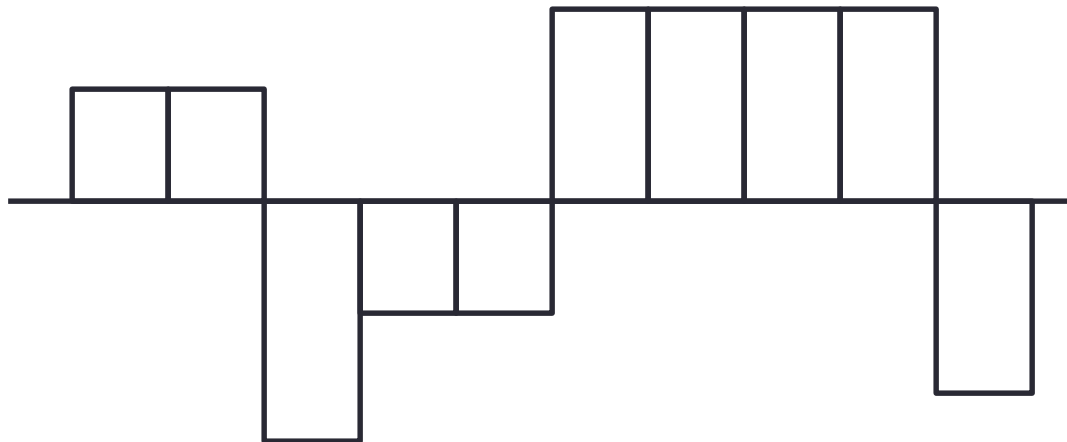
Trivial 2-approx: remove each flat interval one by one.



Improved 2-approx heuristic

If there is a move that reduces the number of flat intervals by 2, then do it. If not, remove any flat interval.

Can only do better than trivial 2-approx.



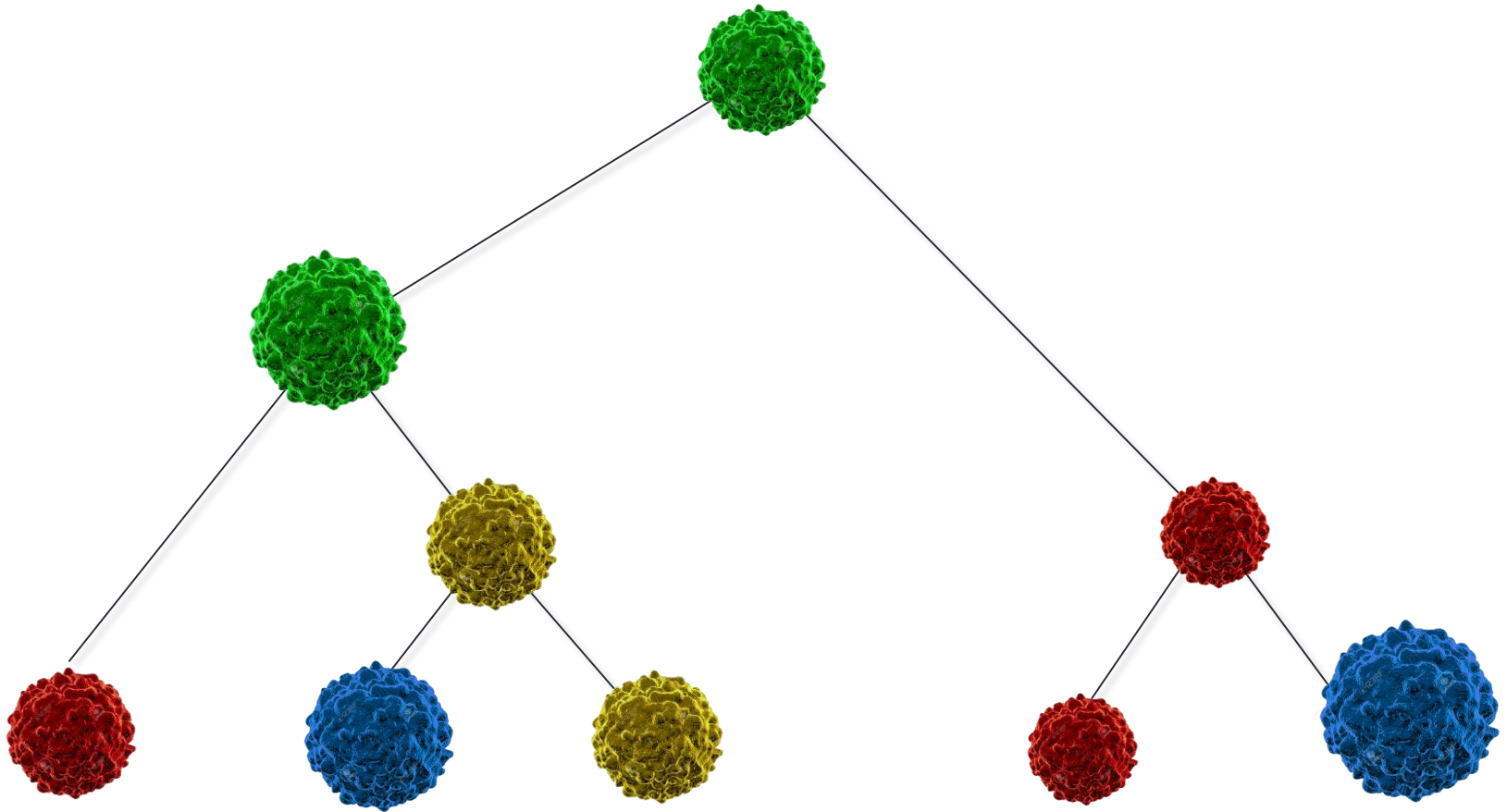
Experiments

- If we only model amplifications and deletions on genomes, can we at least reconstruct phylogenies?

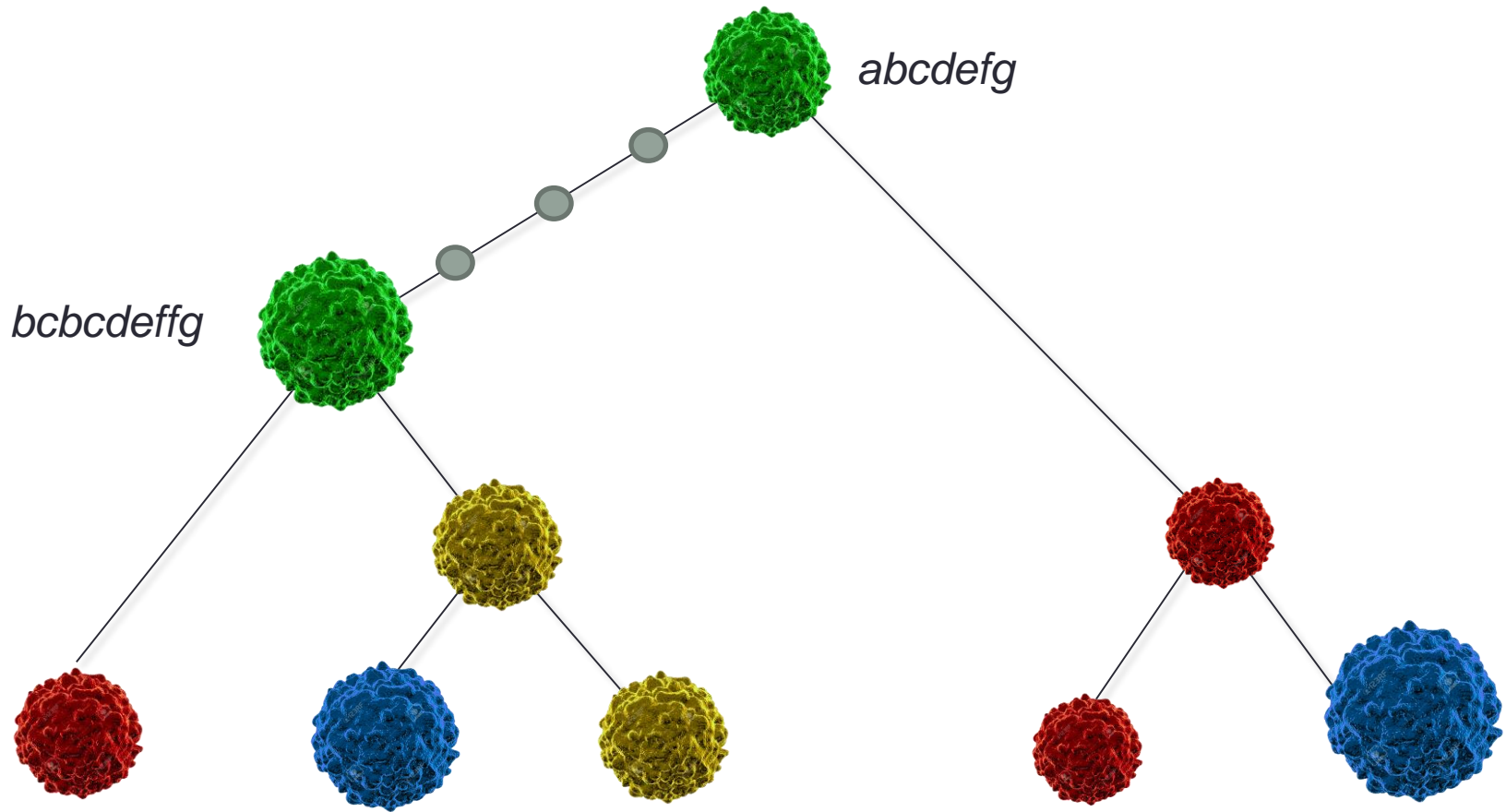
Experiments

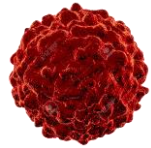
- Simulated trees with random events on each branch
 - Tree topology chosen uniformly at random (non-ultrametric)
 - Simulated chromosomes (not simulated CNPs)
 - Duplications and deletions events
 - ~4000 trees, various parameter combinations
- Goal: given leaf CNPs, reconstruct phylogeny
 - Comparison of (1) flat interval count, (2) improved heuristic, (3) ZZS algorithm/MEDICC model, (4) Euclidean distance
 - Measure: normalized RF distance
(# clades in reconstructed tree that are not in true tree)/(# clades)

n = number of leaves in random tree (def. $n = 100$)
 l = number of genes per chromosome (def. $l = 100$)
 ∂ = prob. of duplication (def. $\partial = 0.5$)
 (e_1, e_2) = range of # of events per branch (def. $[5..10]$)
 (p, q) = control length of events

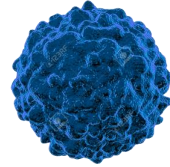


n = number of leaves in random tree (def. $n = 100$)
 l = number of genes per chromosome (def. $l = 100$)
 ∂ = prob. of duplication (def. $\partial = 0.5$)
 (e_1, e_2) = range of # of events per branch (def. $[5..10]$)
 (p, q) = control length of events

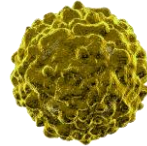




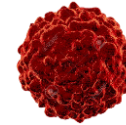
G1 =>
CNP1



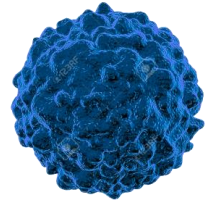
G2 =>
CNP2



G3 =>
CNP3



G4 =>
CNP4



G5 =>
CNP5



Flat interval
count

Improved
approx

ZZS algo
(+1/-1)

Euclidean
distances

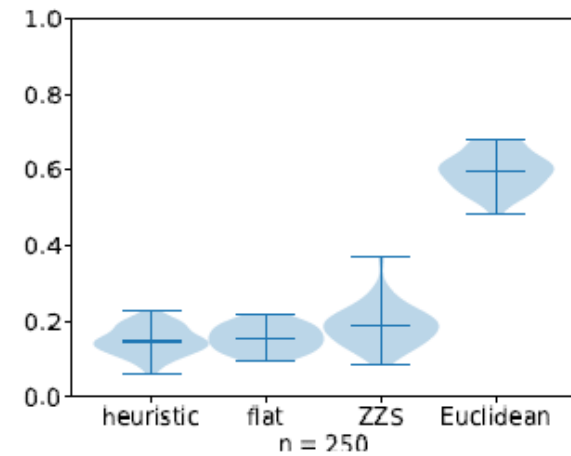
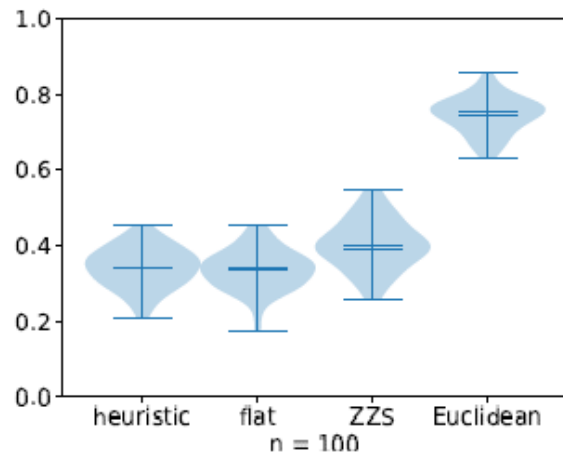
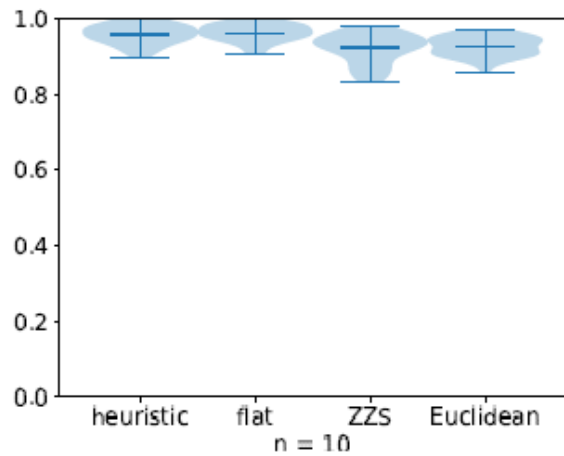
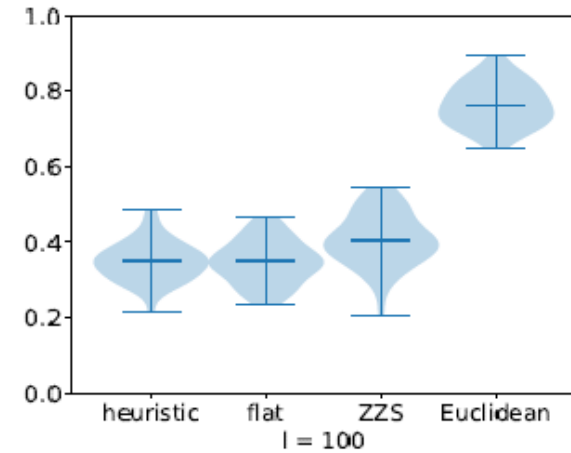
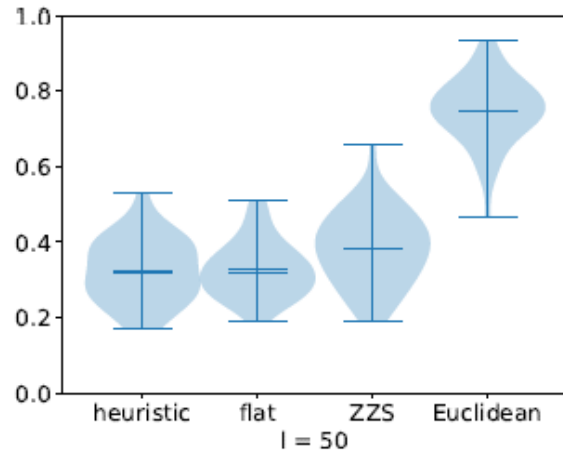
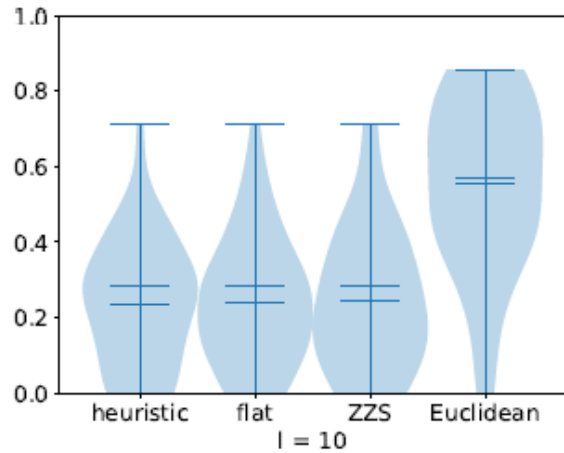


Distance
matrix



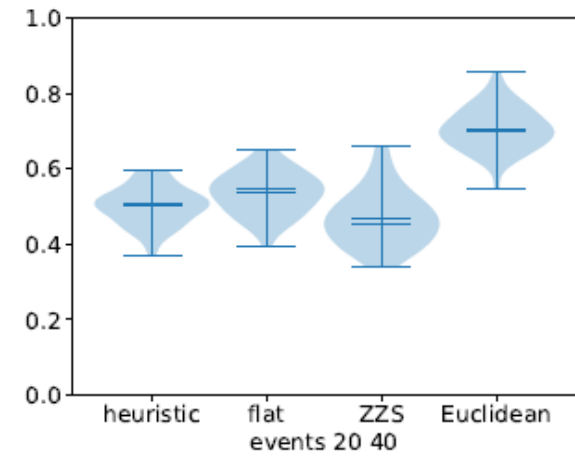
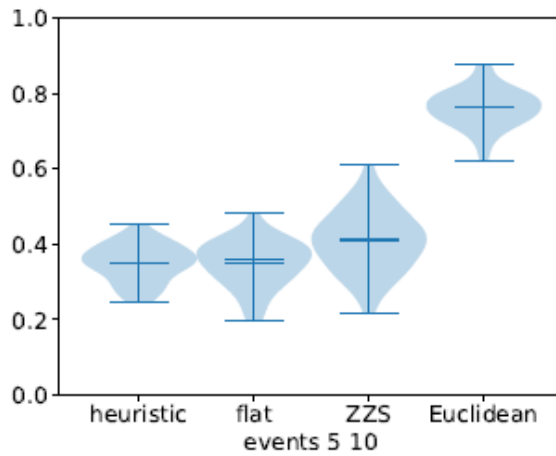
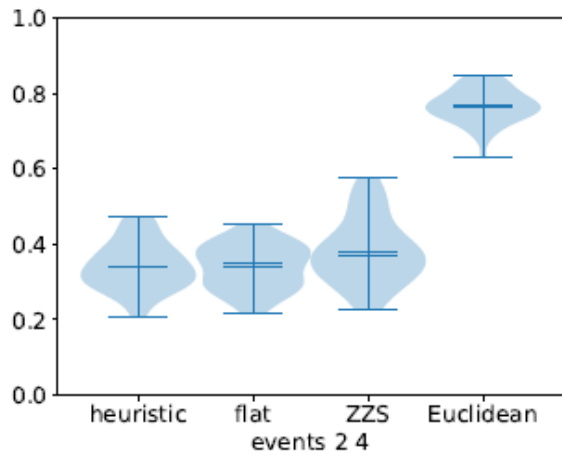
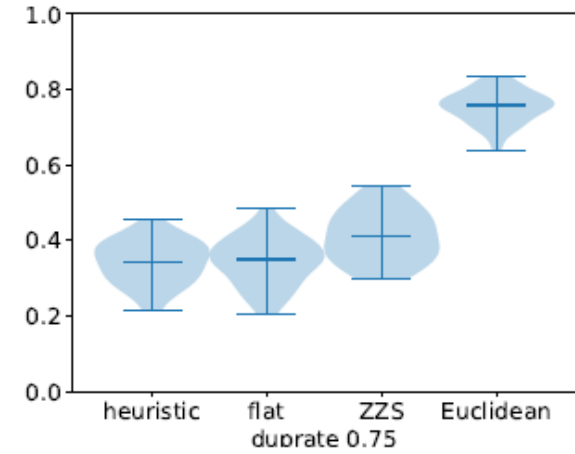
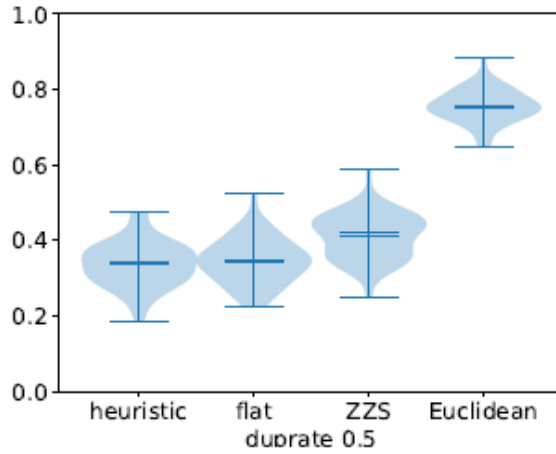
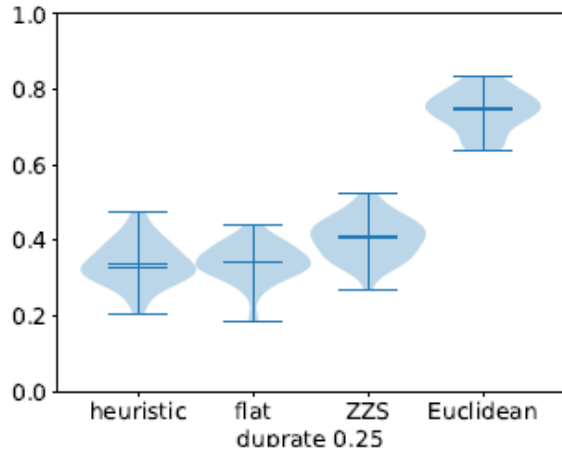
Neighbor-
Joining

More leaves = easier to predict



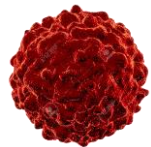
More genes = easier to predict

Dup prob. = no change

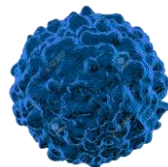


More events = harder to predict

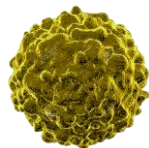
Introducing errors



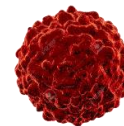
G1 =>
CNP1



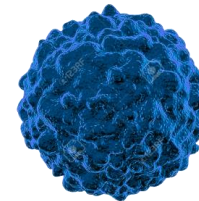
G2 =>
CNP2



G3 =>
CNP3



G4 =>
CNP4



G5 =>
CNP5



Alter each CNP value
 $u_i \Rightarrow \text{choose from } \mathcal{N}(u_i, \alpha)$
 $\alpha \text{ in } \{0.1, 0.25, 0.5, 1\}$



Flat interval
count

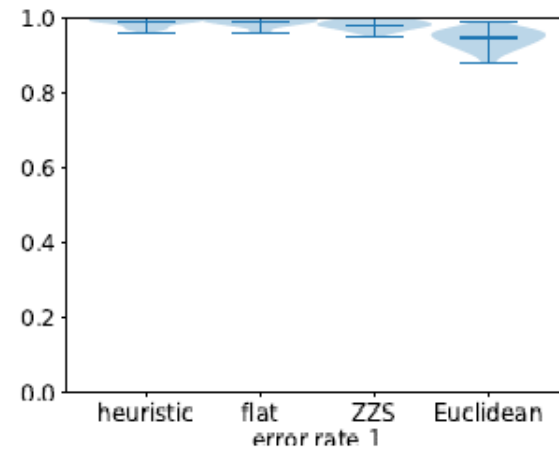
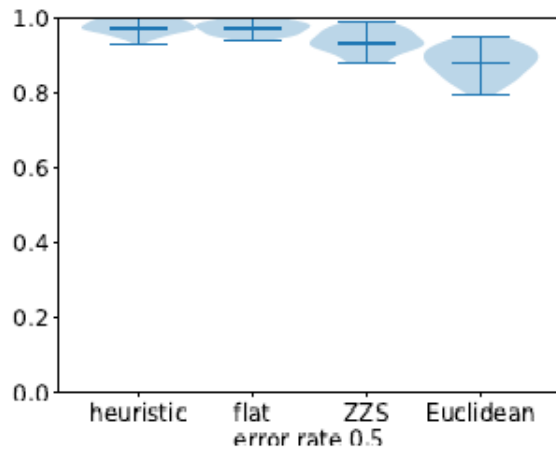
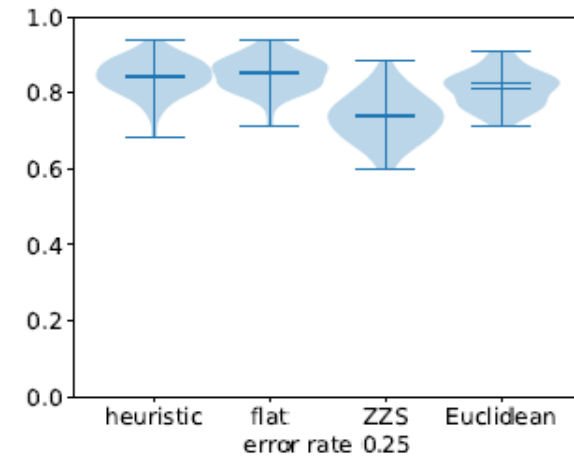
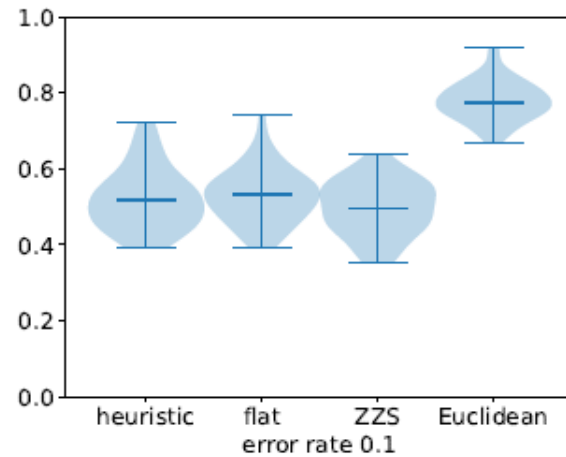
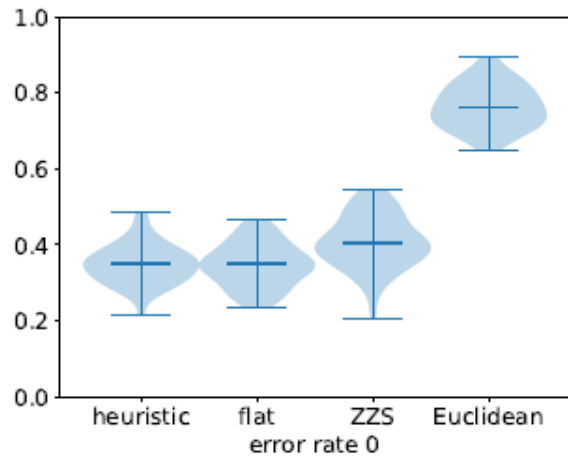
Improved
approx

ZZS algo
(+1/-1)

Euclidean
distances



ZZS a bit more tolerant to error, but accuracy decays rapidly



Discussion

- Max parsimony / statistical reconstructions from CNPs
- How to handle errors in the copy-number calls?
 - Especially for high copy numbers
- Open problems
 - Approx of "DBL" function? Approx with 0-positions? Which cost schemes are poly-time?
- Is the model of "contiguous interval movement" appropriate?
 - Breaks down when heavy rearrangements occur
 - Does not handle amplicons formed by multi-chromosomal genes
 - Mix with SNV/breakpoint/adjacency info?