Algorithms for Pan-genome Computational Analysis

Founder set construction under allelic and non-allelic homologous recombination

Konstantinn Bonnet, Daniel Dörr, Tobias Marschall
Institute for Medical Biometry and Bioinformatics
Düsseldorf, Germany

Data Structures in Bioinformatics, July 14, 2022

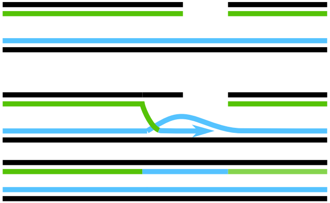
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 956229.

Outline

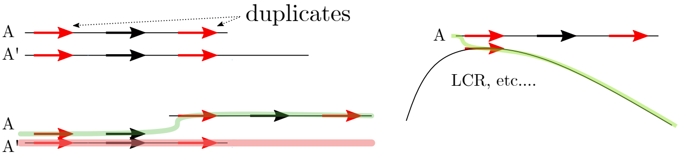
- Homologous recombinations
- A recombination model
- The Founder Set problem
- Minimizing recombinations in founder sets
- Results

EU's Horizon 2020 / ITN MCSA GA No 956229

Homologous recombination



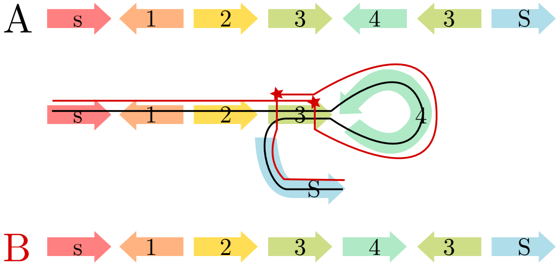
Allelic HR



Non-allelic HR

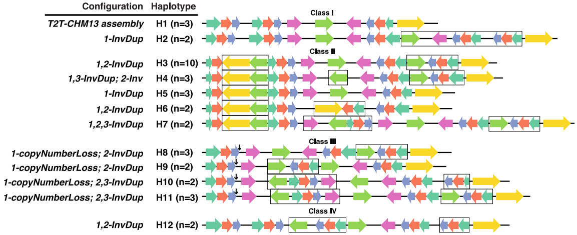
- Double strand-breaks during DNA replication
- Repair: highly similar segments picked by mistake
- Leads to duplications, deletions, and other complex rearrangements

Modeling inversions



- Can be modeled with HR
- Using flanking inverted repeats

Rearrangements in 1p36.13



Configuration	Haplotype
T2T-CHM13 assembly	H1 (n=3)
1-InvDup	H2 (n=2)
1,2-InvDup	H3 (n=10)
1,3-InvDup; 2-Inv	H4 (n=3)
1-InvDup	H5 (n=3)
1,2-InvDup	H6 (n=2)
1,2,3-InvDup	H7 (n=2)
1-copyNumberLoss; 2-InvDup	H8 (n=3)
1-copyNumberLoss; 2-InvDup	H9 (n=2)
1-copyNumberLoss; 2,3-InvDup	H10 (n=2)
1-copyNumberLoss; 2,3-InvDup	H11 (n=3)
1,2-InvDup	H12 (n=2)

- AHR and NAHR play a major role in genomic rearrangements
- Important targets for study in complex loci

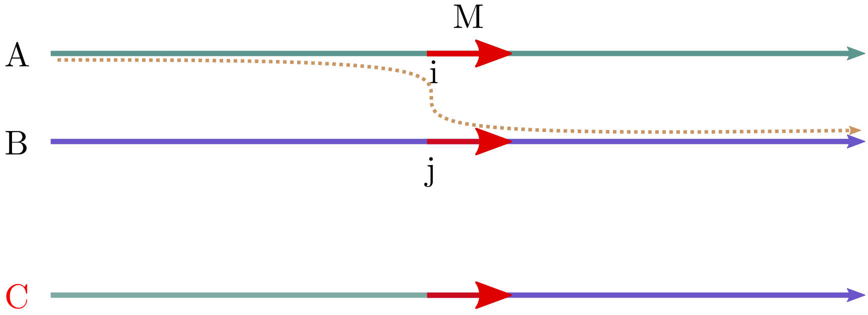
→ New model handling both AHR and NAHR, and which can represent complex rearrangements

[2] Porubsky, Höps, Ashraf et al. Haplotype-resolved inversion landscape reveals hotspots of mutational recurrence associated with genomic disorders Cell, 2022

Outline

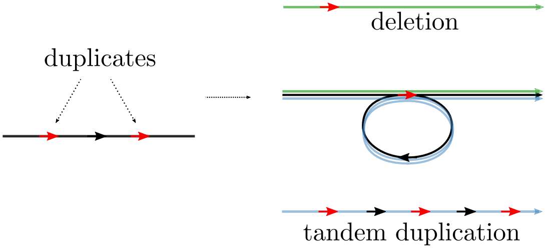
- Homologous recombinations
- A recombination model
- The Founder Set problem
- Minimizing recombinations in founder sets
- Results

The homologous recombination operation



Acts between **two haplotypes** on a **shared segment**
→ Result: **concatenation** of the **suffix** of one and **prefix** of the other

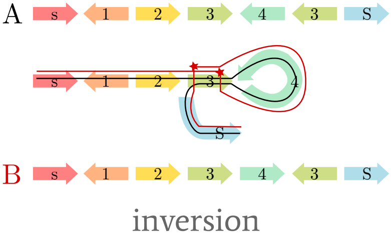
Modeling rearrangements



duplicates

deletion

tandem duplication



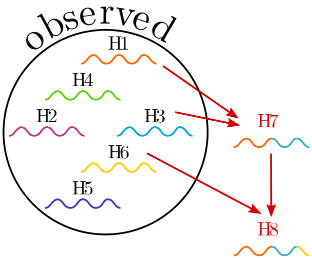
inversion

Outline

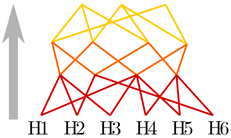
- Homologous recombinations
- A recombination model
- The Founder Set problem
- Minimizing recombinations in founder sets
- Results

EU's Horizon 2020 / ITN MCSA GA No 956229

Recombining in a sample



Recombining pairs
to construct **new haplotypes**

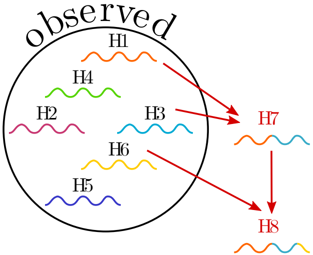


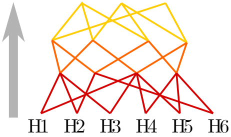
Recombining pairs
to construct **observed haplotypes**

EU's Horizon 2020 / ITN MCSA GA No 956229

Recombining in a sample

Founder Set problem: find a **generator of minimal size** of the set of haplotypes





Recombining pairs to construct **new haplotypes**

Recombining pairs to construct **observed haplotypes**

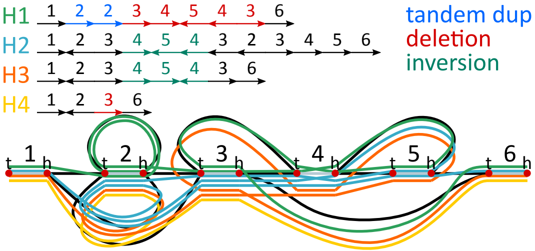
Full algorithmic details: paper

Constructing founder sets under allelic and non-allelic homologous recombination

https://doi.org/10.1101/2022.05.27.493721

Variation graphs

Input: set of haplotypes + their variation graph

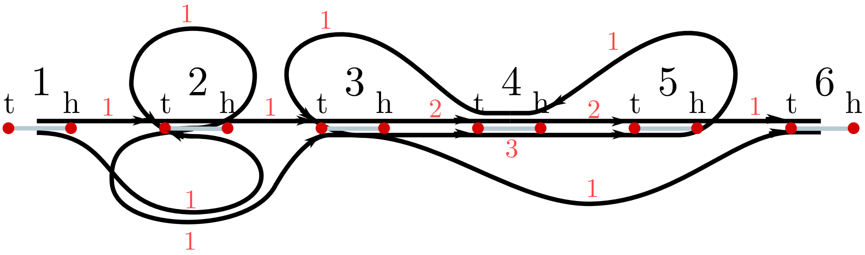


- Nodes and edges: homologous DNA segments and their adjacencies
- Haplotypes: walks from one extremity to the other (source, sink)
- Every edge of the graph is covered by at least one haplotype (by constrution)

A network flow solution

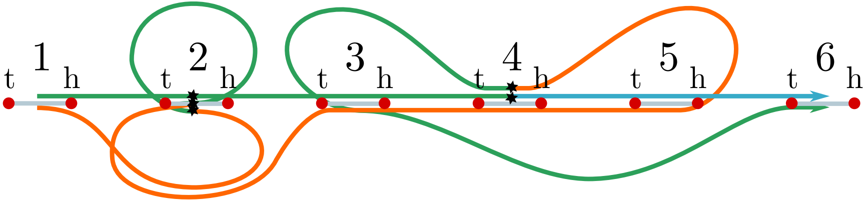
→ Formulation as a network flow problem:
Minimize total flow + constraints to ensure valid haplotypes

Total (incoming) flow at (the tail of) the sink = minimal founder set size



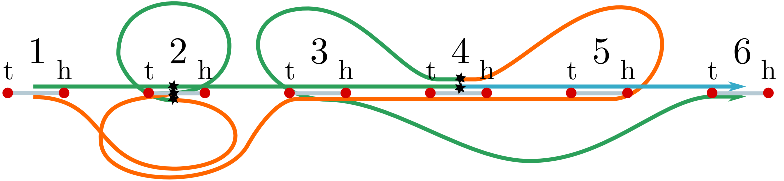
Solution to the Founder Set problem

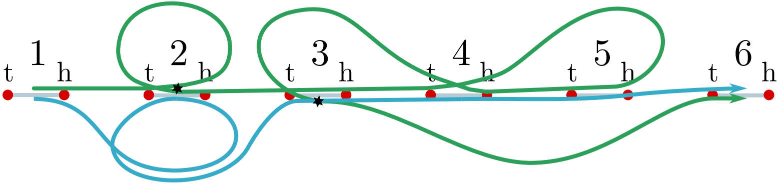
Once founder sequences are constructed → minimal set of founders



Solution to the Founder Set problem

But... is it good enough?



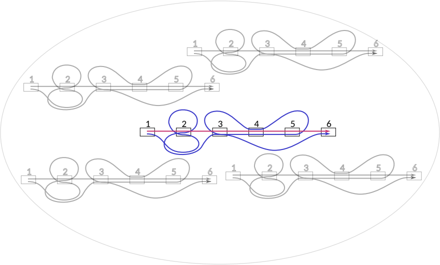


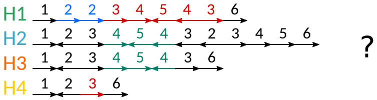
Outline

- Homologous recombinations
- A recombination model
- The Founder Set problem
- Minimizing recombinations in founder sets
- Results

EU's Horizon 2020 / ITN MCSA GA No 956229

One does not simply construct founder sequences...





Structure, number of recombinations?

Haplotype information?

Next question: what is the most parsimonious founder set?

→ Given a flow solution:

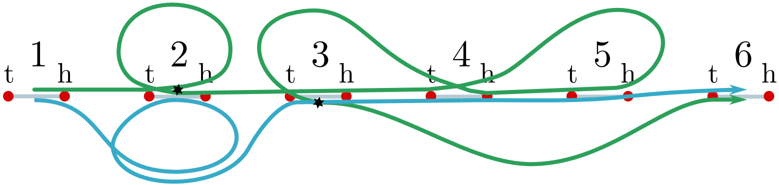
Derive a founder set that **minimizes the recombinations** wrt. the set of haplotypes.

Proposed solution: ILP

Define **minimization problem** on a given flow solution

- Add nodes to the flow graph for marker multiplicities
- Constraints for validity + recombination detection
- Once a solution is computed, extract founder sequences

Solution: **haplotype coloring** and a **minimal number of recombinations**

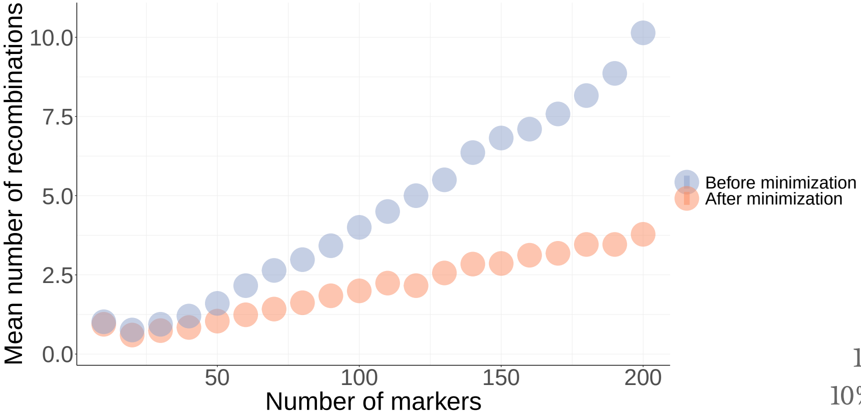


Outline

- Homologous recombinations
- A recombination model
- The Founder Set problem
- Minimizing recombinations in founder sets
- Results

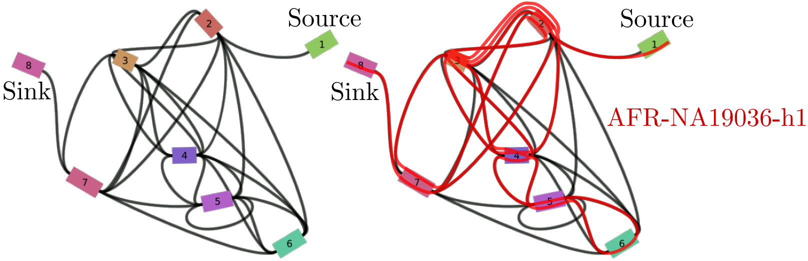
Number of recombinations

Experiments: simulated arbitrary graph (parameterizable) + haplotypes



10 haplotypes
10% dup. and inv.

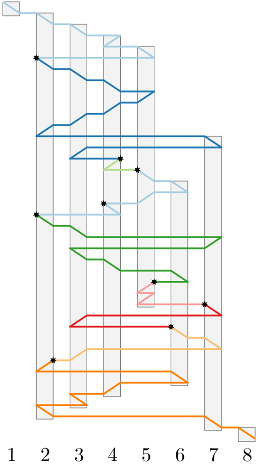
Application: 1p36.13



Visualization with Bandage:
8 nodes, 26 edges, 68 haplotypes + CHM13 reference

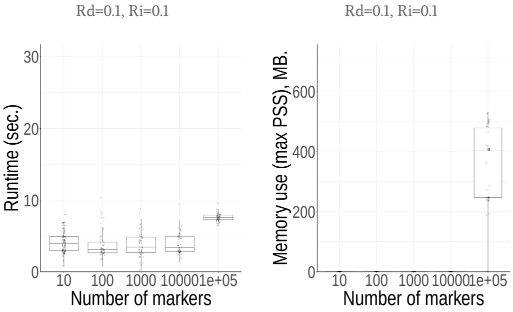
Data provided by Feyza Yilmaz (Jackson Labs).

Application: 1p36.13

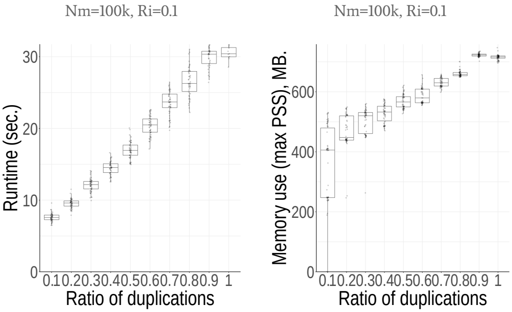


- Single founder sequence
- 9 recombinations between 8 haplotypes
- Minimization time: 60.3 seconds, peak PSS of 225MB.

Benchmark: flow solution

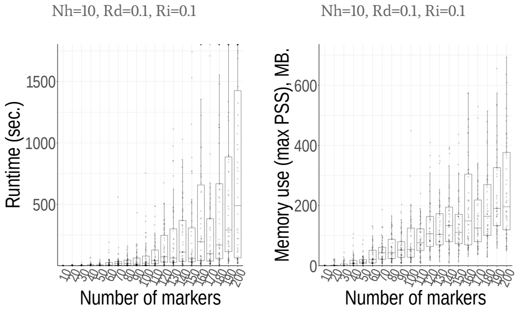


Number of markers

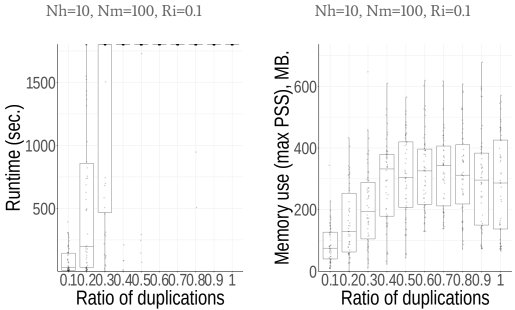


Ratio of duplications

Benchmark: minimization



Number of markers



Ratio of duplications

Conclusion and outlook

A new framework to solve a biologically relevant problem

- A solution to the founder set problem, and a decent one for minimizing the number of recombinations
- Work in progress, but promising with real biological data

Next step:

Find a founder set st. the number of recombinations **within the haplotypes** is minimal wrt. the founder sequences

Availability: `https://github.com/marschall-lab/hrfs`
written in Rust, experiments available as snakemake workflows.

EU's Horizon 2020 / ITN MCSA GA No 956229

Thank you!

Bibliography

[1] R. R. Wick et al. ``Bandage: interactive visualization of de novo genome assemblies''. In: Bioinformatics 31.20 (June 2015), pp. 3350–3352. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv383. eprint: https://academic.oup.com/bioinformatics/article-pdf/31/20/3350/17088082/btv383.pdf. URL: https://doi.org/10.1093/bioinformatics/btv383.

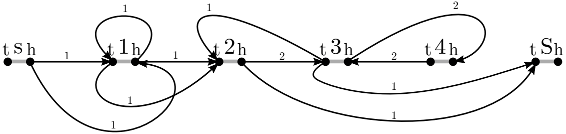
[2] D. Porubsky et al. ``Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders''. In: Cell 185.11 (2022), 1986–2005.e26. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2022.04.017. URL: https://www.sciencedirect.com/science/article/pii/S0092867422004640.

A network flow solution

Note: condense both algo slides into one
Part 1: Given a haplotype variation graph, what is the size of the smallest set of founder sequences? → Formulation as a network flow problem:

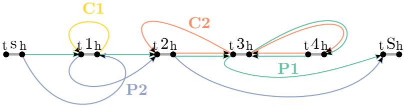
- Total flow of each edge ≥ 1
- Prohibit backtracking to the source ($f_{out_{s^t}} = 0$)

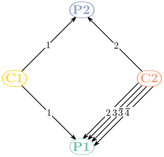
Integer flow at the sink: size of a minimal founder set.

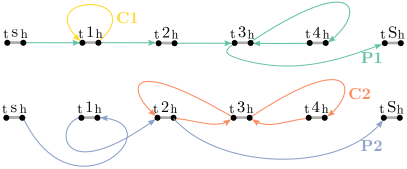


Founder sequences construction

Part 2: Given a solution to the network flow problem, construct a founder set.
→ Idea: decomposition into a component graph







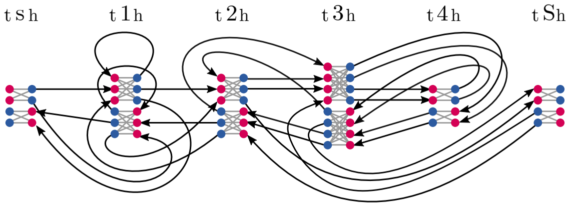
Integrate cycles into paths to obtain a smallest set of founder sequences.

Proposed solution: ILP

Using a previously obtained flow solution:

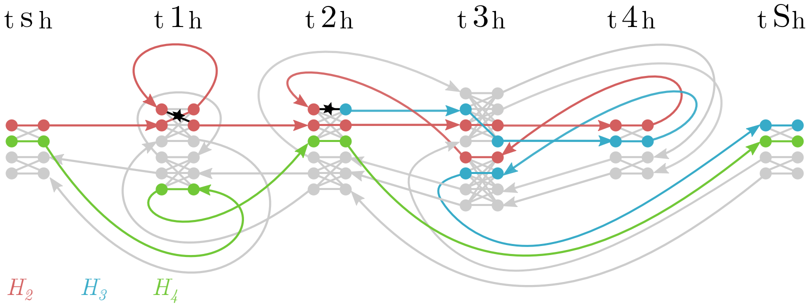
- matching constraints: recombination detected → force a boolean switch on
- flow constraints: push flow from source to sink

Minimize the number of toggled switches, integrating all cycles into valid paths.



ILP solution

Obtained directly from the matchings:


t s h
t 1 h
t 2 h
t 3 h
t 4 h
t S h
H_2
H_3
H_4