Adjacency-constrained hierarchical clustering of a band similarity matrix

with application to GWAS and Hi-C

G. Rigaill – INRA, IPS2 & LaMME

joint with:
C. Ambroise , A. Dehman (Evry),
P. Neuvial, N. Randriamihamison and N. Vialaneix (Toulouse)

AG-MIA

Outline

Motivation

- Genome-Wide Association Studies
- Hi-C studies

Adjacency-constrained hierarchical clustering

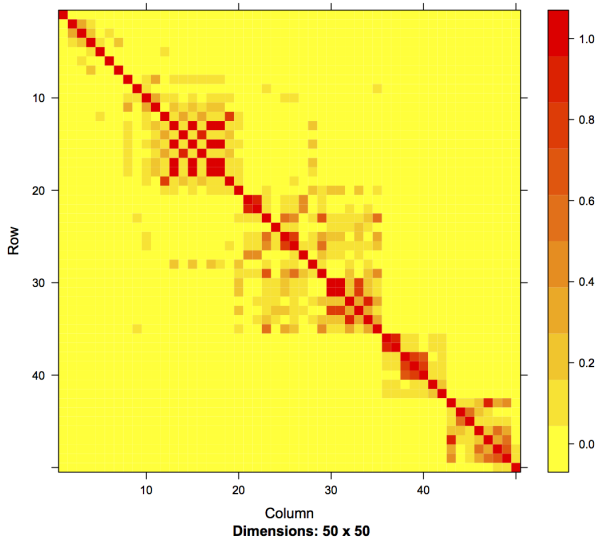
- an improved algorithm
- application to GWAS and Hi-C

Motivation: GWAS and Hi-C

Genome-Wide Association Studies

Goal

Find genetic markers (SNP) significantly associated with a phenotype of interest.



Challenge

Neighboring SNPs are expected to have correlated behavior due to **linkage disequilibrium**

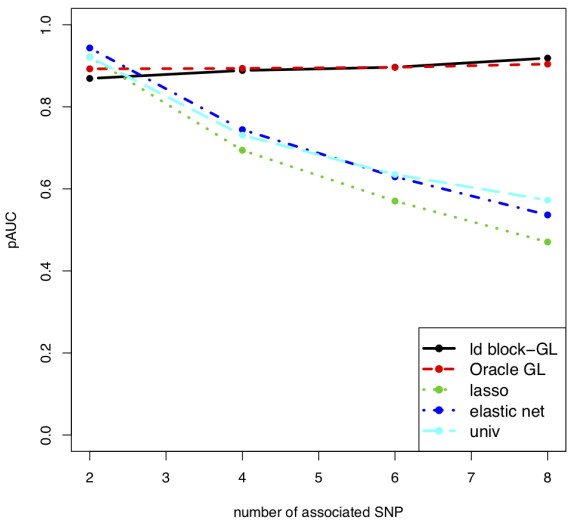
LD map (r^2) for the 50 first SNP of Chromosome 22 in a HIV study (Dalmasso *et al.* 2008):

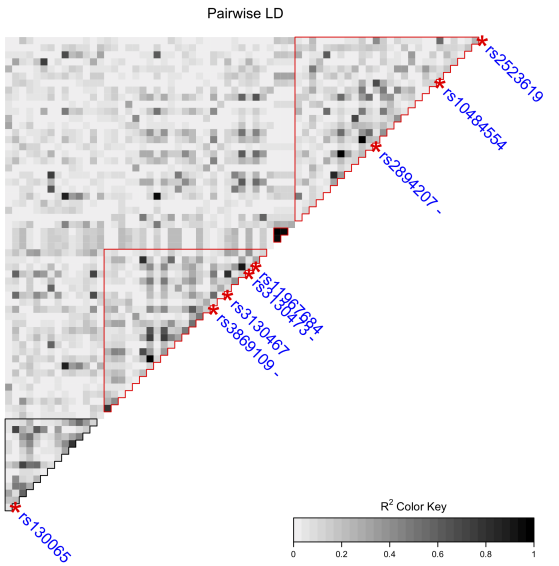
- LD **blocks**
- **hierarchical** structure

Identifying LD blocks associated with a phenotype[1]

A 3-step approach

1. **Hierarchical clustering of the SNPs with adjacency constraint**
2. Estimation of the optimal number of groups using the Gap statistic
3. Selection of the associated blocks using Group Lasso regression





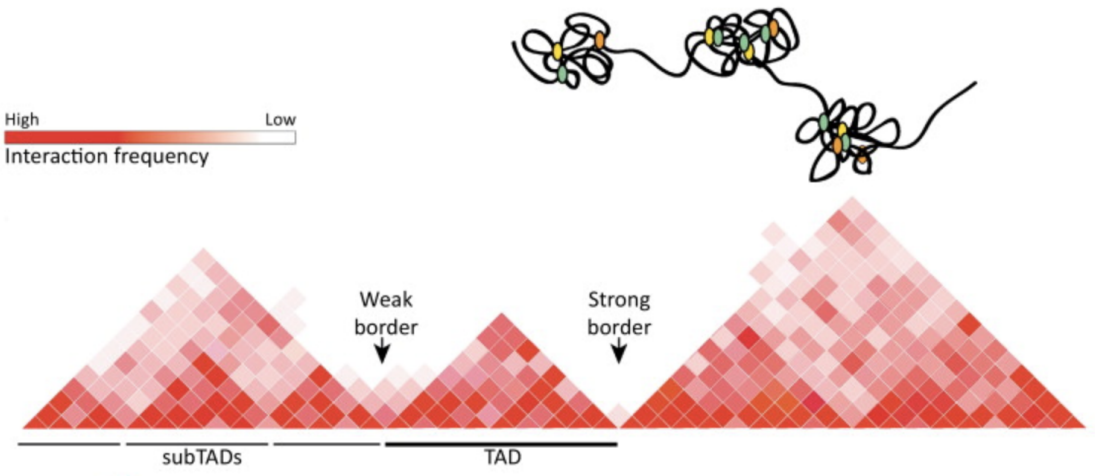
Dehman, Ambroise, & Neuvial, P. (2015). *BMC bioinformatics*, 16(1), 148

TAD detection in Hi-C studies

Goal: identify Topologically Associating Domains (TAD)

TAD = regions whose DNA sequences preferentially contact each other

> Functional relevance of TADs
>
> - in development, cell differentiation
> - in genetic diseases and cancers (modifications of TAD boundaries)



Data: Chromosome contact maps

What is a "good" TAD detection method?

- biological relevance of the TADs
  - size, possible nestedness
  - enrichment in CTCF and cohesin at TAD boundaries
  - enrichment in histone marks involved in transcriptional activity
- robustness to noise
- few, interpretable parameters
- low time/memory footprint
- easy to use implementation

--

Review by Zufferey et al.

Proposal: adjacency-constrained HAC

(HAC = hierarchical agglomerative clustering)

Features

- a generic method to segment the genome into homogeneous regions
- handles similarity data (LD, Hi-C maps)
- adapted to the multi-scale/hierarchical nature of genomic data
- proper statistical interpretation (inertia)

Contributions

- study of mathematical assumptions for the method to be valid
- fast algorithm incorporating a natural biological constraint
- relevance to GWAS and Hi-C studies

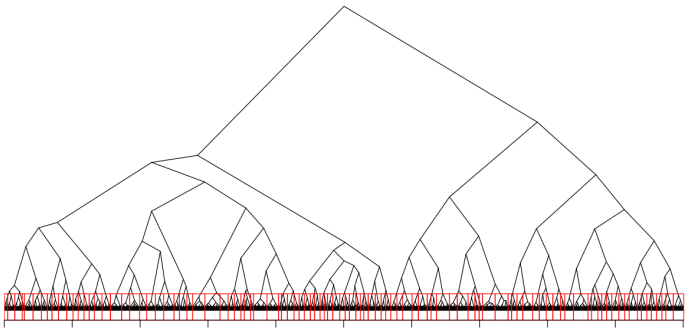
NB: IC-finder uses adjacency-constrained HAC for TAD detection

Adjacency-contrained hierarchical clustering

Hierarchical aggomerative clustering (HAC)

Algorithm

- Input: p objects, similarity S
- Repeat $p - 1$ times: merge the most similar clusters
- Output: A *dendrogram* describing the sequence of merges



Distance between clusters: Ward's linkage

$$\delta(C, C') = \frac{S(C)}{|C|} + \frac{S(C')}{|C'|} - \frac{S(C \cup C')}{|C \cup C'|},$$

where $S(C) = \sum_{i \in C, j \in C} s_{ij}$

Adjacency-constrained HAC

Idea: exploit the natural *ordering* of loci along chromosomes

Algorithm: HAC but only merge **adjacent clusters**

- implemented in the R package `rioja`
- decreased time complexity wrt HAC: "only" $O(p^2)$ operations
- the output is a **set of nested segmentations**

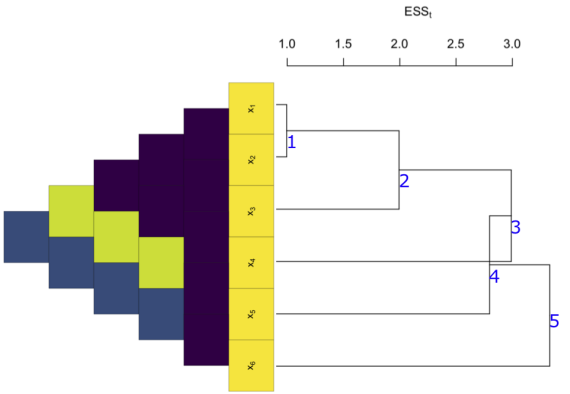
Questions

1. is this method mathematically valid for GWAS and Hi-C?
2. can we further decrease the time complexity?
3. is this method biologically relevant to GWAS and Hi-C?

1. Mathematical validity

Positive definite?

1. If yes, simple interpration as a distance to the "mean element"
2. If not, strange things happen:



--

Miyamoto et al. 2015

1. Always possible to make it positive definite :

$$s'_{ii} = s_{ii} + \lambda$$

2. s and s' yields the same hierarchy

2. A faster algorithm

Quadratic complexity can be too much

Hi-C, GWAS: $p \sim 10^4 - 10^5$ for each chromosome.

Space complexity $(O(p^2))$

could be improved in specific applications

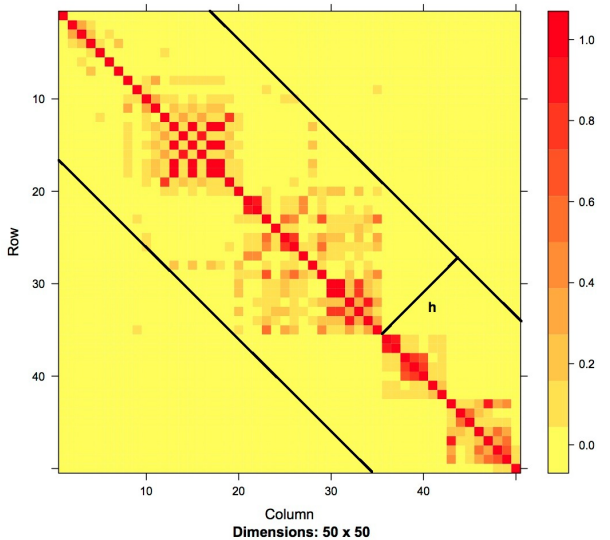
Time complexity $(O(p^2))$

cannot be improved without further assumptions

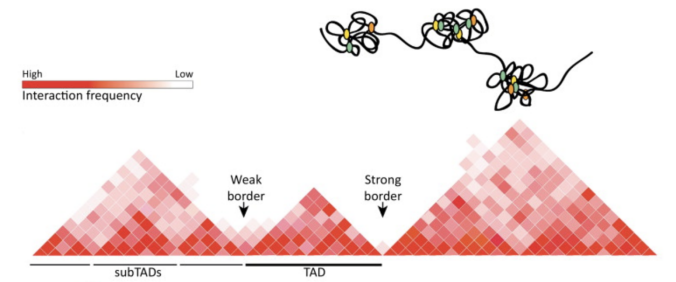
Reduce complexity while maintaining interpretability?

Assumption: **band diagonal similarity**

GWAS



Hi-C



Consequence: at most $p \times h$ items in S (with $h \ll p$)

Description of the algorithm

1. Precomputation of certain cumulative sums

aka: pencil trick ✏️

relies on:

$$\delta(C, C') = \frac{S(C)}{|C|} + \frac{S(C')}{|C'|} - \frac{S(C \cup C')}{|C \cup C'|},$$

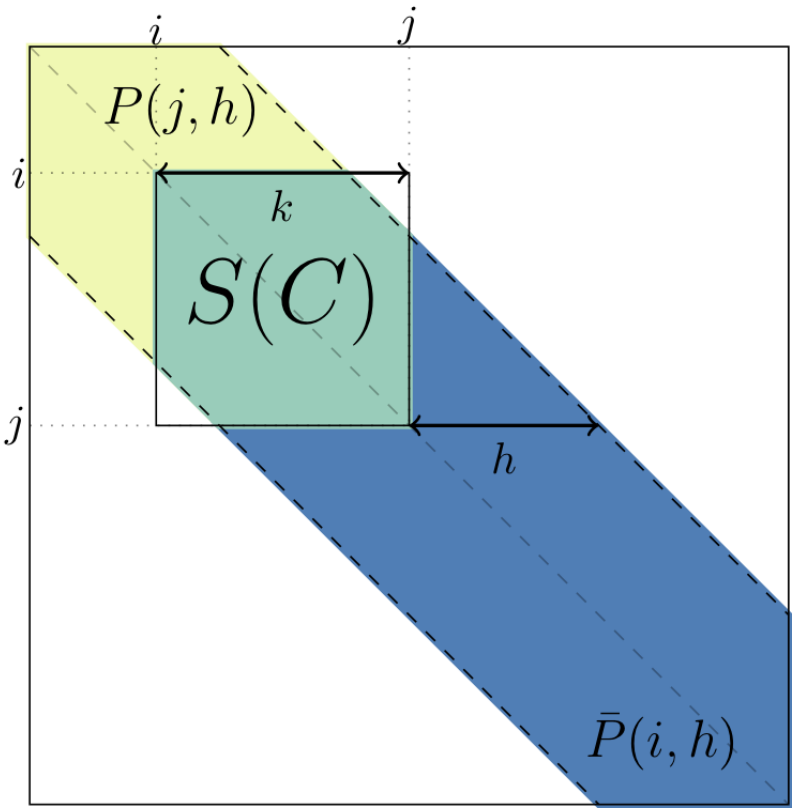
where $S(C) = \sum_{i \in C, j \in C} s_{ij}$

2. Storing candidate merges in a **min heap**

Implementation: R package **adjclust**

1. Pencil trick

$S(C)$ may be written as a simple function of precalculated sums:



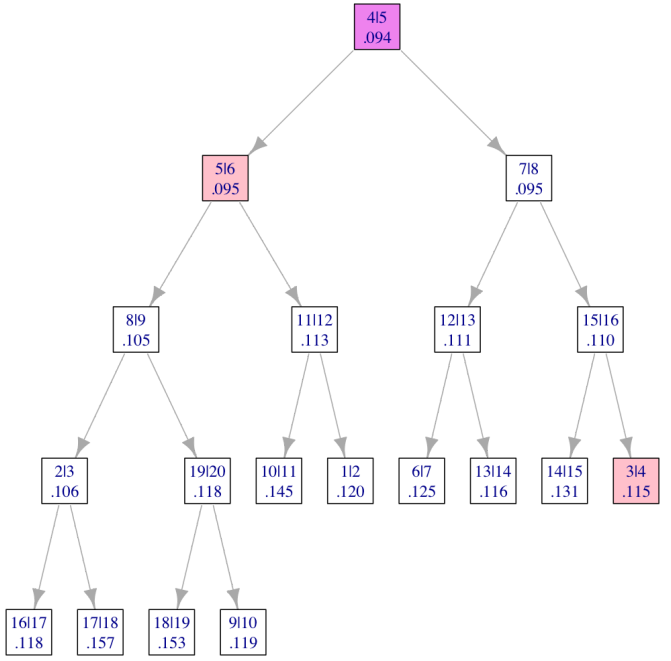
2. Storing candidate fusions in a min heap

A Min heap

- Binary tree
- Property : parent $<$ child
- Updates in $O(log(p))$

Ordering given by the linkage δ

$\rightarrow$ next candidate fusion is the root

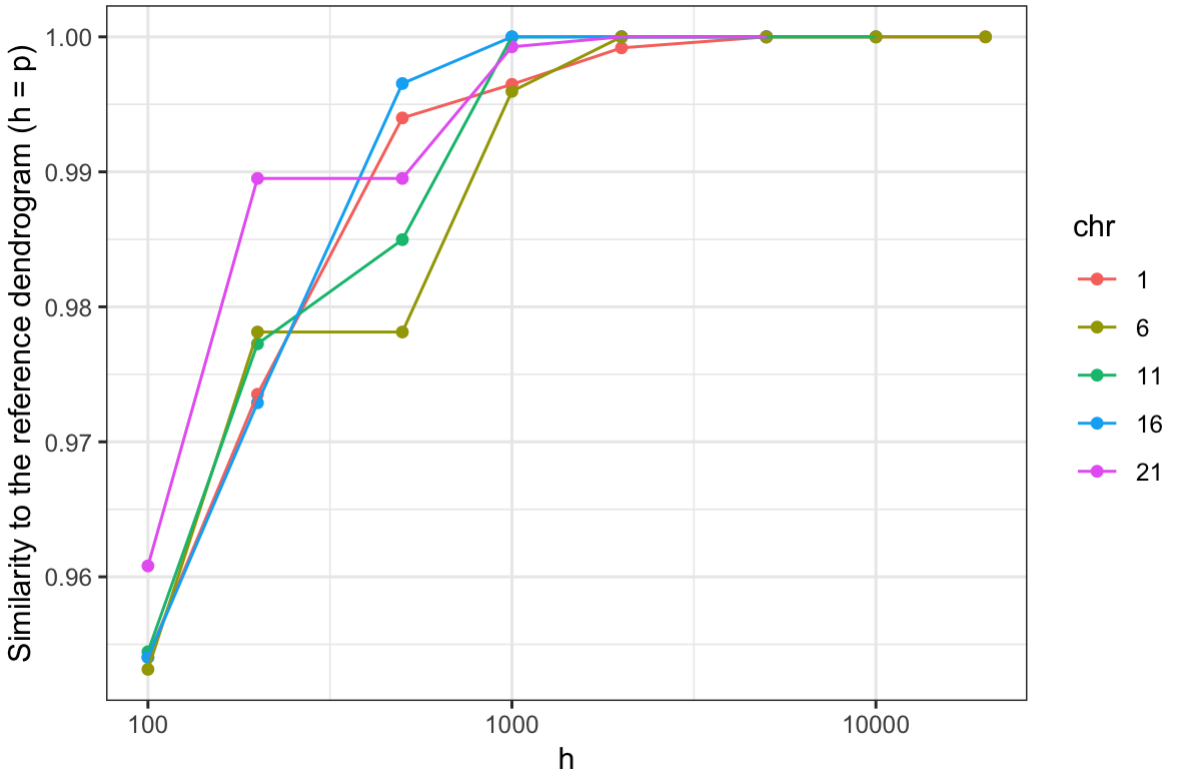


Overall complexity of the algorithm

$O(ph)$ in space and $O(p(h + log(p))$ in time.

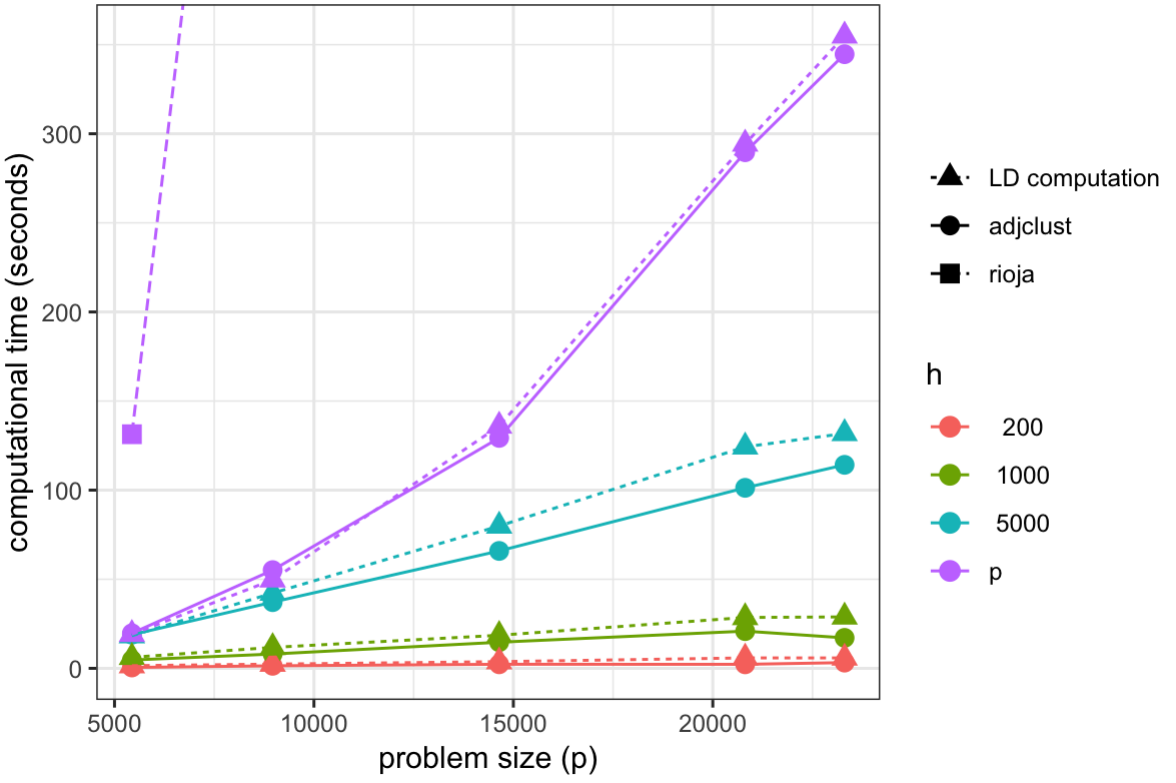
3(a). Application to GWAS

Quality of the band approximation



The clusterings are quasi-identical regardless of the value of h

Improved time complexity



3.(b) Application to Hi-C

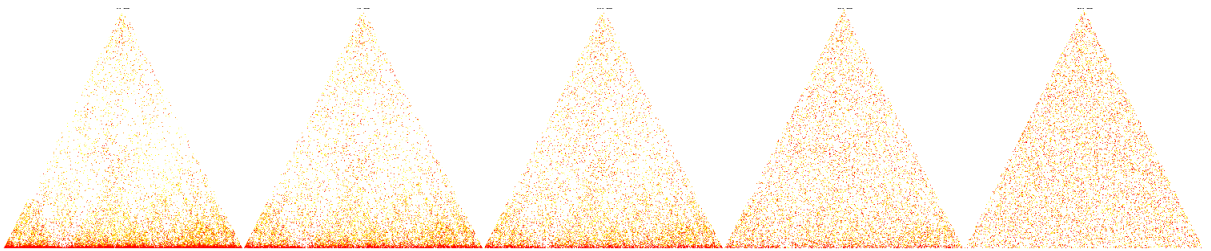
HAC vs constrained HAC

- HAC :a greedy algorithm to minimize the total inertia of a partition
- constrained HAC: greatly reduced search space

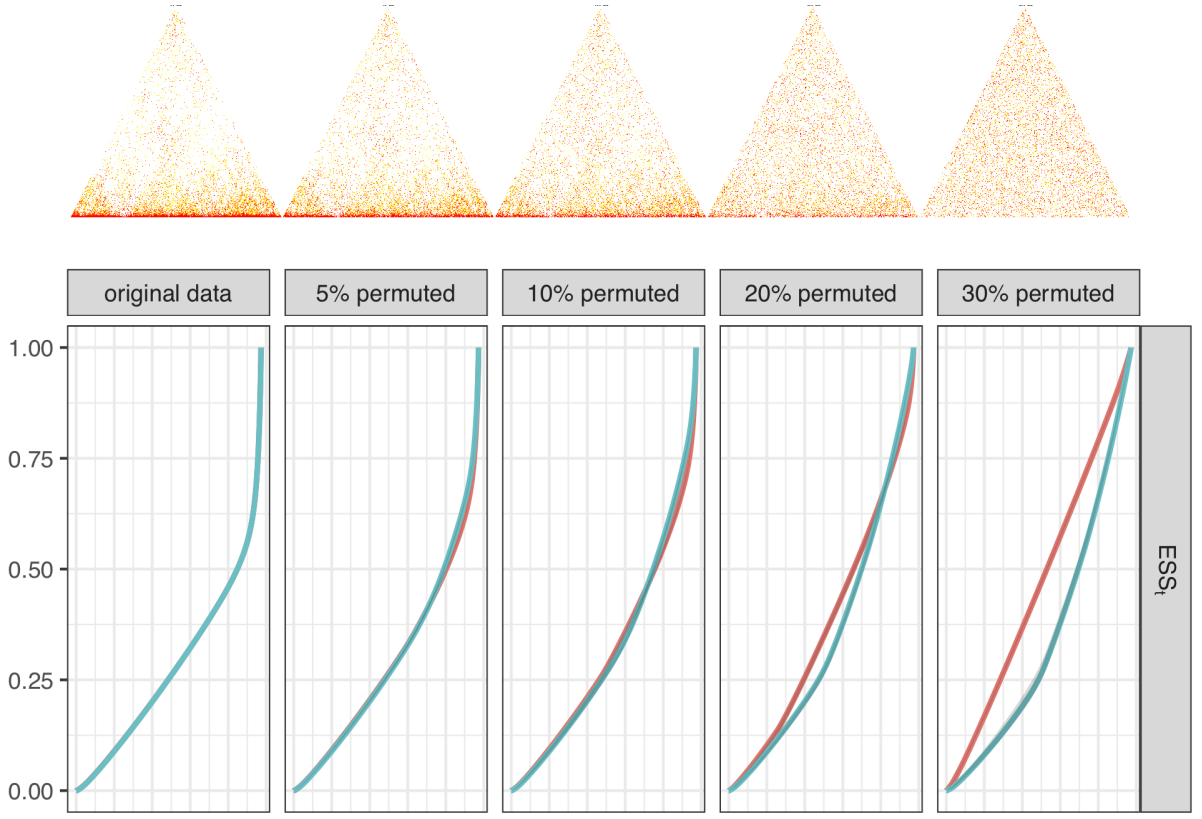
$\Rightarrow$ Expect inertia(HAC) $<$ inertia(constrained HAC) ?

Case study

- Start with Hi-C map of chromosome 3 of ESC in Dixon et al (2012)
- Perturbate the original map by swapping more and more elements:

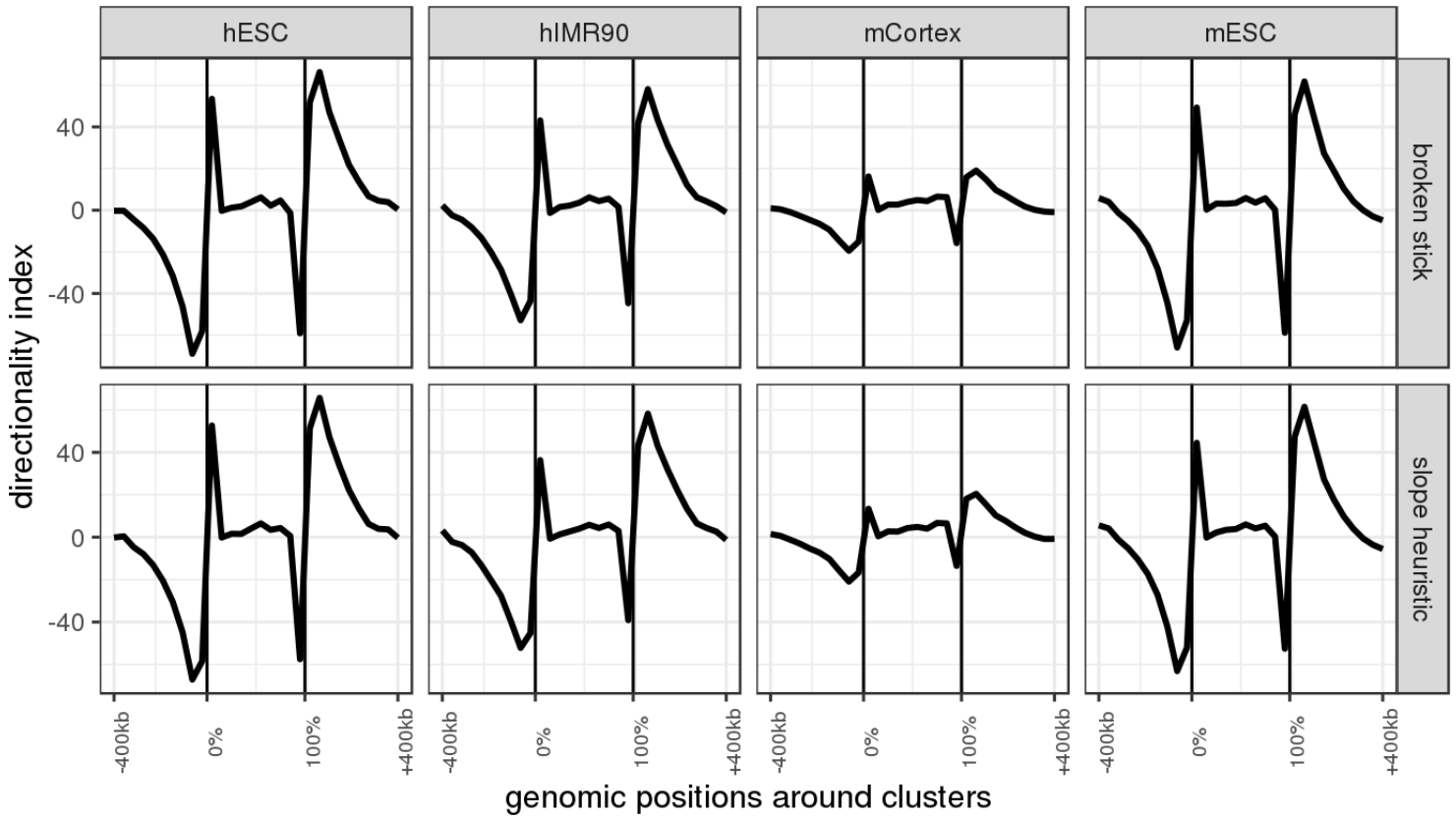


HAC vs constrained HAC: total inertia



Evolution of the Directionality Index around clusters

DI values are expected to show a sharp variation at TADs boundaries



Conclusion and perspective

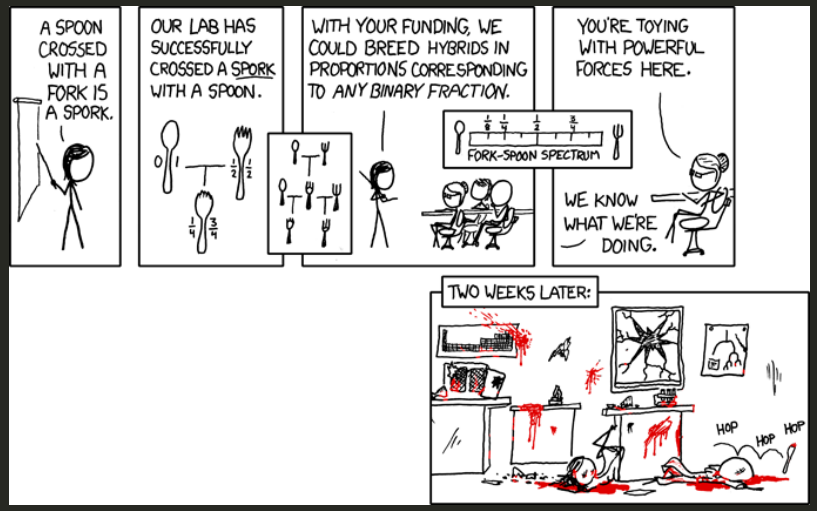
Contributions

- mathematically-sound method
- algorithm achieving quasi-linear complexity
- implementation: R package `adjclust`
  - stable version on CRAN, devel version on github
- genomic applications
  - adding a contiguity constraint can improve the segmentation
  - no loss in performance when taking $h < p$

Perspectives/open questions

- automatic choice of the parameter h
- where to cut the dendrogram? (= model selection)
- differences between two biological conditions

Thank you for your attention!



Toying with powerfull forces : Computer Science, Statistics and Biology!