



**HAL**  
open science

## Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics

Christophe Ambroise, Alia Dehman, Pierre Neuvial, Guillem Rigaiil, Nathalie Vialaneix

► **To cite this version:**

Christophe Ambroise, Alia Dehman, Pierre Neuvial, Guillem Rigaiil, Nathalie Vialaneix. Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics. *Statistical Methods for Post Genomic Data (SMPGD 2019)*, Jan 2019, Barcelona, Spain. 2019. hal-02790995

**HAL Id: hal-02790995**

**<https://hal.inrae.fr/hal-02790995v1>**

Submitted on 5 Jun 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

# Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics

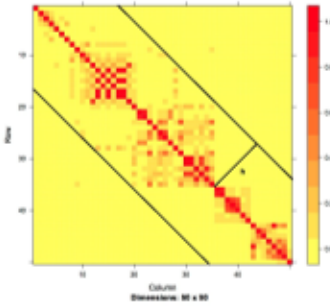
Christophe Ambroise<sup>1</sup>, Alia Dehman<sup>2</sup>, Pierre Neuvial<sup>3</sup>, Guillem Rigail<sup>4</sup> and Nathalie Vialaneix<sup>5</sup>

<sup>1</sup>LaMME, Evry • <sup>2</sup>Hyphen-stat, Toulouse • <sup>3</sup>Institut de Mathématiques de Toulouse/CNRS • <sup>4</sup>IPS2, CNRS/INRA • <sup>5</sup>INRA MIAT •

## Motivation: Regionally-structured genomic data

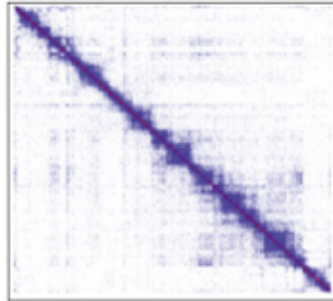
### Genome-Wide Association Studies (GWAS)

- loci: SNP
- similarity: linkage disequilibrium
- regions: LD/haplotype blocks



### Chromosome contact maps (Hi-C)

- loci: binned genome positions
- similarity: contact intensity
- regions: TAD; A/B compartments



## Key 2: Storing candidate fusions in a min-heap

### Min heap

#### A partially ordered binary tree

- nodes = candidate merges
- ordering given by the linkage  $\delta$

→ next candidate fusion is the root of the heap

#### Complexity

- $O(ph)$  in space
- $O(p(h + \log(p)))$  in time



## Implementation

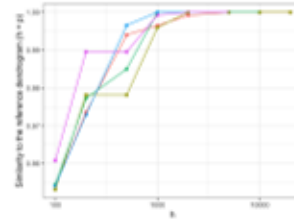
### R package `adjclust`<sup>3</sup>

- plots of similarity, dendrogram and clustering
- wrappers for SNP or Hi-C data analyses
- model selection by broken stick<sup>4</sup> or slope heuristic<sup>5</sup>

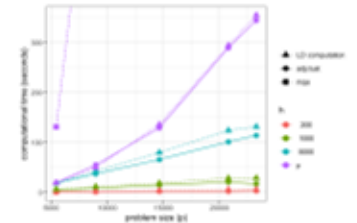
## GWAS: inferring linkage disequilibrium blocks

### Band approximation

Quality index: proportion of approximation vs  $h$



### Scalability

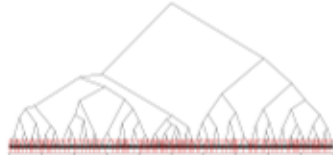


Data from [6]

## Goal: Segmentation by constrained HAC

### Hierarchical Agglomerative Clustering (HAC)

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



### Adjacency-constrained HAC: only merge adjacent clusters

- Improved time complexity: quadratic ( $O(p^2)$ )
- Space complexity ( $O(p^2)$ ): can be improved in specific applications<sup>3</sup>

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

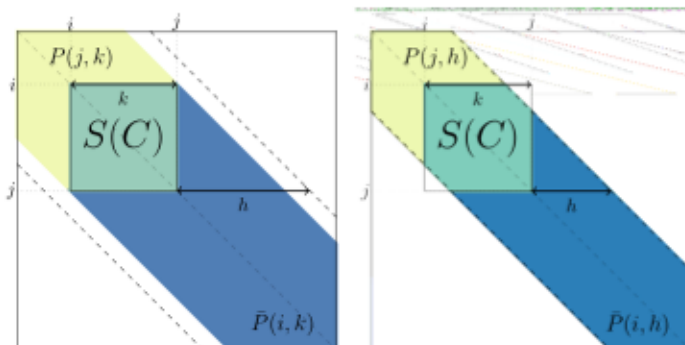
## Contribution: a quasi-linear algorithm<sup>2</sup>

Extra assumption: **band diagonal similarity**

## Key 1: Ward's linkage in constant time

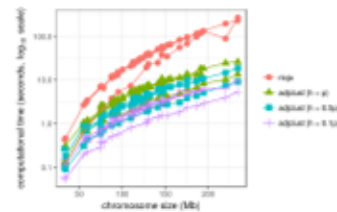
### Distance between clusters: Ward's linkage

$$\delta(C, C') = \frac{S(C)}{|C|} + \frac{S(C')}{|C'|} - \frac{S(C \cup C')}{|C \cup C'|}, \quad S(C) = \sum_{(i,j) \in C^2} s_{ij}$$



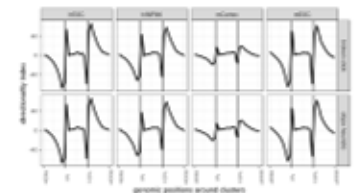
## Hi-C: inferring Topologically Associated Domains

### Influence of bandwidth



Data from [7] and [8]

### DI around clusters



Directionality Index (DI, [7]) values are expected to show a sharp variation at TADs boundaries

## References

- 1 A. Dehman, C. Ambroise, and P. Neuvial, BMC Bioinformatics **16**, 148 (2015).
- 2 C. Ambroise, A. Dehman, P. Neuvial, G. Rigail, and N. Vialaneix, (2019).
- 3 C. Ambroise and others, *Adjclust: Adjacency-Constrained Clustering of a Block-Diagonal Similarity* (2016).
- 4 K.D. Bennett, New Phytologist **132**, 155 (1996).
- 5 S. Ariot, V. Bault, J.-P. Baudry, and others, *Capush: Calibrating Penalties Using Slope Heuristics* (2016).
- 6 C. Dalmasso, W. Carpentier, L. Meyer, and others, PLoS ONE **3**, (2008).
- 7 J. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. Liu, and B. Ren, Nature **485**, 376 (2012).
- 8 Y. Shen, F. Yu, D.F. McCleary, Z. Ye, L. Edsall, and others, Nature **488**, 116 (2012).