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

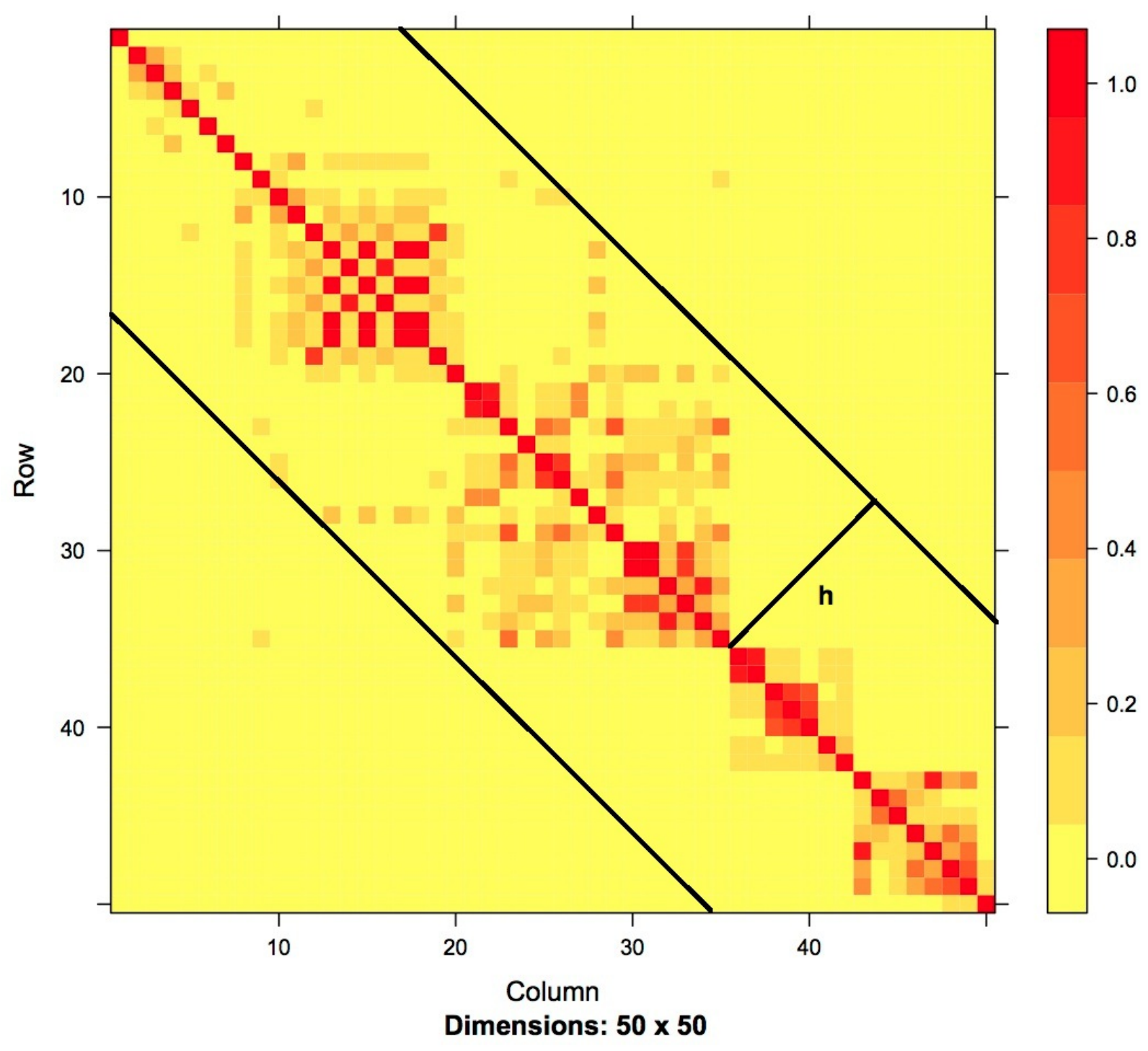
Christophe Ambroise[1], Alia Dehman[2], **Pierre Neuvial[3]**, Guillem Rigaill[4] and Nathalie Vialaneix[5]

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

Motivation: Regionally-structured genomic data

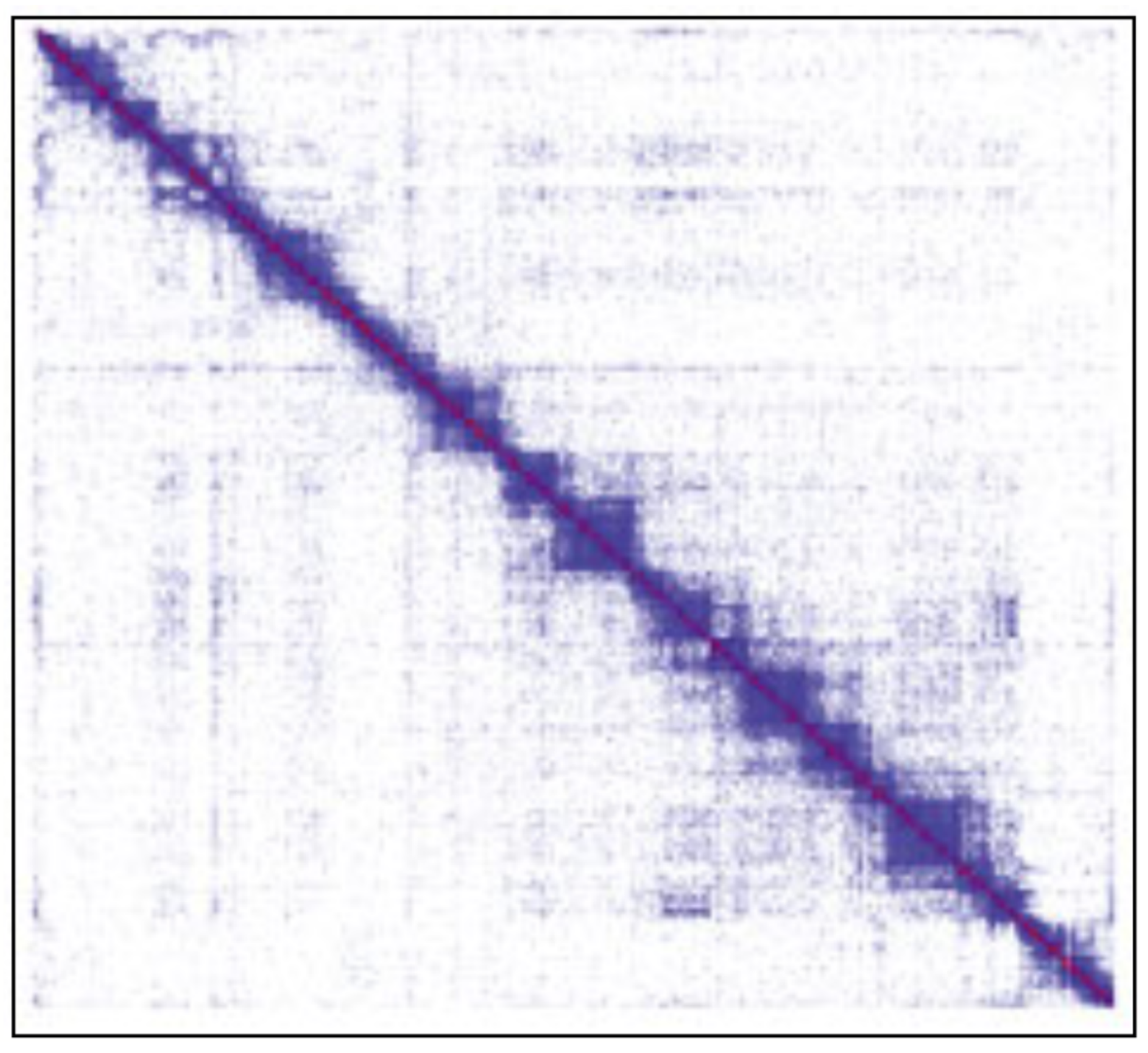
Genome-Wide Association Studies (GWAS)

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



Chromosome contact maps (Hi-C)

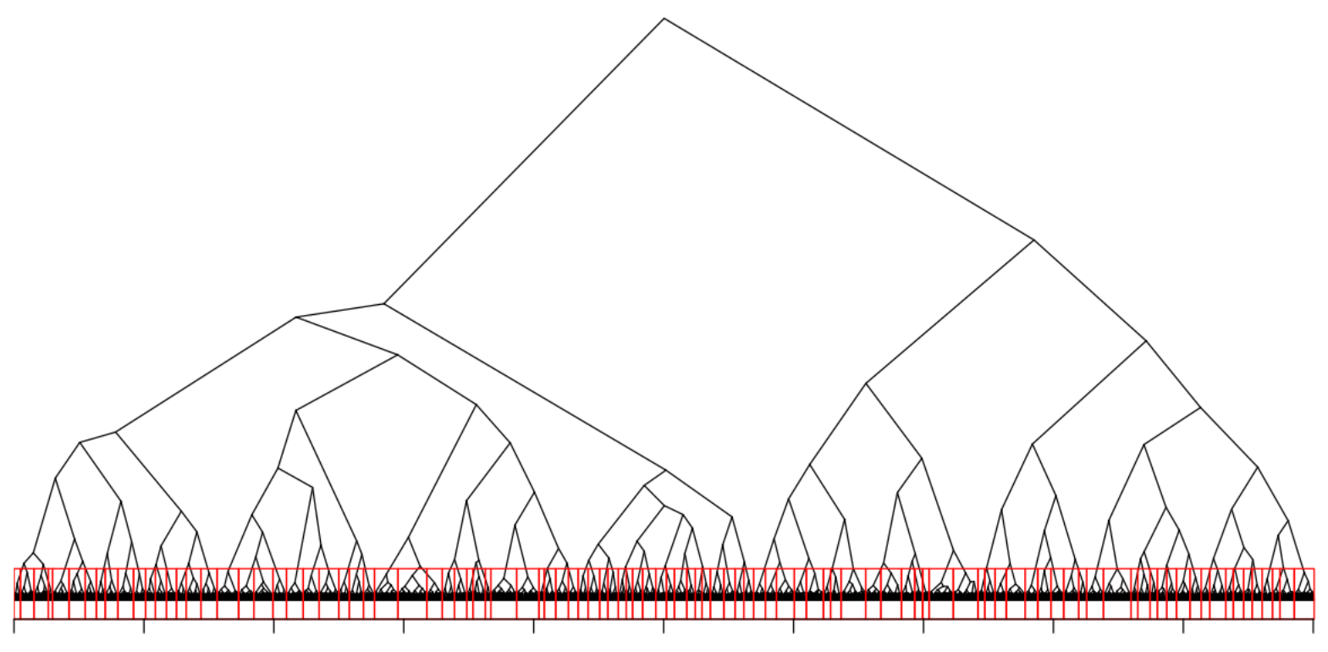
- loci: binned genome positions
- similarity: contact intensity
- regions: TAD; A/B compartments



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[1]

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

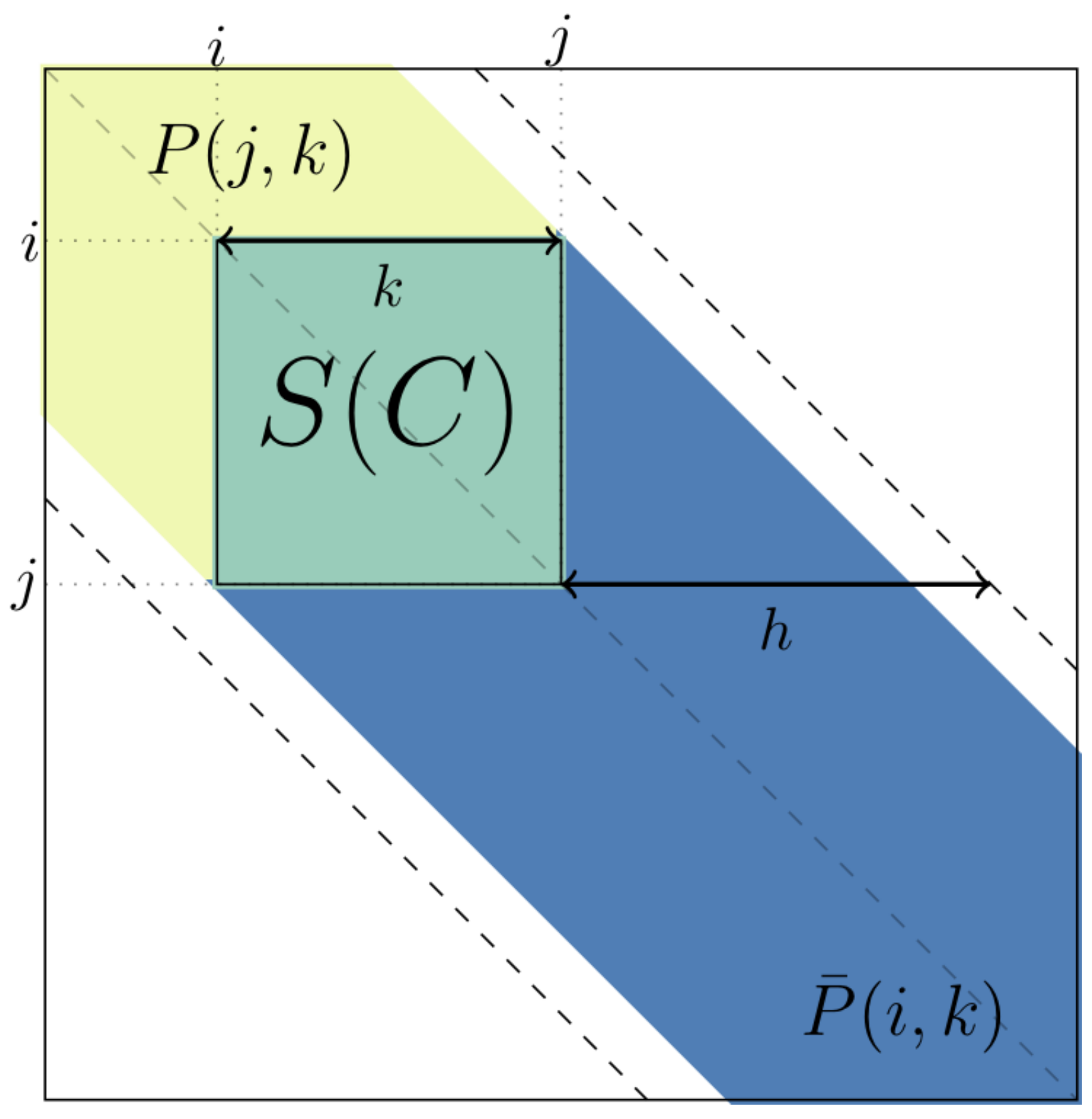
Contribution: a quasi-linear algorithm[2]

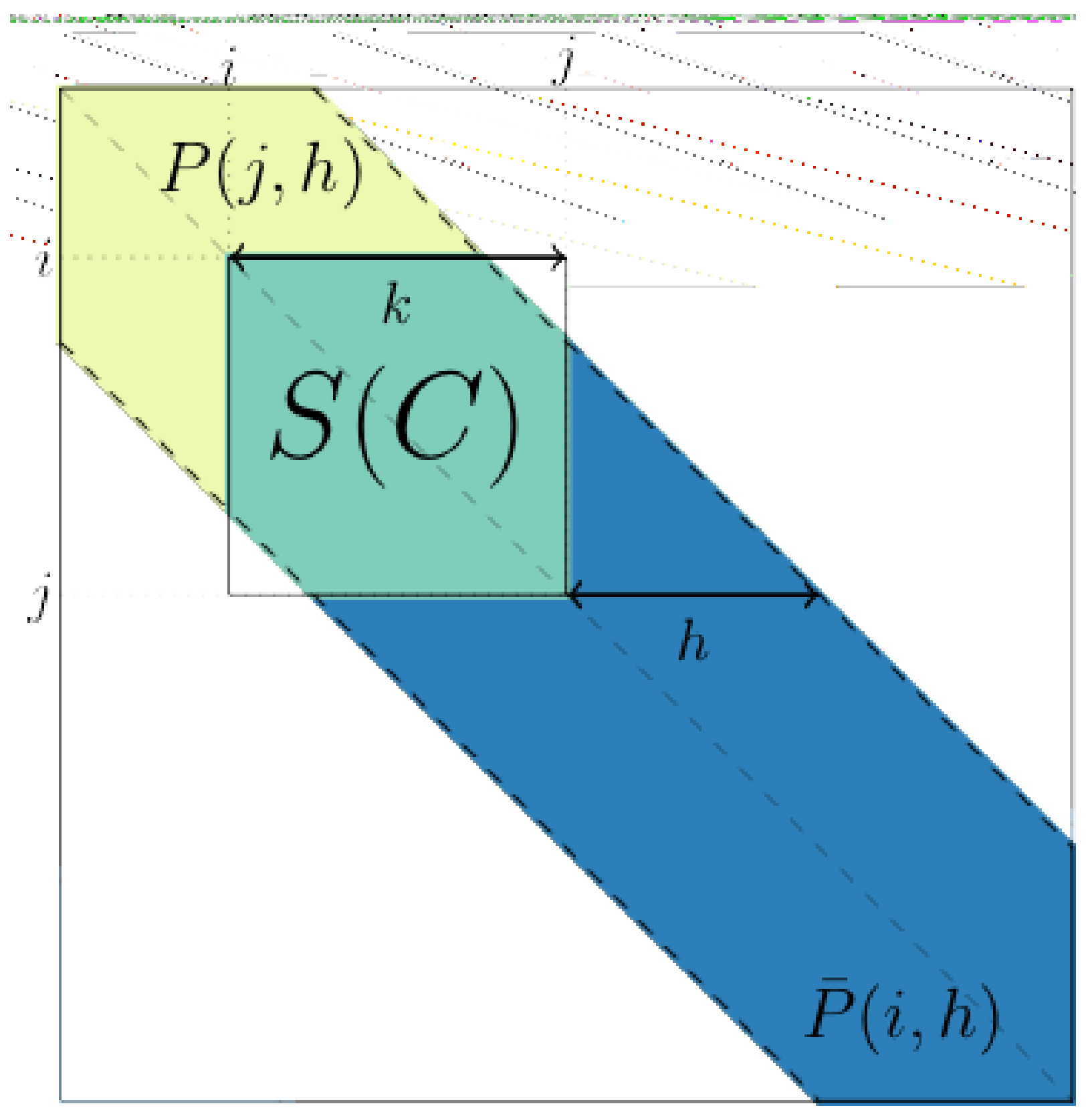
Extra assumption: **band diagonal similarity**

Key 1: Ward's linkage in constant time

Distance between clusters: Ward's linkage

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





Key 2: Storing candidate fusions in a min-heap

Min heap

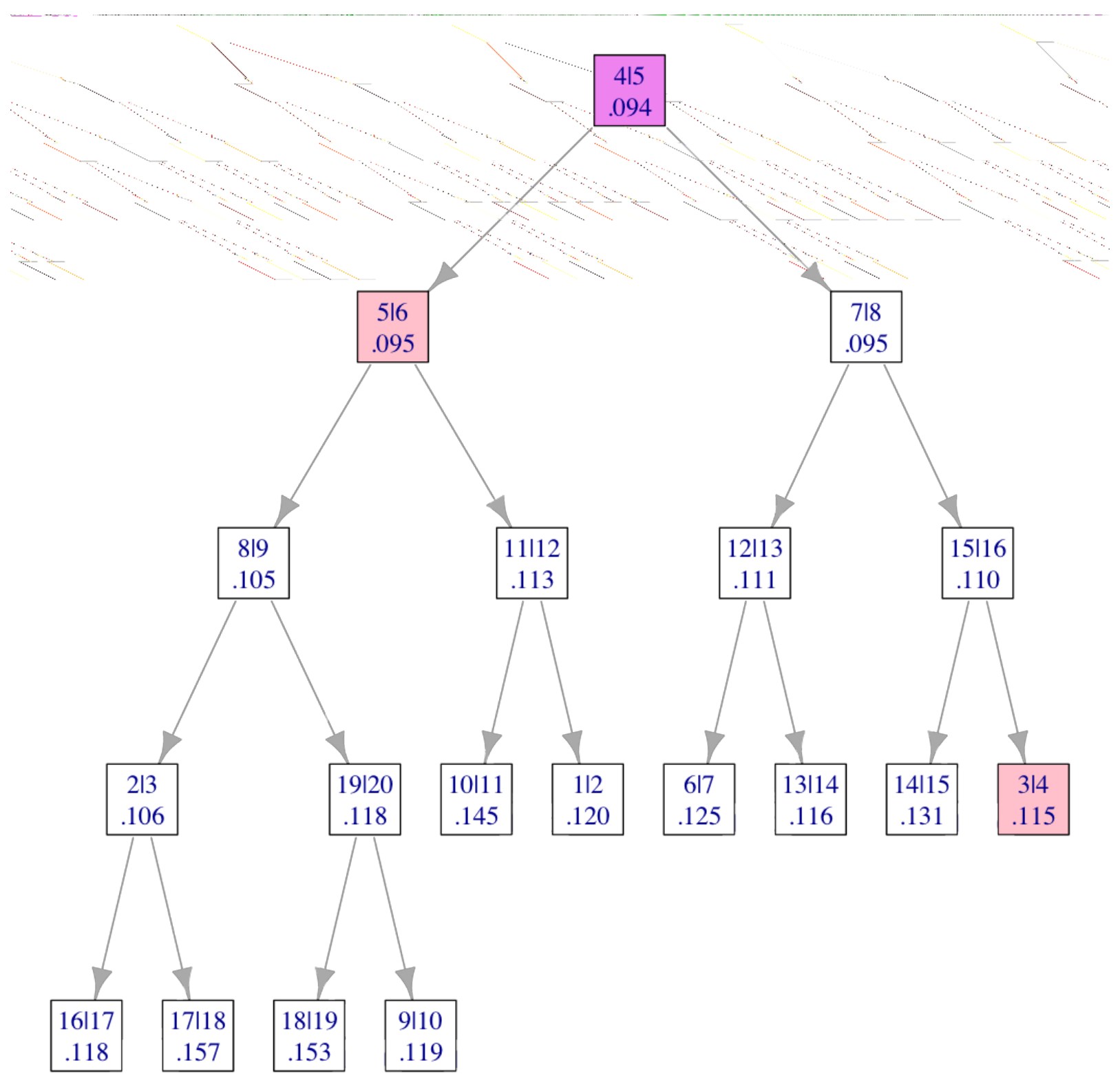
A partially ordered binary tree

- nodes = candidate merges
- ordering given by the linkage δ

→ next candidate fusion is the root of the heap

Complexity

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



Implementation

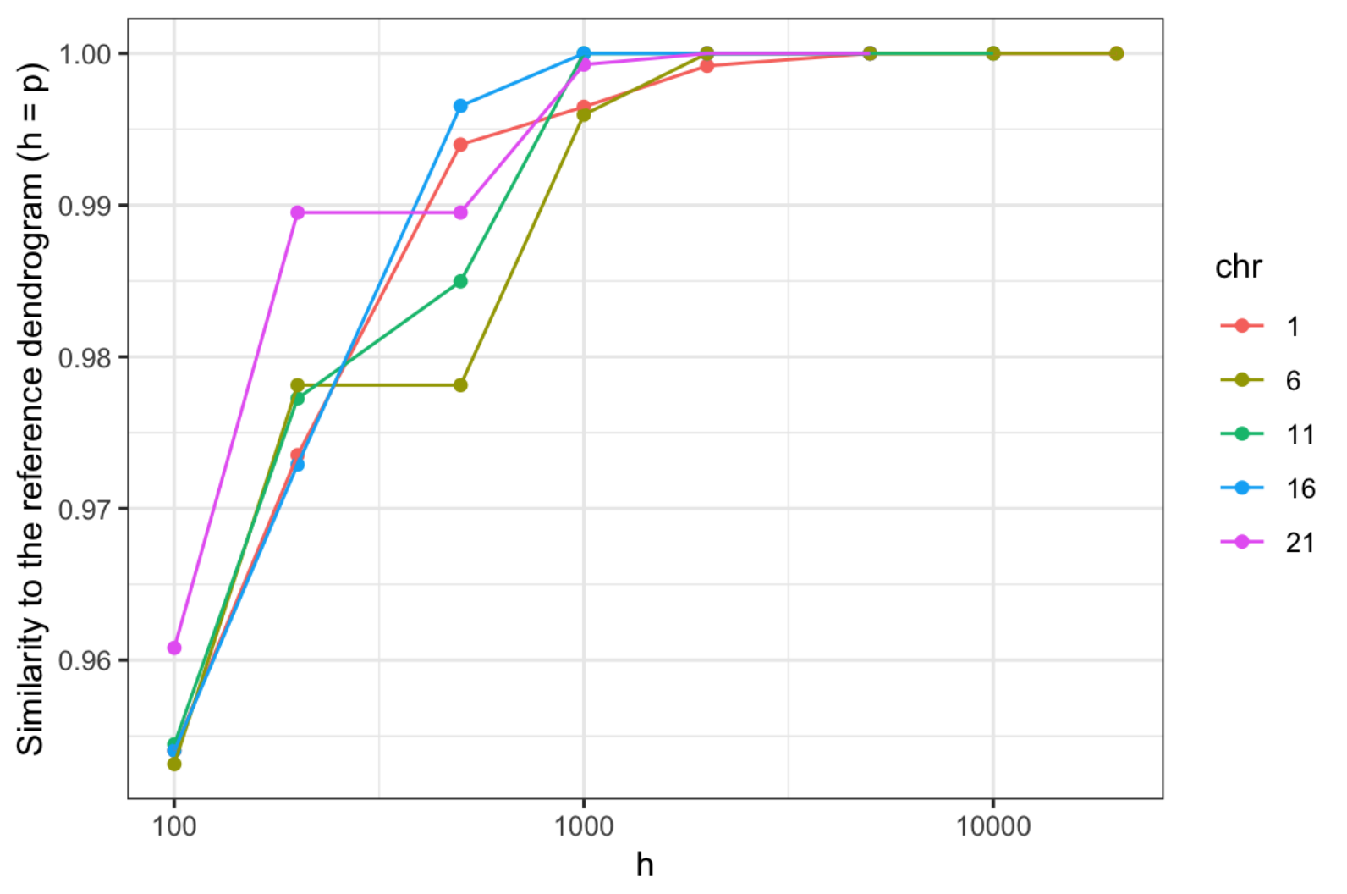
R package `adjclust`[3]

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

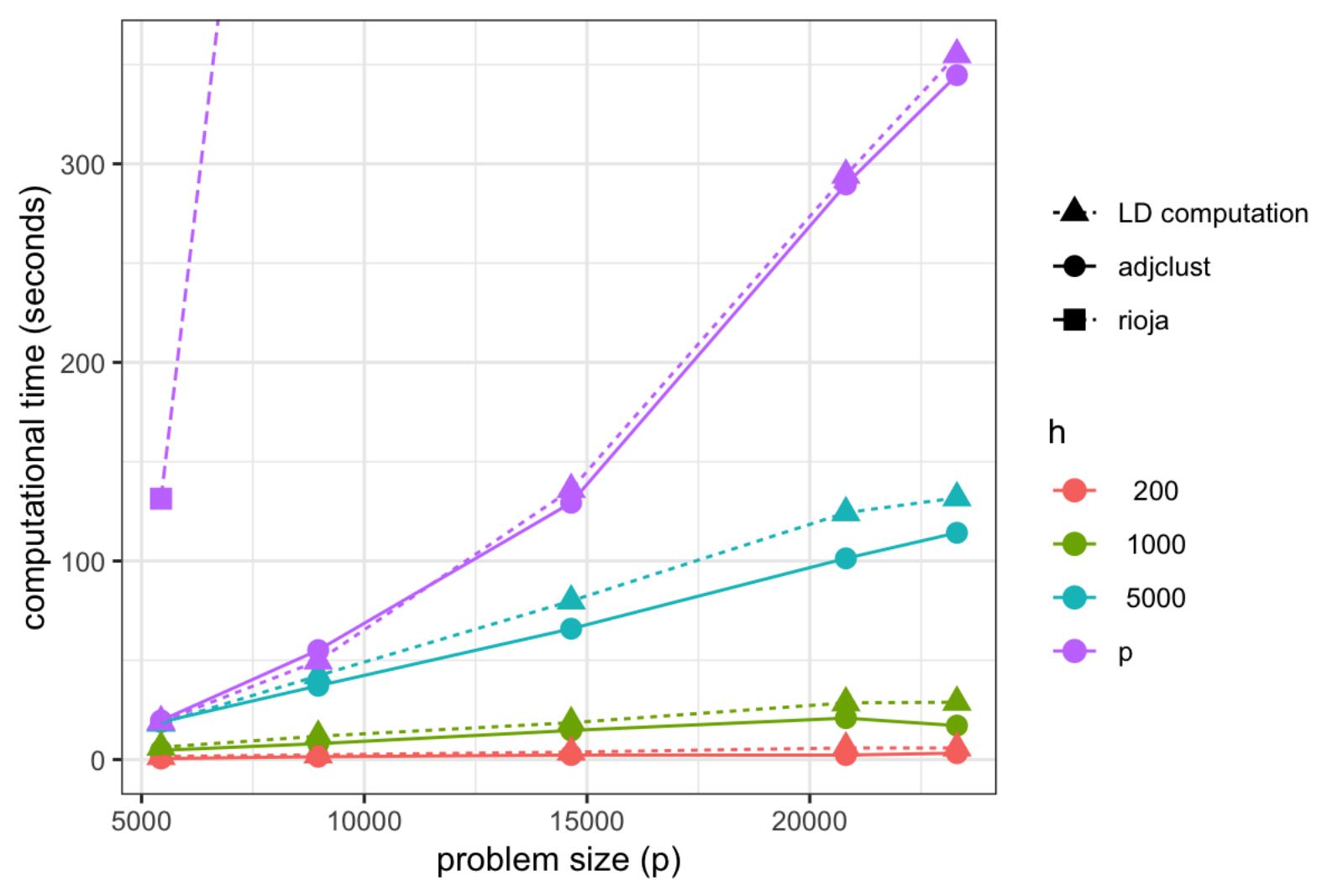
GWAS: inferring linkage disequilibrium blocks

Band approximation

Quality index: proportion of of approximation vs h



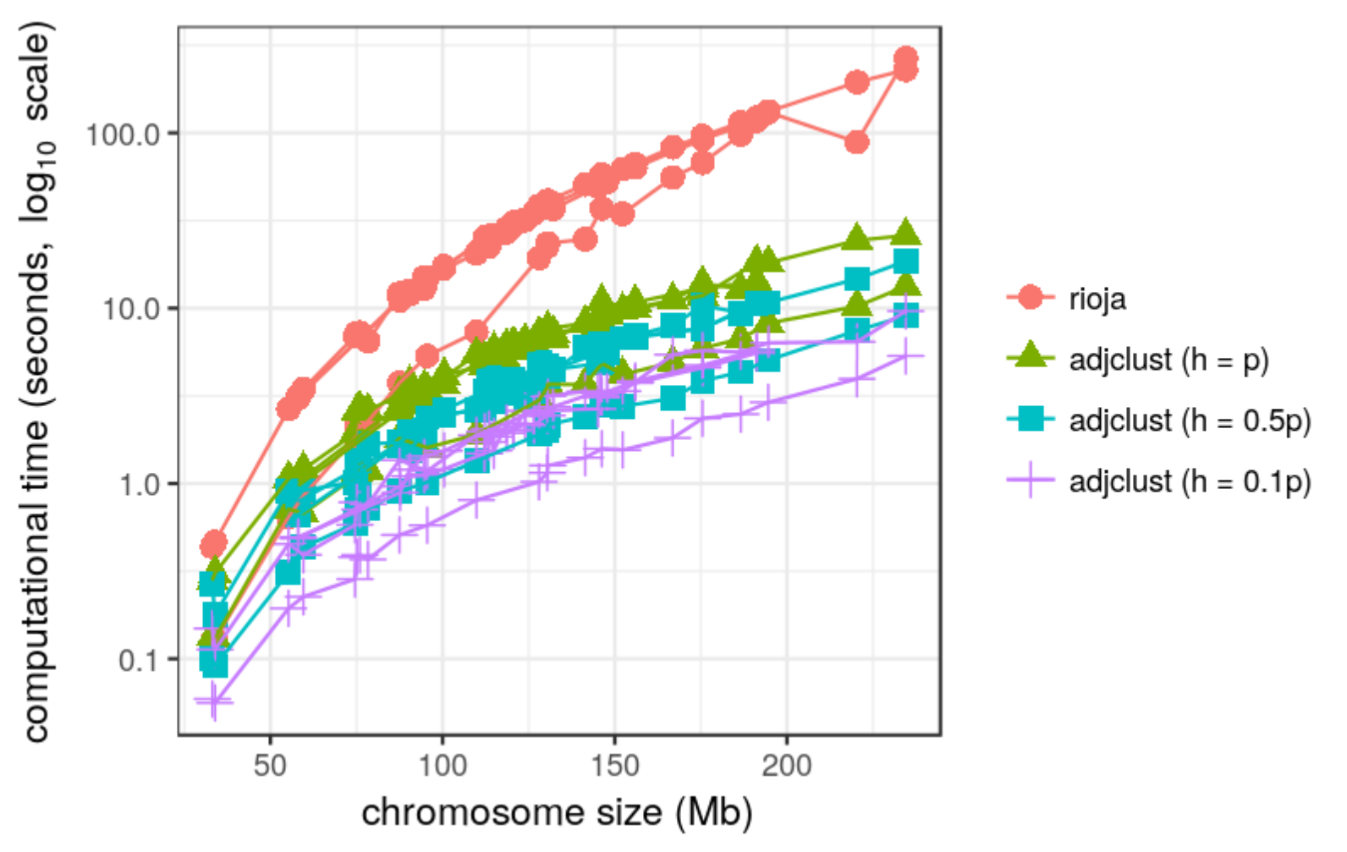
Scalability



Data from [6]

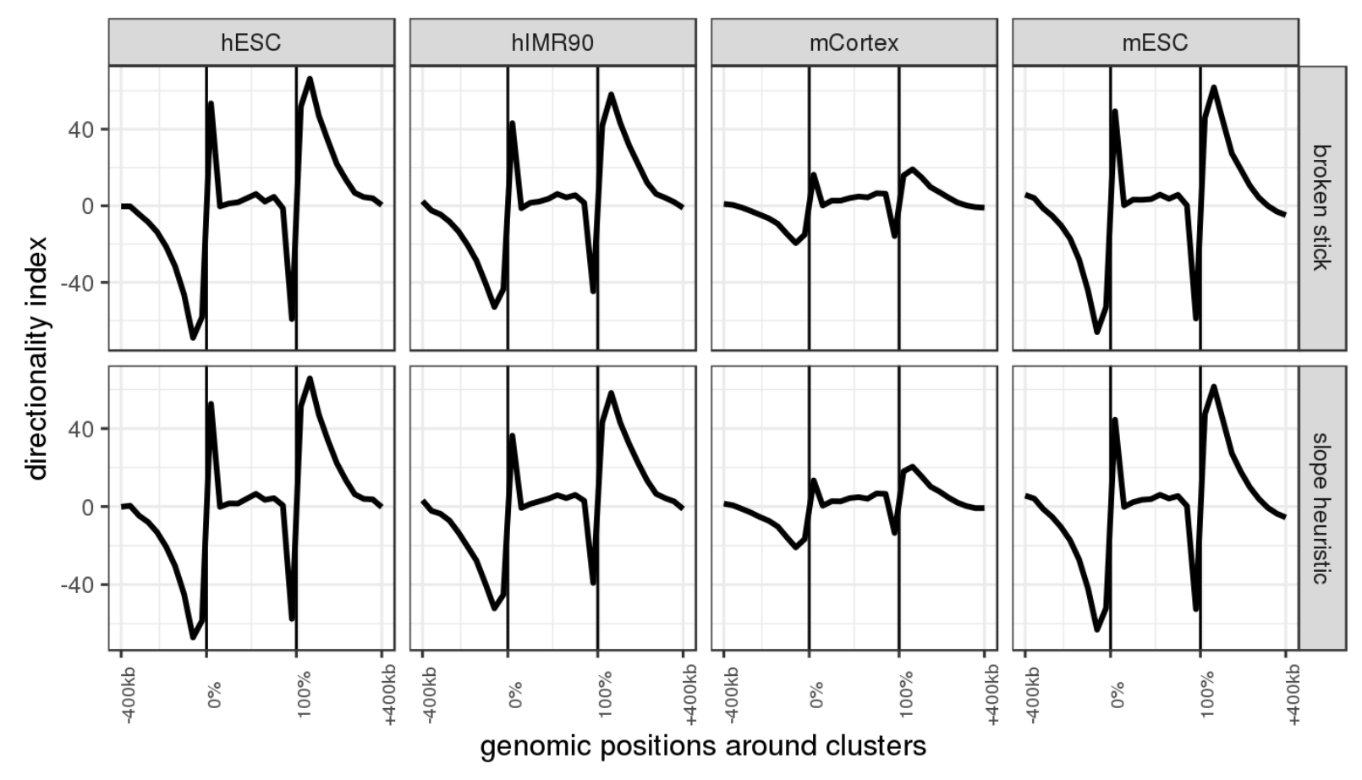
Hi-C: inferring Topologically Associated Domains

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. Rigaill, and N. Vialaneix, (2019).
[3] C. Ambroise and others, *Adjclust: Adjacency-Constrained Clustering of a Block-Diagonal Similarity* (2018).
[4] K.D. Bennett, New Phytologist **132**, 155 (1996).
[5] S. Arlot, V. Brault, J.-P. Baudry, and others, *Capushe: CAlibrating Penalities 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).