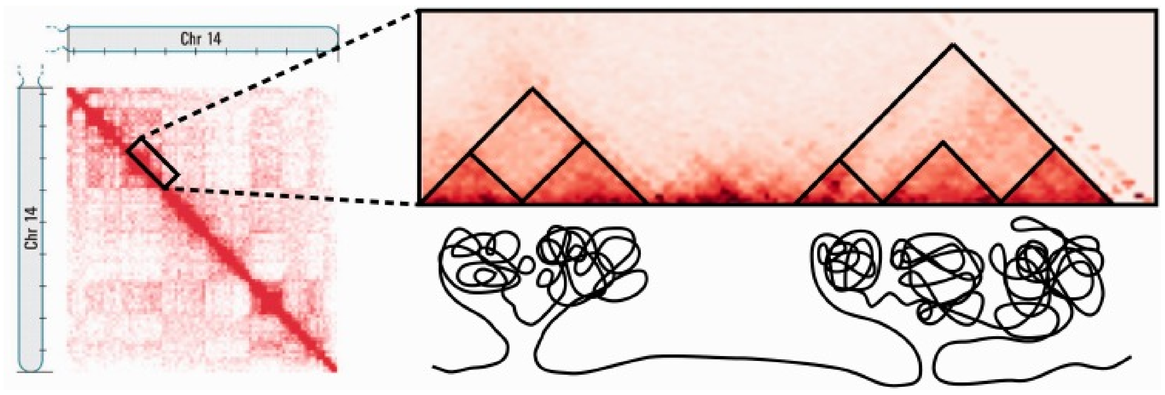
Comparative analysis of 3D genomics data

Elise Jorge[1,2], Pierre Neuvial[3], Nathalie Vialaneix[2] & Sylvain Foissac[1]

[1] GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet-Tolosan, France, [2] Université Fédérale de Toulouse, INRAE, MIAT, 31326 Castanet-Tolosan
[3] Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, CNRS, UPS, F-31062 Toulouse Cedex 9, France
elise.jorge@inrae.fr

3D genomics and Hi-C matrix

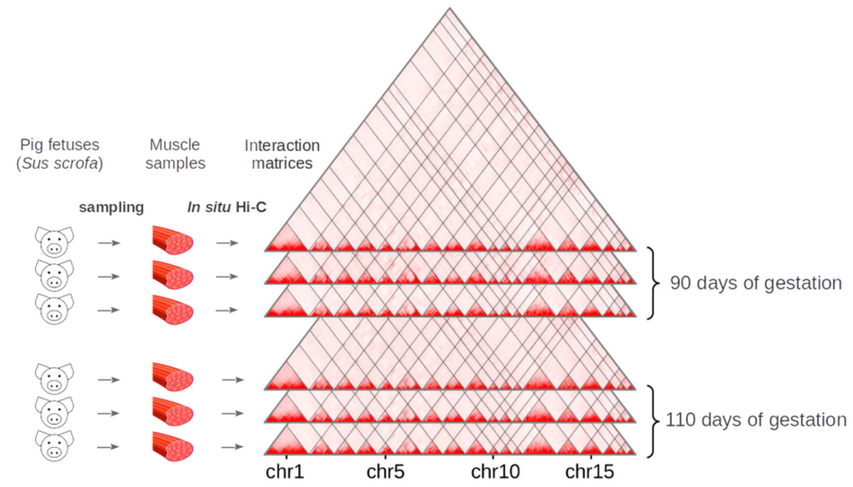
- Hi-C experiment: estimates 3D proximity of genomic positions by measuring interaction frequencies between pairs of positions.
- Result: symmetric matrix of interaction counts (Hi-C matrix).



- Structural changes: critical and linked to neurological pathologies [1], malformations [2], etc.

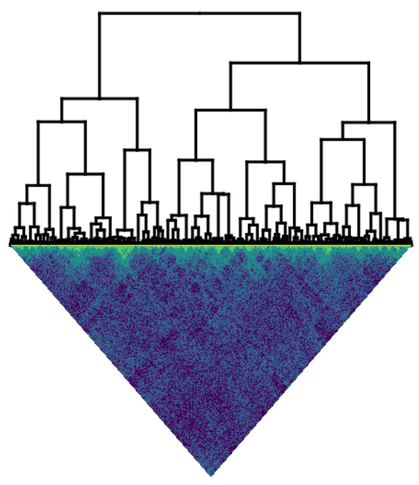
Differential analysis of Hi-C matrices

- Data: Hi-C matrices replicates obtained in two biological conditions.
- Objective: Identify genomic regions with significantly different interactions between the two conditions.



From Hi-C matrices to hierarchical clustering

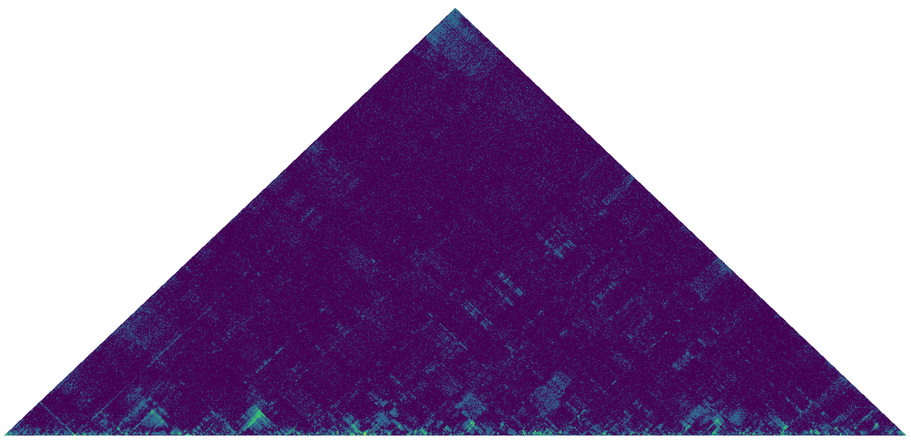
- CCHAC [3]: cluster (dis)similarity data with contiguity constraint.
- Well suited for Hi-C data → similarity data + respects spatial structure.



- Can be used to **browse the data** and identify interesting clusters.

Differential analysis: current methods

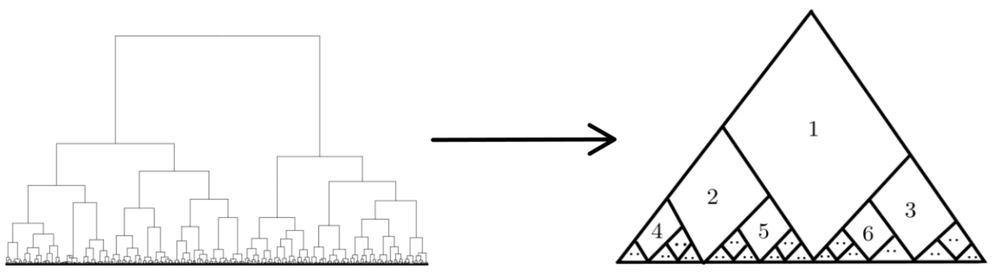
- One interaction → one hypothesis → one p-value.
- *diffHiC* [4]: one of the best available methods [5].



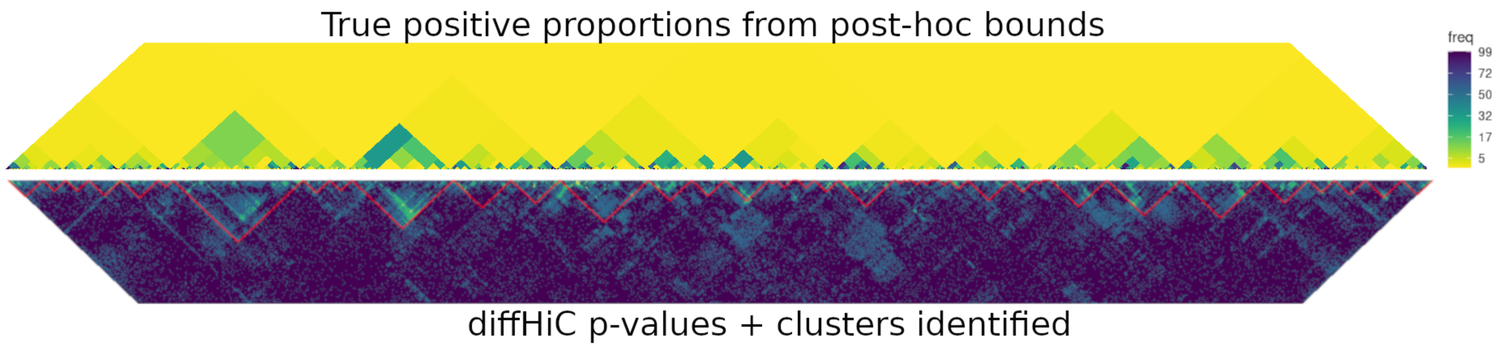
- Problem: interactions ≠ regions → **limited biological interpretation**.
- Need to find a way to "aggregate" p-values and create clusters.

A new method for differential analysis

- Main idea: Use a post-hoc strategy to quantify the number of true positives in a cluster using a hierarchical representation of the data and p-values provided by *diffHiC*.
- STEP 1: Clustering.
  1. Apply a CCHAC to the sum of all matrices → get a **unique clustering**.
  2. Look only at the **rectangles** formed by interactions between two sub-clusters of the same cluster.



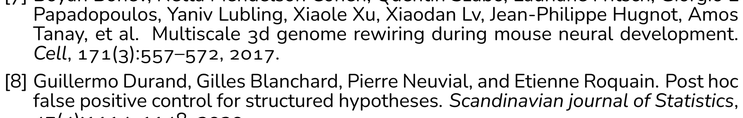
- STEP 2: Post-hoc analysis.
  Post-hoc methods [6] aim at controlling the number of **false positives** in a set **independently** of how the set has been chosen.
  - For each **rectangle**, compute the **post-hoc upper bound** on false positives using *diffHiC* p-values.
  - Declare as significant the clusters for which the **proportion of true positives** in the corresponding rectangle is above a user-defined threshold.



- The method is **data-driven** and allows to detect genomic regions containing a **significant amount of differential interactions**, as shown above with mouse data [7].

Future work

- Studying the **impact** of the **hierarchy**: intra-condition variability + inter-conditions variability → potentially very different hierarchies.
- **Testing different thresholds:** leads to different (but nested) clusters identified.



- Exploring the possibility of using the **nested structure** of hypotheses to compute **tighter** post-hoc bounds [8].

References

[1] Malte Spielmann, Darío G Lupiáñez, and Stefan Mundlos. Structural variation in the 3d genome. *Nature Reviews Genetics*, 19(7):453–467, 2018.

[2] Darío G Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili, John M Opitz, Renata Laxova, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, 2015.

[3] Nathanaël Randriamihamison, Nathalie Vialaneix, and Pierre Neuvial. Applicability and interpretability of Ward's hierarchical agglomerative clustering with or without contiguity constraints. *Journal of Classification*, 38:363–389, 2021.

[4] Aaron T.L. Lun and Gordon K. Smyth. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics*, 16:258, 2015.

[5] Elise Jorge, Sylvain Foissac, Pierre Neuvial, Zytnicki Matthias, and Nathalie Vialaneix. Benchmark of hi-c data differential analysis methods. In prep, 2024.

[6] Gilles Blanchard, Pierre Neuvial, and Etienne Roquain. Post hoc confidence bounds on false positives using reference families. *The Annals of Statistics*, 48(3):1281–1303, June 2020.

[7] Boyan Bonev, Netta Mendelson Cohen, Quentin Szabo, Lauriane Fritsch, Giorgio L Papadopoulos, Yaniv Lubling, Xiaole Xu, Xiaodan Lv, Jean-Philippe Hugnot, Amos Tanay, et al. Multiscale 3d genome rewiring during mouse neural development. *Cell*, 171(3):557–572, 2017.

[8] Guillermo Durand, Gilles Blanchard, Pierre Neuvial, and Etienne Roquain. Post hoc false positive control for structured hypotheses. *Scandinavian journal of Statistics*, 47(4):1114–1148, 2020.