Hi-C Differential Analysis: A new method using tree representation based on Contiguity Constrained Hierarchical Agglomerative Clustering (CCHAC)

N.Randriamihamison, M. Chavent, S. Foissac, P.Neuvial, N.Vialaneix

INSA, Toulouse

December 5, 2019

1. Pratical case and Data
2. State of the art
   - Bin pair level comparisons
   - Alternatives using structural comparisons
3. Differential Analysis method based on CCHAC
   - Hi-C and HAC
   - Method based on CCHAC
   - Preliminary results
4. Conclusion

Pratical case and Data

Introduction

Starting point :

→ work and data of M. Marti-Marimon PhD thesis:

Study of fetal development of piglets using Hi-C data:

→ Data produced by Centre INRA - Occitanie Toulouse :

- 3 Hi-C samples corresponding to 90 days of gestation
- 3 Hi-C samples corresponding to 110 days of gestation

Aim of the hierarchical differential analysis method:

- overcome limits linked to methods based on bin pair level comparisons

State of the art

Introduction and notation

Main question of Hi-C differential analysis:

Given two sets of Hi-C matrices, corresponding respectively to two biological conditions, how can we compare those two biological conditions with statistical guarantees ?

Notation:

- Considered biological conditions: $\mathcal{C}_i$ for $i \in \{1, 2\}$
- Hi-C matrices: H^t for $t \in \{1, \ldots, T\}$
- Interaction Counts: $H^t = (h^t_{ij})_{1 \leq i,j \leq p}$ where p is the number of bins

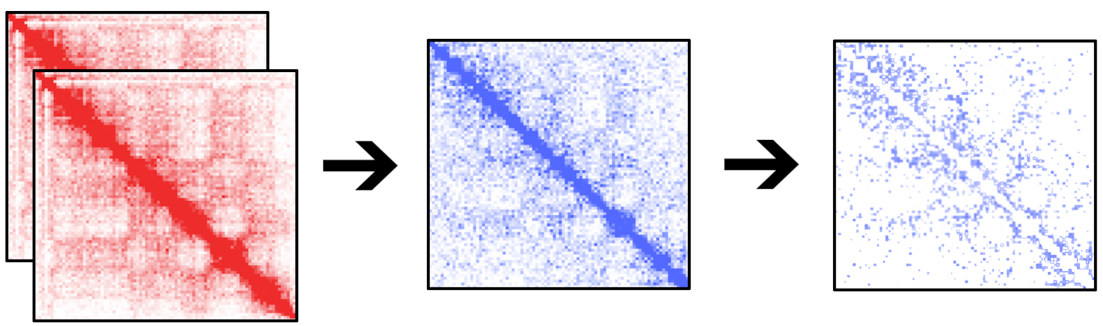
We have

- $\mathcal{C}_1 \cup \mathcal{C}_2 = \{1, \ldots, T\}$
- $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$

Bin pair level comparisons

Most methods realize comparisons at a **bin pair level**:

1. For each bin pair, compute a certain statistic
2. For each bin pair, deduce from the statistic a *p*-value
3. Apply correction for multiple testing
4. Obtain a list of differential bin pairs between the two conditions



Using Z scores

[Stansfield et al., 2018] developed a method implemented in the R package HiCcompare :
$\rightarrow$ cannot use replicate ($\mathcal{C}_1 = \{1\}$ and $\mathcal{C}_2 = \{2\}$)

1. For each bin pair (i, j), compute $m_{ij} = \log_2\left(\frac{h_{ij}^{\mathbf{2}}}{h_{ij}^{\mathbf{1}}}\right) = \log_2\left(h_{ij}^2\right) - \log_2\left(h_{ij}^1\right)$
2. For each bin pair, compute the associated Z-score:

$$z_{ij} = \frac{m_{ij} - m}{\sigma}$$

where m is the mean of the m_{ij}'s and σ their standard deviation

$\rightarrow$ deduce p-**values**

Limits:

- statistical guarantees are very limited
- does not account for intra-condition variability (no replicates)

Using $\mathcal{NB}$ distribution

[Lun and Smyth, 2015] developed a method implemented in the R package diffHic :
$\rightarrow$ can use replicates (at least 2 replicates by conditions)

1. Hi-C entries are modeled using negative binomial distributions:

$$h_{ij}^{t} \sim \mathcal{NB}(\mu_{ij}, \phi_{ij})$$

2. Test is performed identically as for RNA-seq

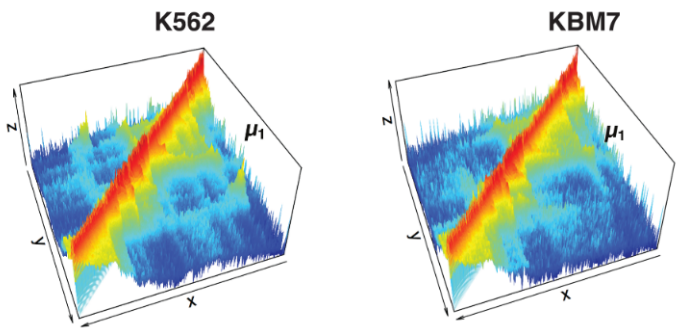
Limits:

- does not account for the depedency between bin pairs

Using the neighbouring structure of Hi-C maps

[Djekidel et al., 2018] developed a method implemented in the R package FIND :
$\rightarrow$ can use replicates (at least 2 replicates by conditions)

1. Represent counts h_{ij}^t by the triplet $(i, j, h_{ij}^t) \in \mathbb{R}^3$ and define $(i, j, \mu_{1/2})$ where $\mu_{1/2}$ is the mean of counts for the first/second condition



2. Statistical test based on a homogeneous spatial Poisson process
$\rightarrow$ similar to what is done in neuro-imaging comparisons.

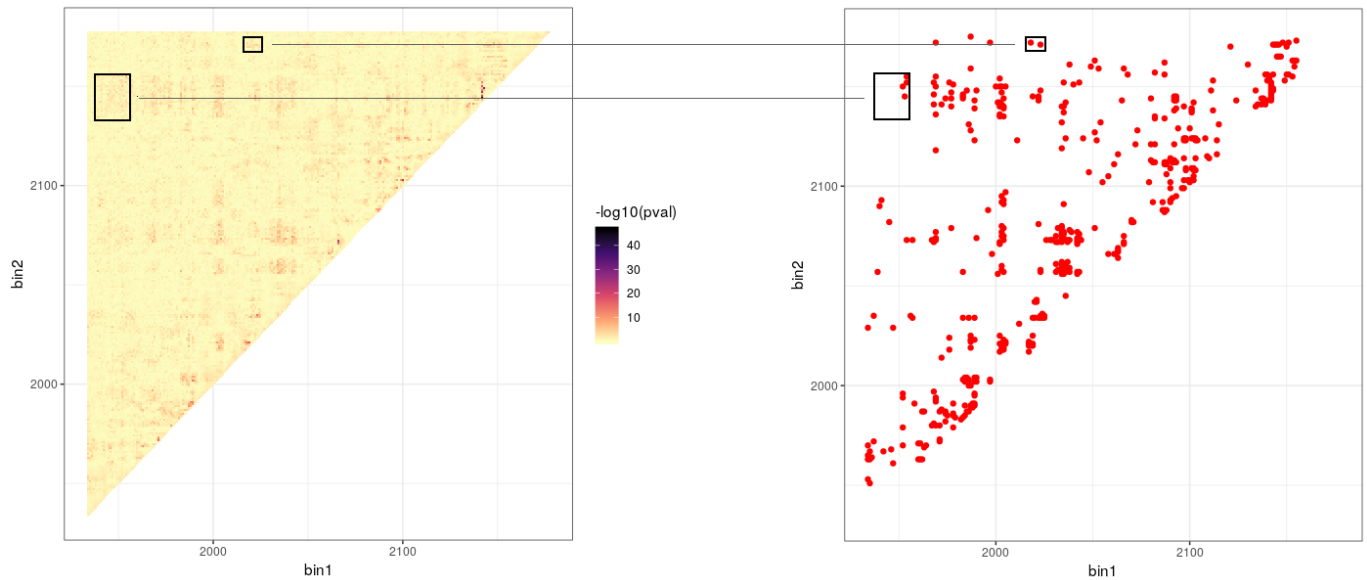
Limits:

- works well only if bin resolution is very high
- unsure that the model is well-suited for Hi-C data

Limits of comparisons at bin pair level

Results:

List of bin pairs (i, j) corresponding to differential interactions between conditions



Limits: These approaches do not account for:

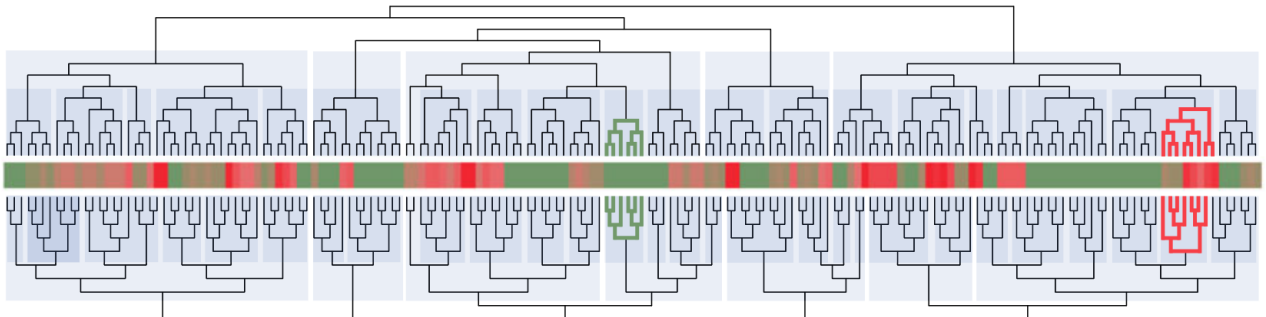
- Dependency between bin pairs
- Hierarchical structure of Hi-C data

$\Rightarrow$ Lack of interpretability in terms of structural differences

[Fraser et al., 2015]'s alternative

[Fraser et al., 2015] developed an approach based on tree structures which account for structural differences:
$\rightarrow$ cannot use replicate ($\mathcal{C}_1 = \{1\}$ and $\mathcal{C}_2 = \{2\}$)

1. For each Hi-C matrix, H^1 and H^2, obtain a clustering of the genome (e.g. TAD clustering)
2. Find common clusters between the two obtained clusterings
3. Apply a hierarchical clustering on those common clusters using the mean of interaction counts as a similarity measure:
   $\rightarrow$ Result : Tree of common clusters spatial organization for each sample
4. A score based on the comparison of path distances within the trees is associated to each cluster (Local Tree Changes measure) and Z-score are computed



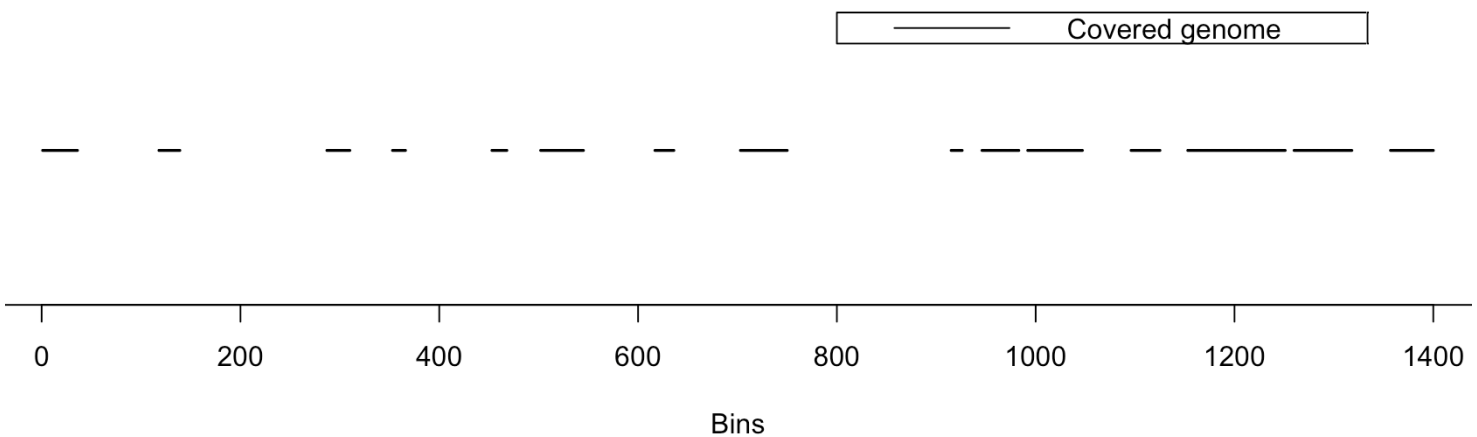
Limits of [Fraser et al., 2015]'s alternative

Results:

List of clusters of bins with differential reciprocal structural organization between conditions

Limits:

- does not account for intra-condition variability (no replicates)
- common structures typically represent a narrow part of the genome:
  $\rightarrow$ Differences probably also lie in regions that are rejected by this approach



Overcoming some of those limits ?

In order to overcome some previously listed limits, a method should be able to:

- perform **structural comparisons**
- use replicates in order to take into account **intra-condition variability**

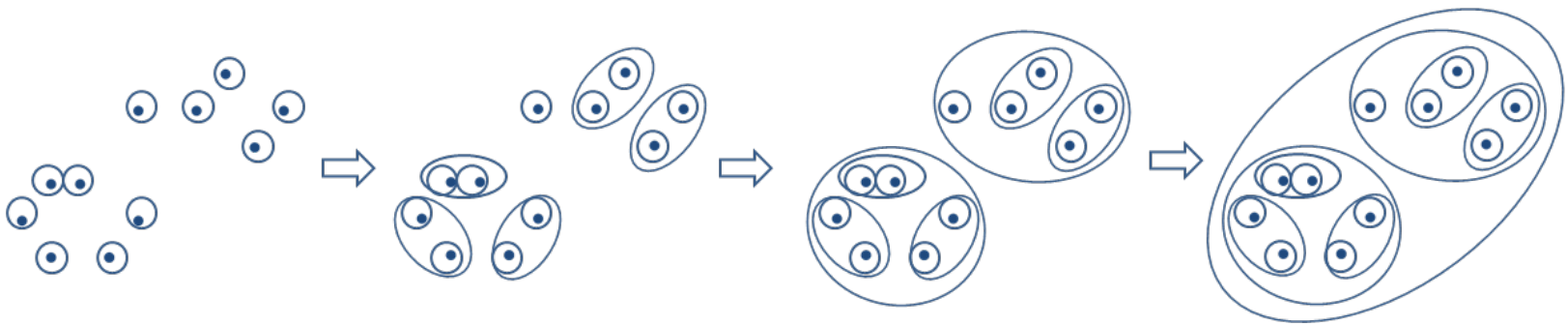
$\rightarrow$ The method proposed in the sequel is also based the **comparisons of tree structures** and can use **replicates**

Differential Analysis method based on CCHAC

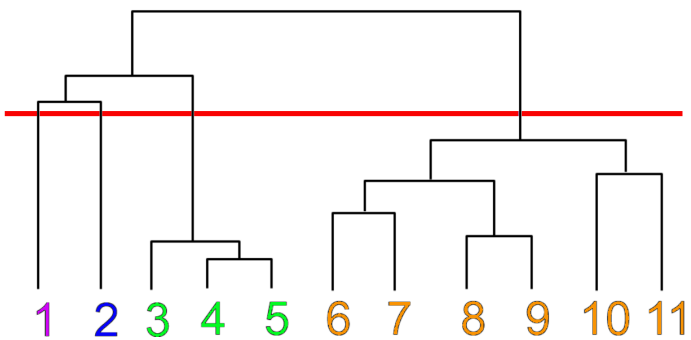
Hierarchical Agglomerative Clustering (HAC)

A multiscale approach to study hierarchical structure:

Initialisation: **For** $t = 1, \ldots, n$ **:** **End:**



Graphical representation of HAC results: $\rightarrow$ **Dendrograms**



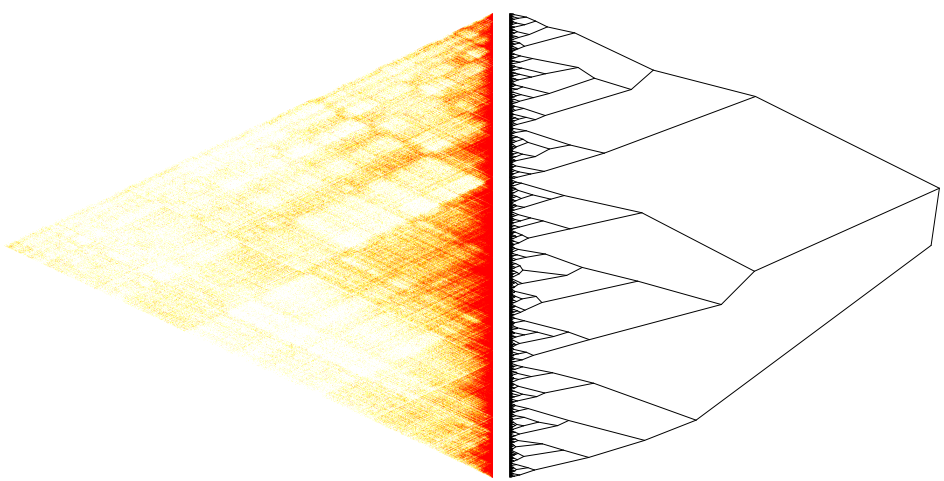
Hi-C and CCHAC

Hi-C data are **3D-proximity measure** $\leftrightarrow$ **similarity** data
$\Rightarrow$ Statistically founded possibility to use HAC on Hi-C matrices
[Randriamihamison et al., 2019]

Contiguity **C**onstrained **H**ierarchical **A**gglomerative **C**lustering:
$\rightarrow$ only adjacent bins can be merged

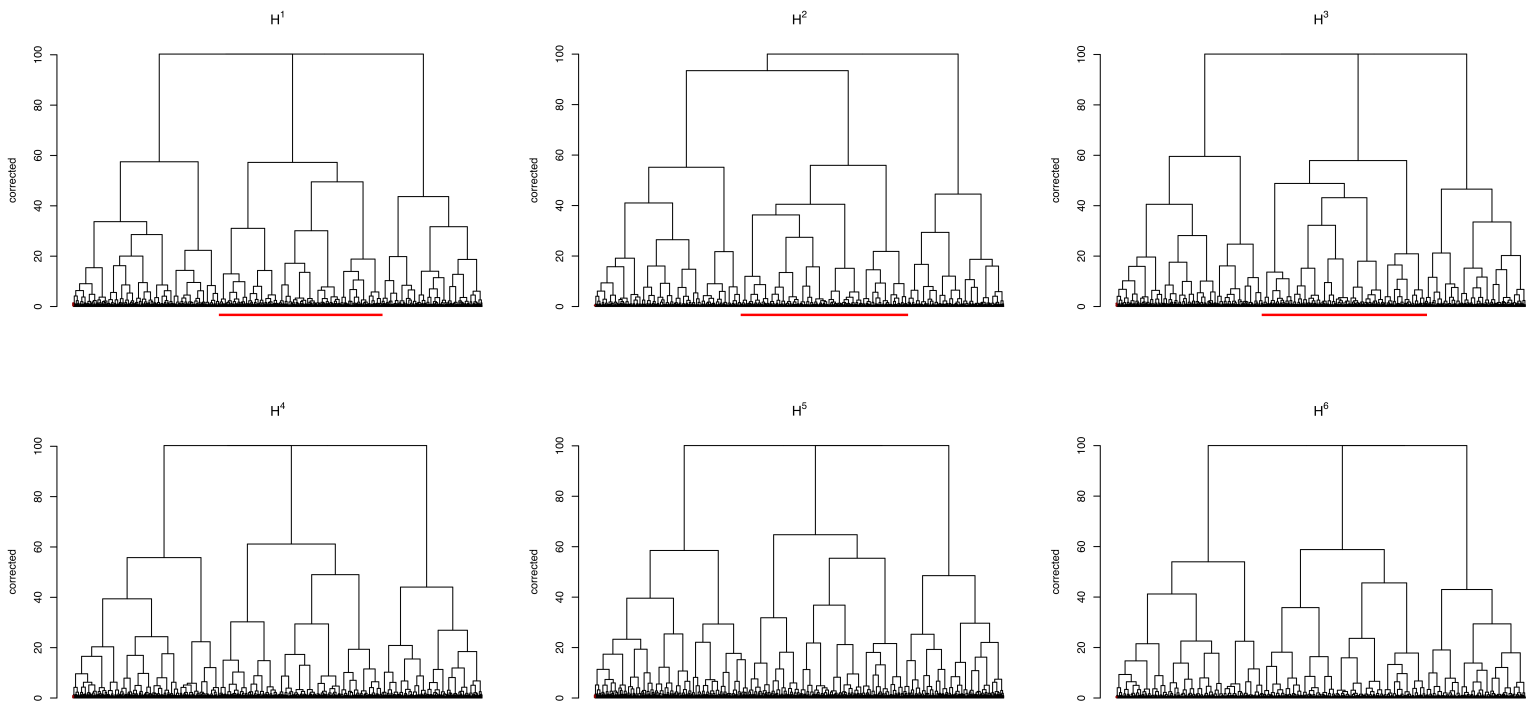
Implementation: R package adjclust

Using CCHAC on Hi-C matrices produces binary trees:



Overview of the method

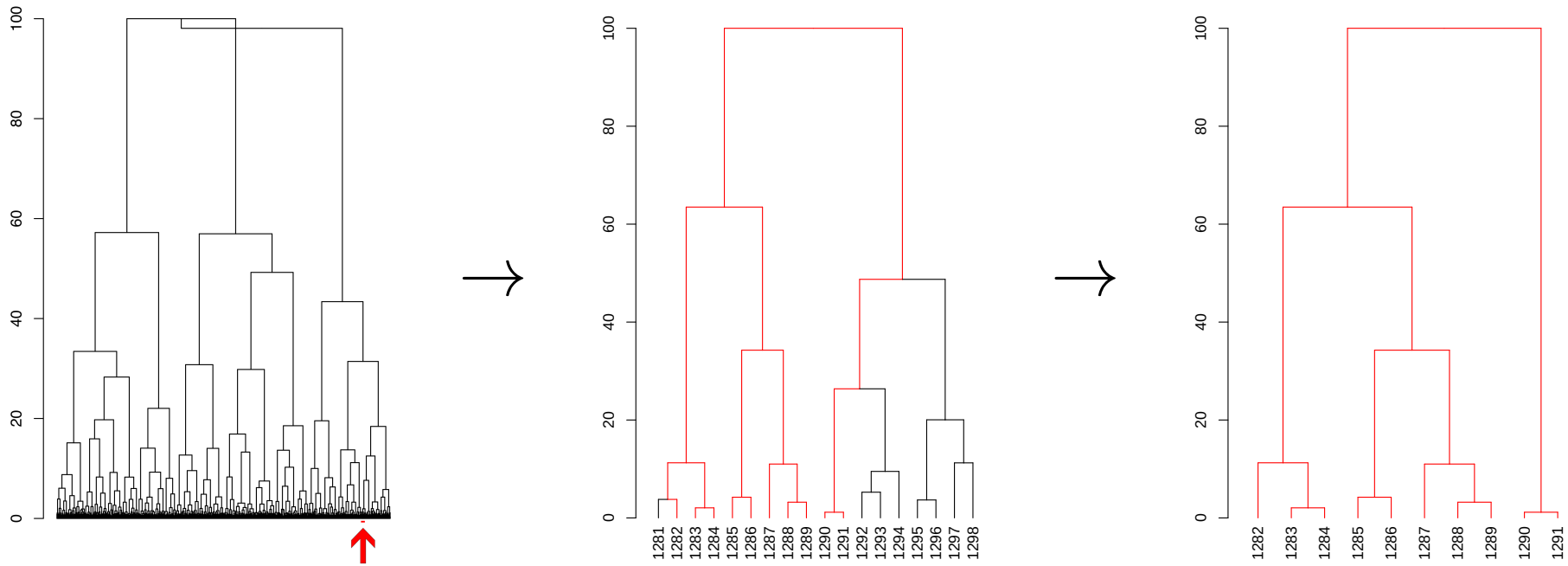
1. For each Hi-C Matrix, obtain a dendrogram using CCHAC
2. For each dendrogram and for each genomic region under study (e.g. all genomic intervals of a fixed bin size), consider the associated induced subtrees
3. Using distances between induced subtrees, compute a statistic to compare biological conditions on the genomic region



Defining induced subtrees

Given a dendrogram and a genomic interval, we can define an **induced subtree**:

→ Example for genomic interval $[1282, 1291]$:



→ Result: a set of 6 induced subtrees (one for each sample) defined on the same genomic interval

Comparing induced subtrees

Comparison of 6 corresponding induced subtrees (defined on the same genomic interval) $\Rightarrow$ Need for a **tree distance**

A lot of possible tree distances:

- R package ape
- R package distory

Simulation $\rightarrow$ **Weighted Path Difference Metric (WPD)**

Practical case (2×3 samples):
For each genomic interval, we obtain:

- 6 intra-conditions distances
- 9 inter-conditions distances

Defining a statistic [work in progress]

A solution might be to consider a statistic such as:

$$W_l := \frac{\bar{d}_l^{inter} - \bar{d}_l^{intra}}{\sigma_{d_l}}$$

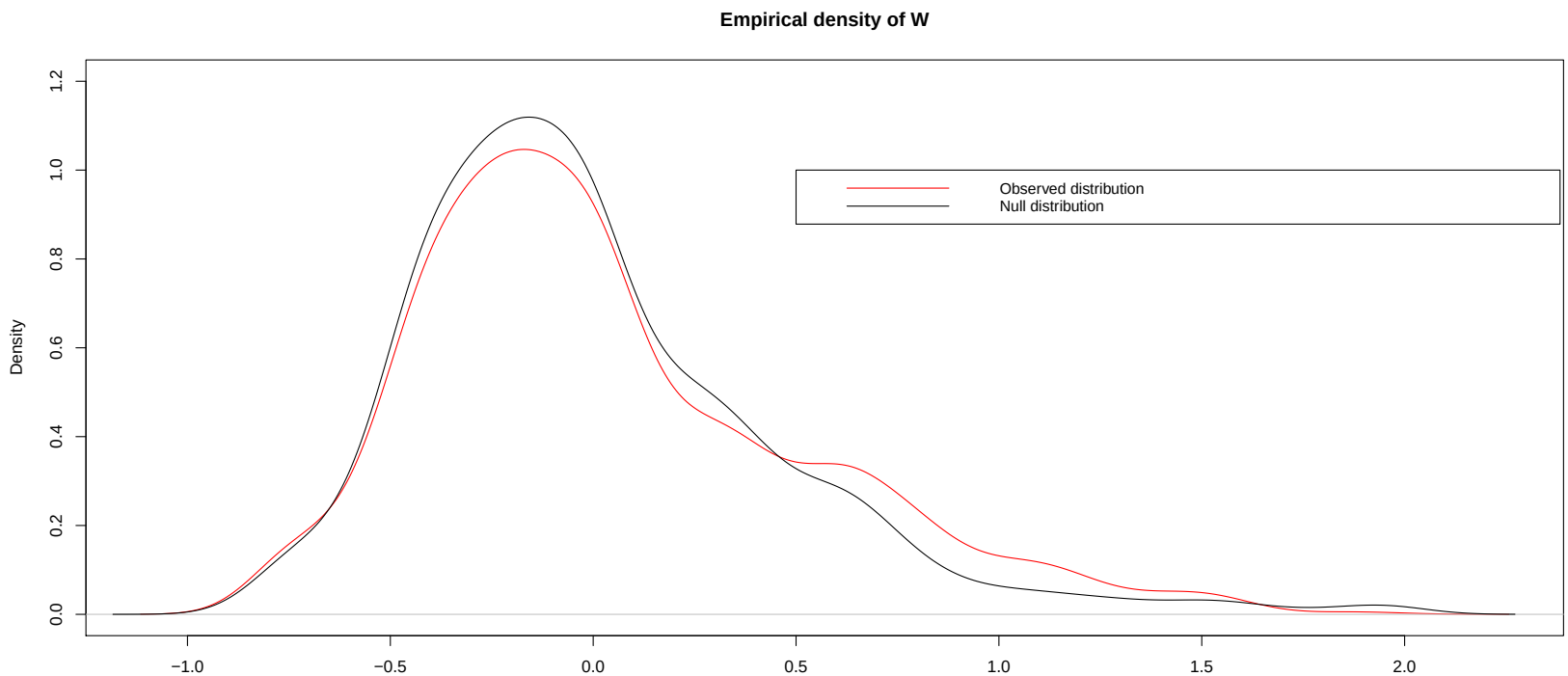
where

- $\bar{d}_l^{inter}$ is the mean of d_l entries corresponding to inter-conditions distances
- $\bar{d}_l^{intra}$ is the mean of d_l entries corresponding to intra-conditions distances
- σ_{d_l} is the standard deviation of d_l entries

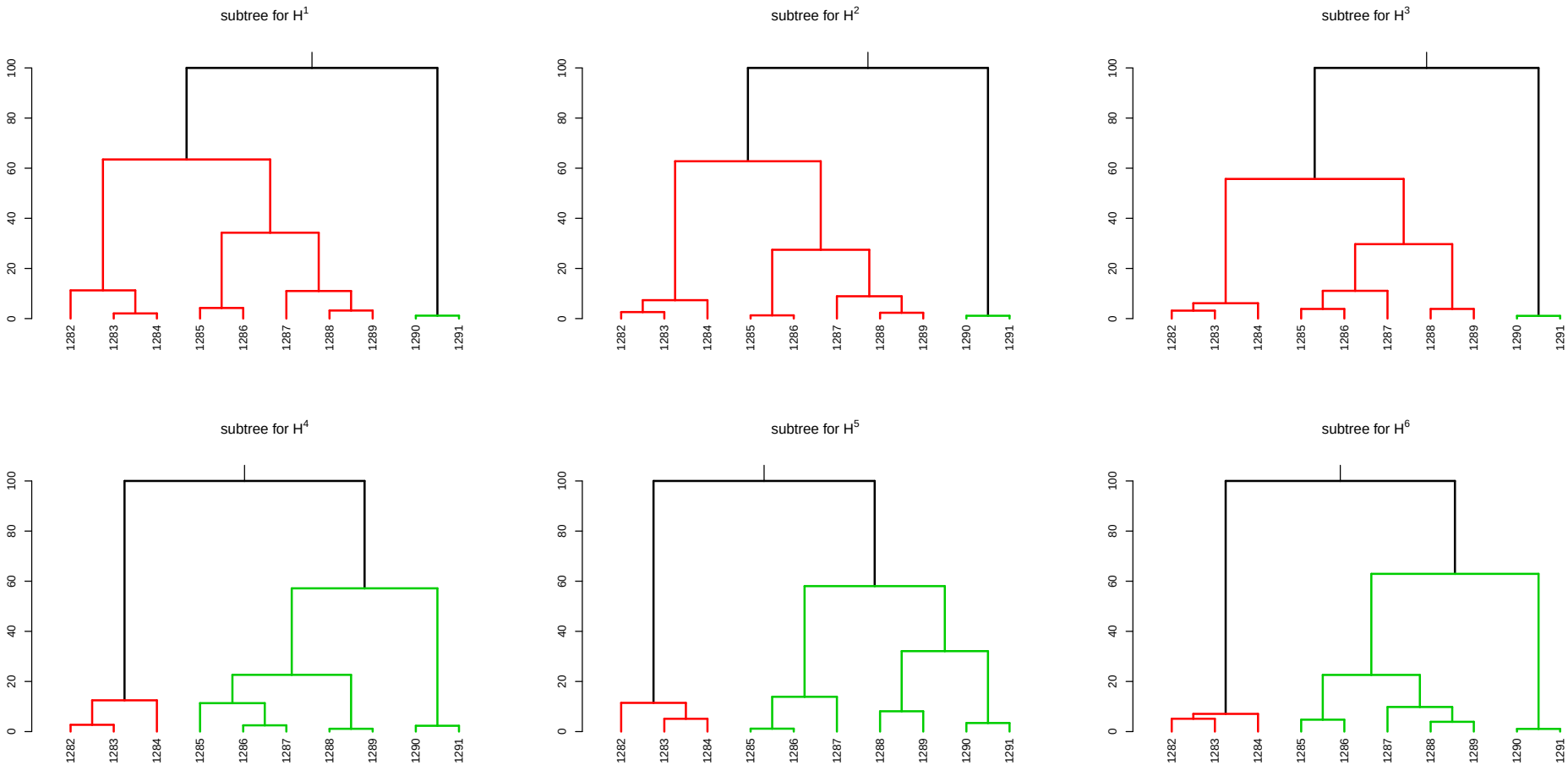
Empirical distribution of W

Setting:

- data from fetal pig development ($\mathcal{C}^1 = \{1, 2, 3\}$, $\mathcal{C}^2 = \{4, 5, 6\}$)
- bin resolution: 40 kb
- chromosome 18
- genomic intervals defined by sizes: 10 bins, 20 bins



Example of a "differential structure"



Conclusion

- What we wanted: a method that would allow to:
  - structurally interpret differences
  - use replicates

- The answer: Differential Analysis based on CCHAC [work in progress]:
  - based on tree representation of Hi-C data obtained via CCHAC
  - focus on genomic intervals in order to allow local comparisons
  - select genomic intervals over which the 3D-structure of genome is differential

Further investigations:

- How to choose a relevant set of genomic intervals for the analysis ?
- Alternative choice of the test statistic (percentage of explained inertia ?)
- Extension of the study to whole genome

Thank you for your attention!

Djekidel, M. N., Chen, Y., and Zhang, M. Q. (2018).

FIND: difFerential chromatin INteractions detection using a spatial poisson process.

Genome Research, 28(3):412–422.

Fraser, J., Ferrai, C., Chiariello, A. M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B. L., Kraemer, D. C., Aitken, S., Xie, S. Q., Morris, K. J., Itoh, M., Kawaji, H., Jaeger, I., Hayashizaki, Y., Carninci, P., Forrest, A. R., The FANTOM Consortium, Semple, C. A., Dostie, J., Pombo, A., and Nicodemi, M. (2015).

Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation.

Molecular Systems Biology, 11:852.

Lun, A. T. and Smyth, G. K. (2015).

diffHic: a bioconductor package to detect differential genomic interactions in hi-c data.

BMC Bioinformatics, 16(1).

Randriamihamison, N., Vialaneix, N., and Neuvial, P. (2019).

Applicability and interpretability of hierarchical agglomerative clustering with or without contiguity constraints.

arXiv preprint arXiv:1909.10923v1.

Stansfield, J. C., Cresswell, K. G., Vladimirov, V. I., and Dozmorov, M. G. (2018).

HiCcompare: an r-package for joint normalization and comparison of HI-c datasets.

BMC Bioinformatics, 19(1).

Empirical density of W for biological conditions defined as different cell lines: