Comparative analysis of 3D genomic data

Nathalie Vialaneix

nathalie.vialaneix@inrae.fr
http://www.nathalievialaneix.eu

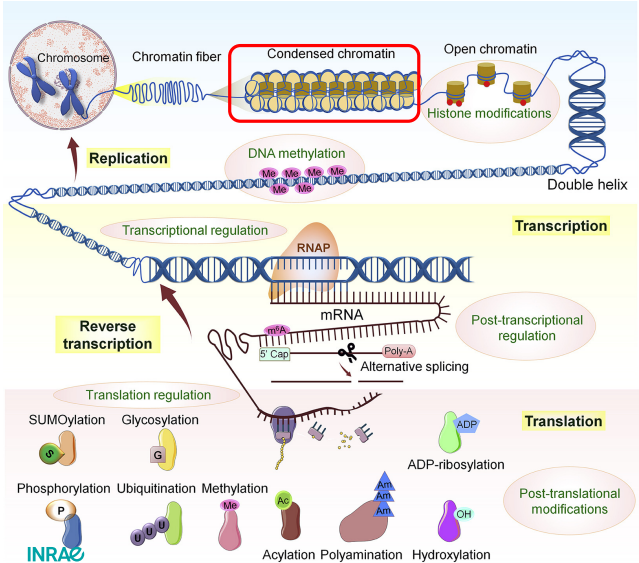
SMPGD 2026: Statistical Methods for Post Genomic Data
January 30th, 2026

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

INRAE

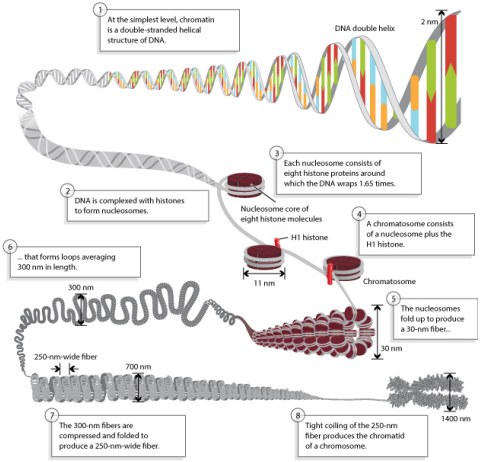
Outline

3D organization of DNA as a regulatory layer



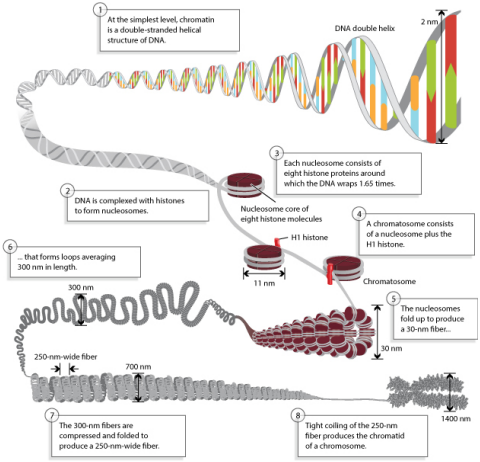
[Pramanik et al., 2021]

DNA organization

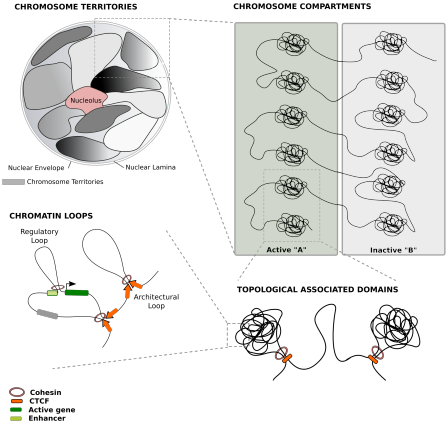


[Annunziato, 2008] 2 meters compressed in 6 μm

DNA organization

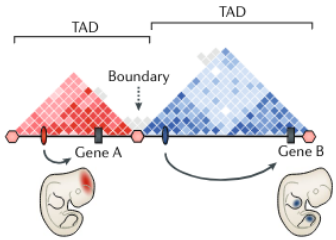


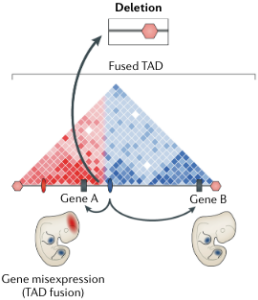
[**Annunziato, 2008**] 2 meters compressed in 6 μm

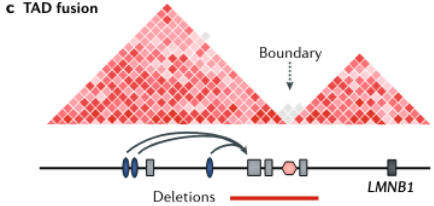


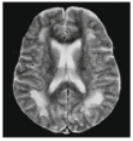
[**Servant, 2017**] adapted from
[**Bonev and Cavalli, 2016**]

Impact on DNA functionning [Spielmann et al., 2018]



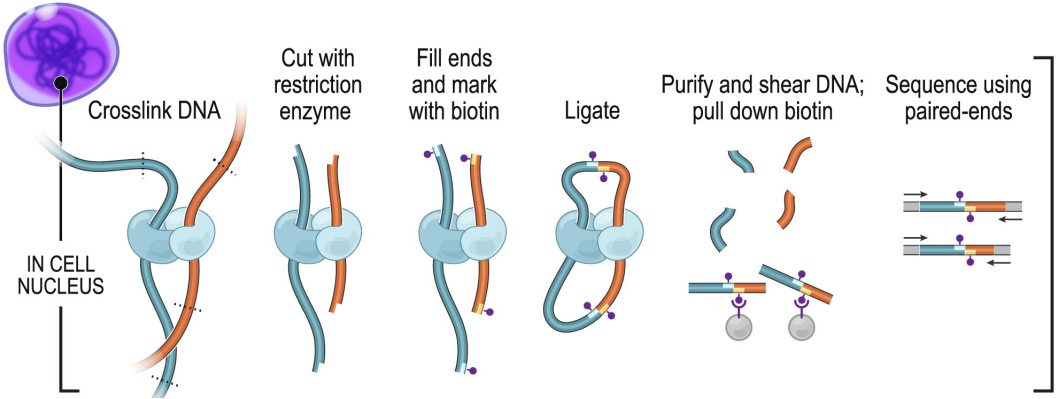




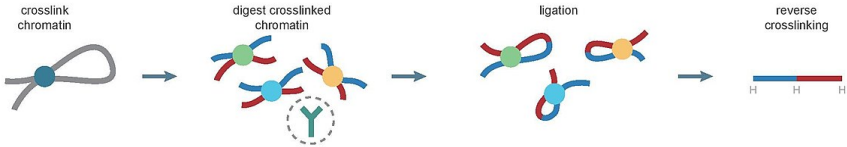


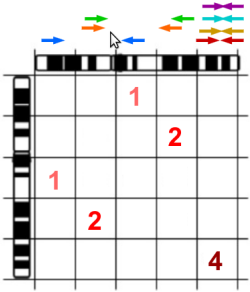
Adult-onset demyelinating leukodystrophy

Hi-C experiment



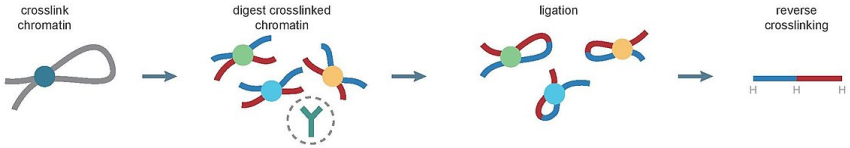
Hi-C data

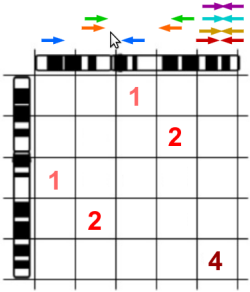


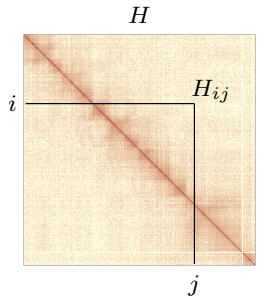


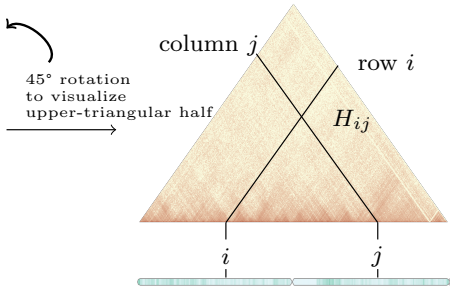
Images: adapted from **[Li et al., 2014]** (top) and courtesy of Sylvain Foissac + own work (bottom).

Hi-C data









Images: adapted from **[Li et al., 2014]** (top) and courtesy of Sylvain Foissac + own work (bottom).

Outline

Where Statistics comes into play

Data: $(H_{ij})_{ij}$, symmetric matrix where entries are counts

H_{ij}: proxy of the spatial proximity between genomic position i and j (call "bins"; (i, j) is a "bin pair" or a "pixel")

Where Statistics comes into play

Series of works on:

Data: $(H_{ij})_{ij}$, symmetric matrix where entries are counts

H_{ij}: proxy of the spatial proximity between genomic position i and j (call "bins"; (i, j) is a "bin pair" or a "pixel")

Hi-C data clustering







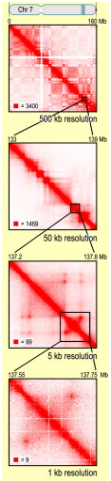
INRAE Inria cnrs

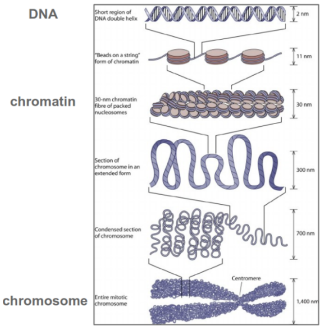
Statistical tests for Hi-C data

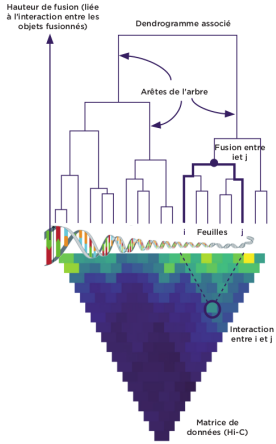



INRAE

1. Hi-C matrices and hierarchical clustering [Neuvial et al., 2023]







Idea used in: [**Fraser et al., 2015**, **Soler-Vila et al., 2020**, **Zhang et al., 2021b**] among others

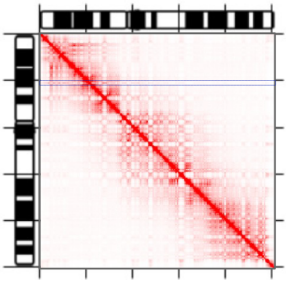
Clustering based on Hi-C matrices

Summary of the work performed in: Projet CNRS SCALES (with P. Neuvial & S. Foissac). Thèse INRAE/Inria Nathanaël Randriamihamison (also with M. Chavent)

Clustering based on Hi-C matrices

Summary of the work performed in: Projet CNRS SCALES (with P. Neuvial & S. Foissac). Thèse INRAE/Inria Nathanaël Randriamihamison (also with M. Chavent)

How is hierarchical clustering usually performed on Hi-C matrices?



1. a bin =
2. Euclidean distance + HC

Clustering based on Hi-C matrices

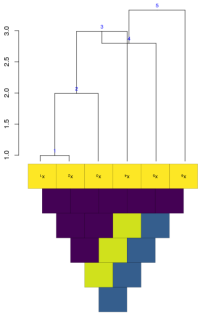
Summary of the work performed in: Projet CNRS SCALES (with P. Neuvial & S. Foissac). Thèse INRAE/Inria Nathanaël Randriamihamison (also with M. Chavent)

Why is it not satisfactory?

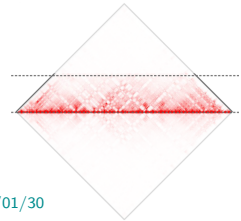
1. Hi-C matrices are already distances!
2. sparse vs full representations

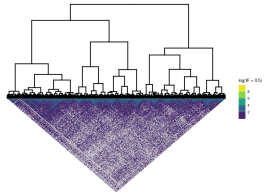
Clustering based on Hi-C matrices

Contributions

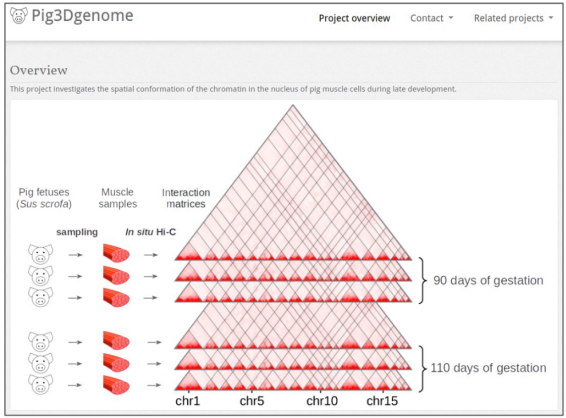


1. extend Ward's HC to similarity matrices as Hi-C with adjacency constraint **[Randriamihamison et al., 2021]**
2. study of the mathematical properties **[Randriamihamison et al., 2021]**
3. study of reversals in practice on Hi-C matrices **[Randriamihamison et al., 2021]**
4. fast version using a band computation **[Ambroise et al., 2019]**, implemented in the R package **adjclust**





2. Motivation example for tests in Hi-C data: Pig3D genome project



Marti-Marimon et al, 2021

Pig3Dgenome

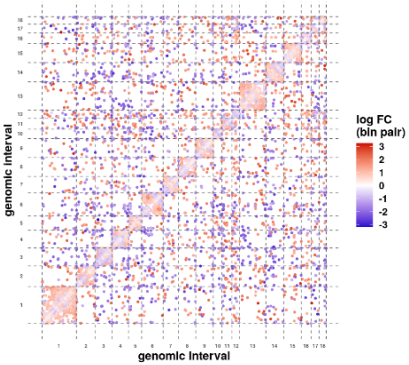
Differences between conditions?

[Marti-Marimon et al., 2018] (courtesy of Sylvain Foissac)

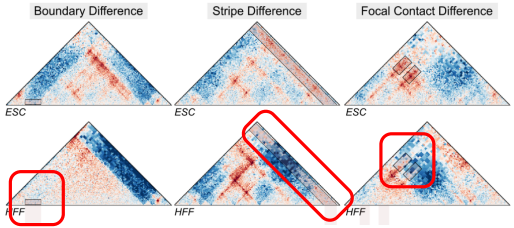
Outline

Standard methods for differential analysis of Hi-C data

Pixel-based methods



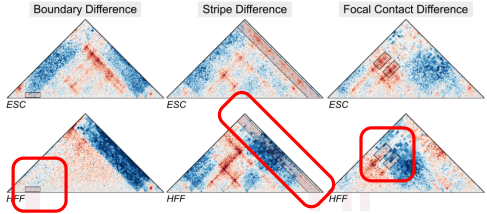
Structure-based methods



[Gjoni et al., 2025]

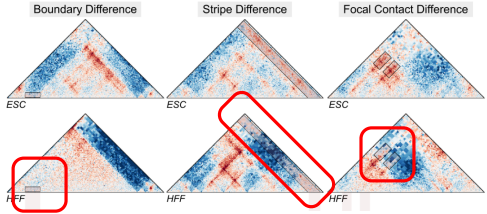
Structure-based methods

- target a specific structure



Structure-based methods

- target a specific structure



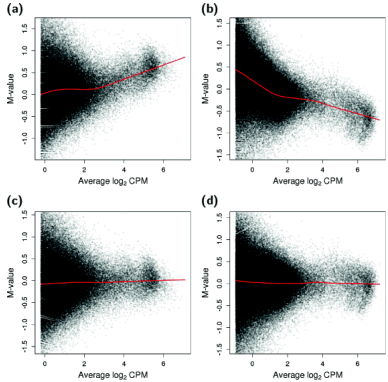
- limited to this specific structure and sensitive to differences in methods for structure-discovery **[Forcato et al., 2017]**

Pixel-based methods [Jorge et al., 2025]

	Design		Preprocessing		Model				p-values	
Tool	Accept $K > 2$ conditions	Usage of covariates	Filtering	Normalization	Assumption on data distribution	Usage of replicates	Structure-aware	Type of method	raw p-values provided	p-value correction
ACCOST	✘	✘	bad quality	ICE	NB	✔	✘	exact test	✔	✘
CHESS	✘	✘	bad quality	ICE	✘	✘	✔	Z-score	✔	✘
diffHic	✔	✔	low counts	MA	NB	✔	✘	GLM	✔	BH
FIND	✘	✘	✘	KR	spatial Poisson process	✔	✔	exact test	✘	BH/distance
HiCcompare	✘	✘	user-defined	MD	✘	✘	✘	Z-score	✔	BH/distance
HiCDCPlus	✘	✘	bad quality	ICE	NB	✔	✘	exact test	✔	BH
HOMER	✘	✘	✘	ICE	✘	✘	✘	Z-score	?	?
multiHiCcompare	✔	✔	low counts	MD	NB	✔	✘	GLM	✔	BH/distance
Selfish	✘	✘	low counts	KR	✘	✘	✔	Z-score	✘	BH
sslHiC	✘	✘	large distance	min/max	✘	✘	✔	GNN	✘	BH

Most accurate method: **diffHic** [Lun and Smyth, 2015]

1. normalize data with MA trend correction by cyclic LOESS



Most accurate method: **diffHic** [Lun and Smyth, 2015]

1. normalize data with MA trend correction by cyclic LOESS

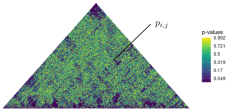
2. fit a BN GLM: $H_{i,j}^{\ell} \sim \mathrm{BN}(\lambda_k, \phi_{i,j,k})$, $k := \mathrm{Condition}(\ell)$, $\phi_{i,j,k}$: dispersion parameter ($+$ moderated tests)

Most accurate method: **diffHic** [Lun and Smyth, 2015]

1. normalize data with MA trend correction by cyclic LOESS

2. fit a BN GLM: $H^{\ell}_{i,j} \sim \mathrm{BN}(\lambda_k, \phi_{i,j,k})$, $k := \mathrm{Condition}(\ell)$, $\phi_{i,j,k}$: dispersion parameter ($+$ moderated tests)

$\Rightarrow \mathbf{P} = (p_{ij})_{i,j=1,\ldots,p,\, i \leq j}$



Limitations...

- poor interpretability
- spatial structure of the matrix is neither used

Limitations...

- poor interpretability
- spatial structure of the matrix is neither used

First approach (**diffHic**): "clustering" of contiguous significant p-values

But:

- no statistical guarantees
- badly handles false negative in positive regions

Limitations...

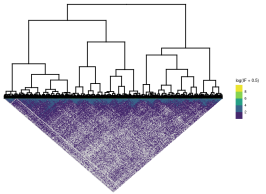
- poor interpretability
- spatial structure of the matrix is neither used

First approach (**diffHic**): "clustering" of contiguous significant *p*-values

But:

- no statistical guarantees
- badly handles false negative in positive regions

Second approach **[Neuvial et al., 2024]**



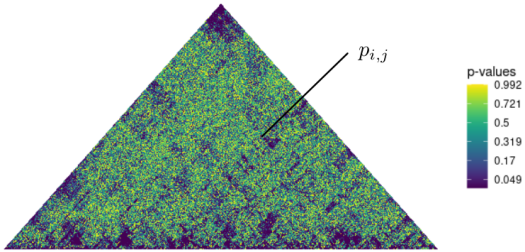
But:

- only partially exploits the spatial structure
- limited to TAD-like structures

Outline

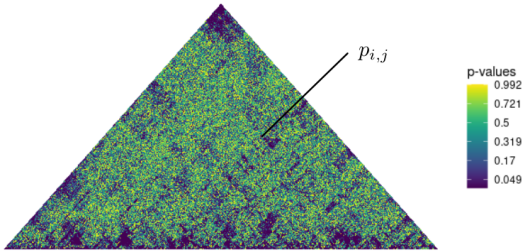
General objective

What do we start from?



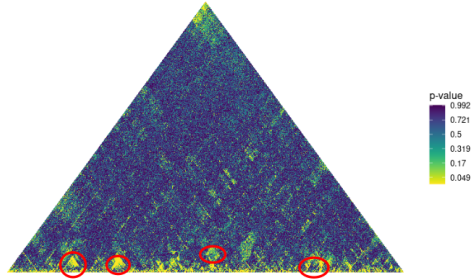
General objective

What do we start from?



What do we want to achieve?

1. provide statistical guarantees for any (arbitrary) cluster



2. build clusters in a data-driven manner

Why post-hoc inference?

$\mathcal{H} = \{(i,j) : \ i,j = 1,\dots,p, \ i \leq j\}$: hypotheses to test
$\mathcal{H}_0 \subset \mathcal{H}$: true null hypotheses

Multiple testing correction:

- $\mathcal{R}$: rejected hypotheses
- False positives: $\mathcal{R} \cap \mathcal{H}_0$
- False Discovery Rate: $\mathbb{E}\big(\frac{|\mathcal{R} \cap \mathcal{H}_0|}{|\mathcal{R}| \vee 1}\big)$.

Why post-hoc inference?

$\mathcal{H} = \{(i,j) : \ i,j = 1,\dots,p, \ i \leq j\}$: hypotheses to test
$\mathcal{H}_0 \subset \mathcal{H}$: true null hypotheses

Multiple testing correction:

- $\mathcal{R}$: rejected hypotheses
- False positives: $\mathcal{R} \cap \mathcal{H}_0$
- False Discovery Rate: $\mathbb{E}\big(\frac{|\mathcal{R} \cap \mathcal{H}_0|}{|\mathcal{R}| \vee 1}\big)$.
- But: **Global** FDR control $\not\Longrightarrow$ FDR control in **subset of pixels**

Why post-hoc inference?

$\mathcal{H} = \{(i,j) : \ i,j = 1,\dots,p, \ i \leq j\}$: hypotheses to test
$\mathcal{H}_0 \subset \mathcal{H}$: true null hypotheses

Multiple testing correction:

- $\mathcal{R}$: rejected hypotheses
- False positives: $\mathcal{R} \cap \mathcal{H}_0$
- False Discovery Rate: $\mathbb{E}\left(\frac{|\mathcal{R} \cap \mathcal{H}_0|}{|\mathcal{R}| \vee 1}\right)$.
- But: **Global** FDR control $\not\Longrightarrow$ FDR control in **subset of pixels**

Post-hoc inference

[Goeman and Solari, 2011]:

- For $S \subset \mathcal{H}$, quantify $\mathrm{TDP}(S) = 1 - \frac{|S \cap \mathcal{H}_0|}{|S|}$?
- post-hoc bound $V_\alpha : S \subset \mathcal{H} \to \mathbb{R}$ st:

Why post-hoc inference?

$\mathcal{H} = \{(i,j) : \ i,j = 1,\dots,p, \ i \leq j\}$: hypotheses to test
$\mathcal{H}_0 \subset \mathcal{H}$: true null hypotheses

Multiple testing correction:

- $\mathcal{R}$: rejected hypotheses
- False positives: $\mathcal{R} \cap \mathcal{H}_0$
- False Discovery Rate: $\mathbb{E}\left(\frac{|\mathcal{R} \cap \mathcal{H}_0|}{|\mathcal{R}| \vee 1}\right)$.
- But: **Global** FDR control $\not\Longrightarrow$ FDR control in **subset of pixels**

Post-hoc inference

[Goeman and Solari, 2011]:

- For $S \subset \mathcal{H}$, quantify $\mathrm{TDP}(S) = 1 - \frac{|S \cap \mathcal{H}_0|}{|S|}$?
- post-hoc bound $V_\alpha : S \subset \mathcal{H} \to \mathbb{R}$ st:

$$\mathbb{P}(\forall S \subset \mathcal{H}, |S \cap \mathcal{H}_0| \leq V_\alpha(S)) \geq 1 - \alpha$$

Why post-hoc inference?

$\mathcal{H} = \{(i,j) : \ i,j = 1,\dots,p, \ i \leq j\}$: hypotheses to test
$\mathcal{H}_0 \subset \mathcal{H}$: true null hypotheses

Multiple testing correction:

- $\mathcal{R}$: rejected hypotheses
- False positives: $\mathcal{R} \cap \mathcal{H}_0$
- False Discovery Rate: $\mathbb{E}\left(\frac{|\mathcal{R} \cap \mathcal{H}_0|}{|\mathcal{R}| \vee 1}\right)$.
- But: **Global** FDR control $\not\Longrightarrow$ FDR control in **subset of pixels**

Post-hoc inference
[Goeman and Solari, 2011]:

- For $S \subset \mathcal{H}$, quantify $\mathrm{TDP}(S) = 1 - \frac{|S \cap \mathcal{H}_0|}{|S|}$?
- post-hoc bound $V_\alpha : S \subset \mathcal{H} \to \mathbb{R}$ st:

$$\mathbb{P}(\forall S \subset \mathcal{H}, \mathrm{TDP}(S) \geq \gamma_\alpha(S)) \geq 1 - \alpha$$

with $\gamma_\alpha(S) = 1 - \frac{V_\alpha(S)}{|S|}$.

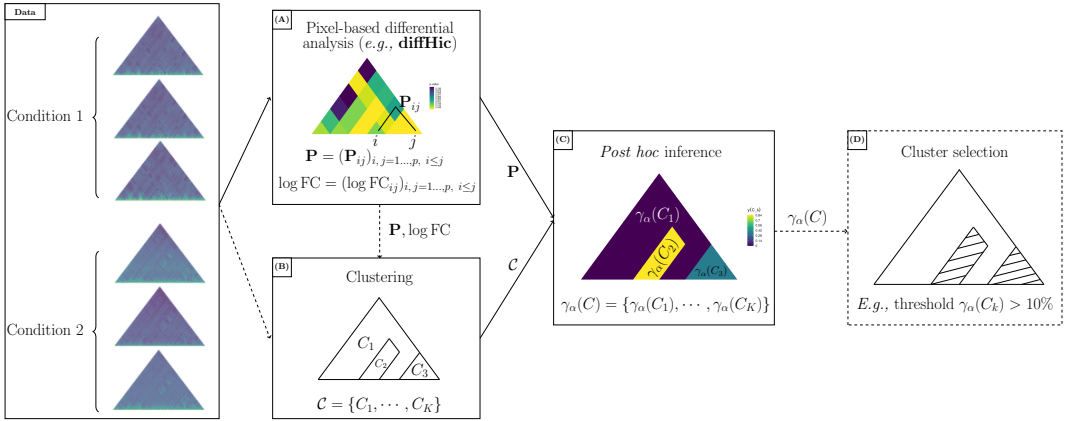
A very general post-hoc bound

Under independence or PRDS **[Benjamini and Yekutieli, 2001]**, Simes' post hoc bound:

$$V_\alpha^{\text{Simes}}(S) = \min_{1 \le \ell \le |S|} \left[\sum_{(i,j) \in S} \mathbb{1}_{\{\mathbf{P}_{ij} > \frac{\alpha \ell}{m}\}} + \ell - 1 \right]$$

is valid **[Goeman and Solari, 2011]**.

In short, proposed procedure



Data driven clustering

Desirable properties of clusters:

- contiguous 2D regions in Hi-C map
- related to bin pair test results
- also accounts for direction of the change (p-value not sufficient)

Data driven clustering

2D constrained hierarchical clustering

[Lebart, 1978, Grimm, 1987, Gordon, 1996, Thirion et al., 2014]

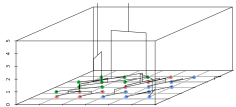
1. **Neighboring constraint:**

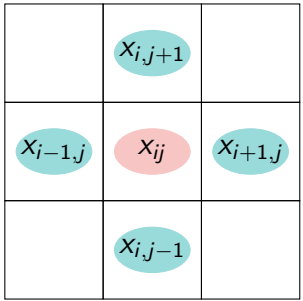
$$(i,j) \text{ and } (i',j') \text{ neighbors } \iff |i-i'|+|j-j'|=1$$

2. **Agglomerative clustering** based on
   - **counts**:

$$x_{ij} := (H_{ij}^{C_1,r_1}, H_{ij}^{C_1,r_2}, H_{ij}^{C_1,r_3}, H_{ij}^{C_2,r_1}, H_{ij}^{C_2,r_2}, H_{ij}^{C_2,r_3})$$

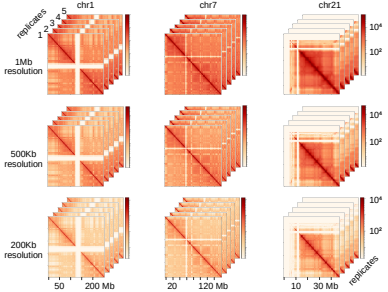
   - **diff**: $x_{ij} = (\log \mathrm{FC}_{ij}, \mathbf{P}_{ij})$

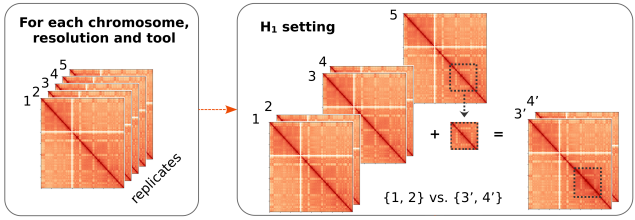




Simulation study

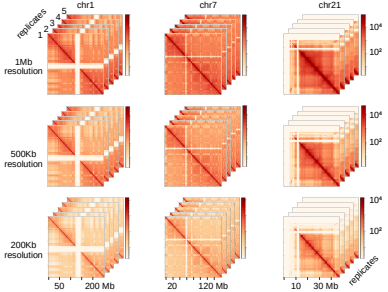
Semi-simulated data from ENCODE data:

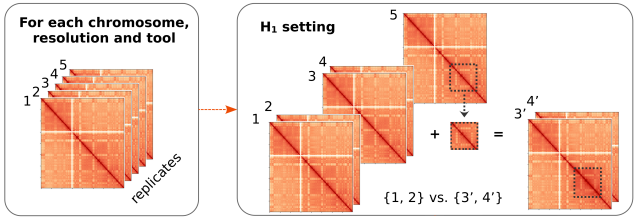




Simulation study

Semi-simulated data from ENCODE data:

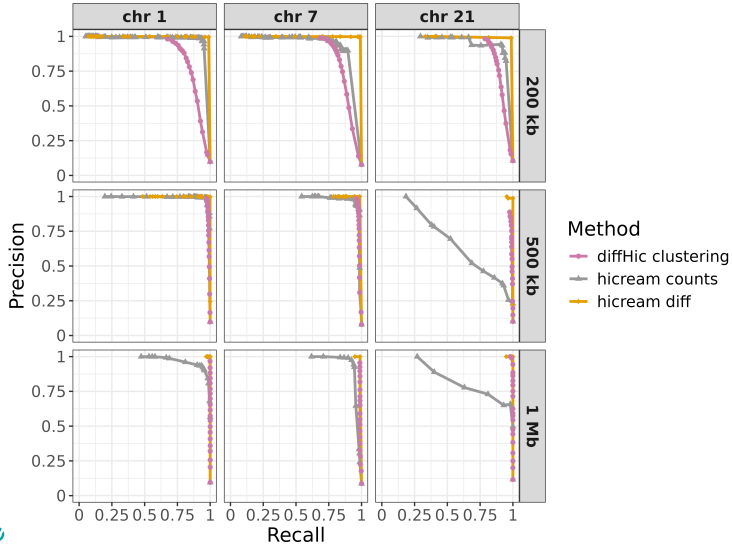




Tested methods:

- our : constrained 2D clustering + $\gamma_\alpha(S)$ cutoff
- **diffHic** clustering based on adjusted p-values + cluster-level FDR cutoff

Results: PR curve



Results: numerical performance of "best" results

		diffHic clustering					hicream diff				
Chr.	Res.	Cluster-level FDR	Nb. of clusters	Median cluster size	Precision	Recall	TDP	Nb. of clusters	Median cluster size	Precision	Recall
1	200 kb	0.40	14198	3	0.879	0.802	0.014	120	36	0.996	0.989
1	500 kb	0.05	3037	4	0.988	0.971	0.086	72	70	0.998	0.999
1	1 Mb	0.05	2634	1	0.967	0.999	0.2	76	10.5	1	1
7	200 kb	0.45	3994	4	0.904	0.802	0.005	76	33	0.997	0.991
7	500 kb	0.05	1131	4	0.987	0.962	0.258	153	10	1	0.986
7	1 Mb	0.05	1109	1	0.955	0.987	0.698	42	11	1	0.988
21	200 kb	0.25	314	5	0.939	0.833	0.291	21	34	0.989	0.988
21	500 kb	0.05	151	2	0.888	0.975	0.083	19	10	0.988	1
21	1 Mb	0.10	100	1	0.99	0.99	0.2	7	12	1	1

Results: numerical performance of "best" results

		diffHic clustering					hicream diff				
Chr.	Res.	Cluster-level FDR	Nb. of clusters	Median cluster size	Precision	Recall	TDP	Nb. of clusters	Median cluster size	Precision	Recall
1	200 kb	0.40	14198	3	0.879	0.802	0.014	120	36	0.996	0.989
1	500 kb	0.05	3037	4	0.988	0.971	0.086	72	70	0.998	0.999
1	1 Mb	0.05	2634	1	0.967	0.999	0.2	76	10.5	1	1
7	200 kb	0.45	3994	4	0.904	0.802	0.005	76	33	0.997	0.991
7	500 kb	0.05	1131	4	0.987	0.962	0.258	153	10	1	0.986
7	1 Mb	0.05	1109	1	0.955	0.987	0.698	42	11	1	0.988
21	200 kb	0.25	314	5	0.939	0.833	0.291	21	34	0.989	0.988
21	500 kb	0.05	151	2	0.888	0.975	0.083	19	10	0.988	1
21	1 Mb	0.10	100	1	0.99	0.99	0.2	7	12	1	1

Real dataset with external biological knowledge

Data from: **[Zhang et al., 2021a]** (CTCF depletion study)

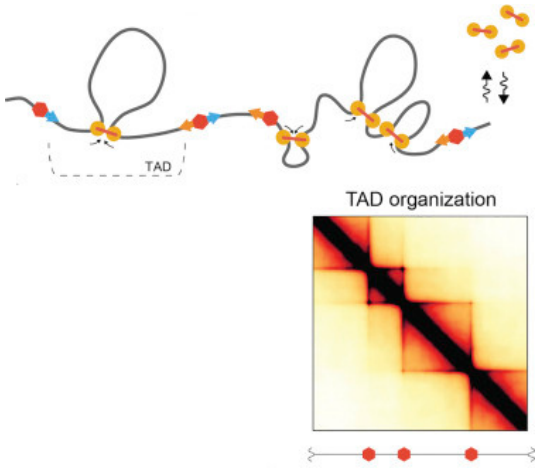
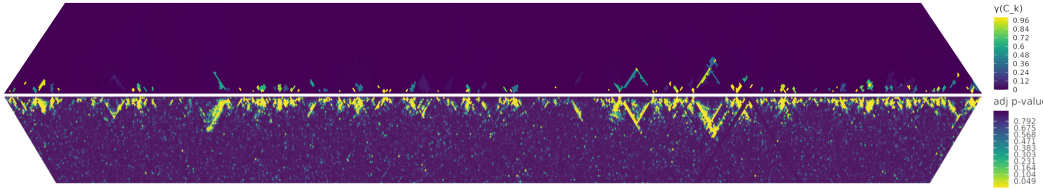


image adapted from: **[Fudenberg et al., 2016]**

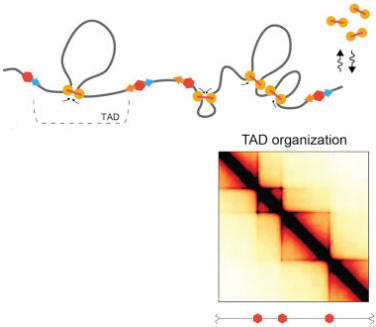
Smoothing property of **hicream**

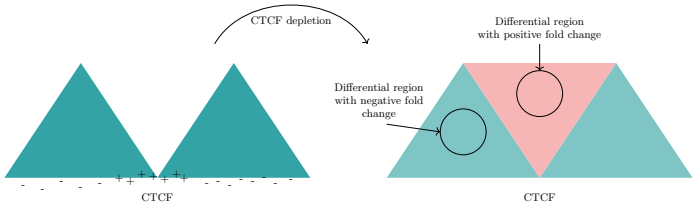
Minimal True Discovery Proportions from *post hoc* bounds



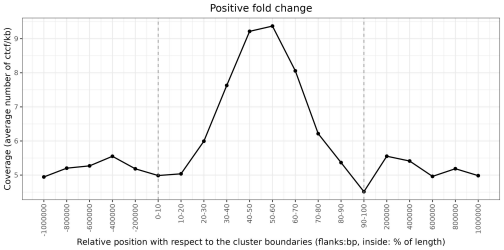
diffHic differential analysis adjusted p-values

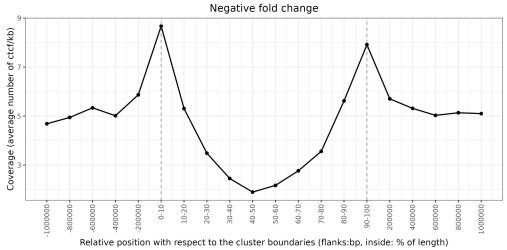
What to expect from CTCF depletion?

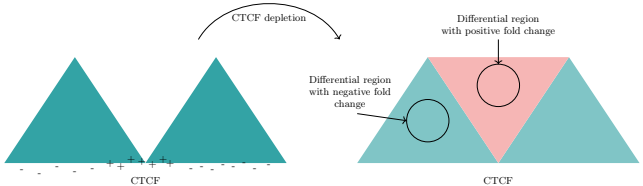




Results based on identified clusters







Conclusion and perspectives

Take home messages

- Hi-C data are challenging (hierarchical, spatial and counts)
- Post-hoc bound for genomics give interesting exploratory results with quantifiable statistical evidence (best of two worlds)

Conclusion and perspectives

Take home messages

- Hi-C data are challenging (hierarchical, spatial and counts)
- Post-hoc bound for genomics give interesting exploratory results with quantifiable statistical evidence (best of two worlds)

What's next?

- Define tighter bounds taking advantage of the hierarchical structure
- Perform clustering and pixel-based differential analysis together?

Conclusion and perspectives

Take home messages

- Hi-C data are challenging (hierarchical, spatial and counts)
- Post-hoc bound for genomics give interesting exploratory results with quantifiable statistical evidence (best of two worlds)

What's next?

- Define tighter bounds taking advantage of the hierarchical structure
- Perform clustering and pixel-based differential analysis together?

Thank you for your attention!
Questions?

References

Ambroise, C., Dehman, A., Neuvial, P., Rigaill, G., and Vialaneix, N. (2019).
Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics.
Algorithms for Molecular Biology, 14:22.

Annunziato, A. T. (2008).
DNA packaging: nucleosome and chromatin.
Nature Education, 1(1):26.

Benjamini, Y. and Yekutieli, D. (2001).
The control of the false discovery rate in multiple testing under dependency.
Annals of Statistics, 29(4):1165–1188.

Bonev, B. and Cavalli, G. (2016).
Organization and function of the 3D genome.
Nature Reviews Genetics, 17(11):661–678.

Forcato, M., Nicoletti, C., Pal, K., Livi, C., Ferrari, F., and Bicciato, S. (2017).
Comparison of computational methods for Hi-C data analysis.
Nature Methods, 14(7):679–685.

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.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L. (2016).
Formation of chromosomal domains by loop extrusion.

Cell Reports, 15(9):2038–2049.

Gjoni, K., Gunsalus, L. M., Shuzhen, K., McArthur, E., Pittman, M., Capra, J. A., and Pollard, K. S. (2025).
Comparing chromatin contact maps at scale: methods and insights.
Nature Methods, 22:824–833.

Goeman, J. J. and Solari, A. (2011).
Multiple testing for exploratory research.
Statistical Science, 26(4):584–597.

Gordon, A. (1996).
A survey of constrained classification.
Computational Statistics & Data Analysis, 21(1):17–29.

Grimm, E. C. (1987).
CONISS: a FORTRAN 77 program for stratigraphically constrained analysis by the method of incremental sum of squares.
Computers & Geosciences, 13(1):13–35.

Jorge, E., Foissac, S., Neuvial, P., Zytnicki, M., and Vialaneix, N. (2025).
A comprehensive review and benchmark of differential analysis tools for Hi-C data.
Briefings in Bioinformatics, 26(2):bbaf074.

Lebart, L. (1978).
Programme d'agrégation avec contraintes.
Les Cahiers de l'Analyse des Données, 3(3):275–287.

Li, G., Cai, L., Chang, H., Hong, P., Zhou, Q., Kulakova, E. V., Kolchanov, N. A., and Ruan, Y. (2014).
Chromatin interaction analysis with paired-end tag (ChIA-PET) sequencing technology and application.
BMC Genomics, 15(Suppl 12):S11.

Lun, A. T. and Smyth, G. K. (2015).
diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data.
BMC Bioinformatics, 16:258.

Marti-Marimon, M., Vialaneix, N., Voillet, V., Yerle-Bouissou, M., Lahbib-Mansais, Y., and Liaubet, L. (2018).
A new approach of gene co-expression network inference reveals significant biological processes involved in porcine muscle development in late gestation.
Scientific Report, 8:10150.

Neuvial, P., Foissac, S., and Vialaneix, N. (2023).
Comprendre l'organisation spatiale de l'ADN à l'aide de la statistique.
In Knoop, M., Blanc, S., and Bouzeghoub, M., editors, *L'Interdisciplinarité. Voyage au-delà des Disciplines*, pages 172–176. CNRS.

Neuvial, P., Randriamihamison, N., Chavent, M., Foissac, S., and Vialaneix, N. (2024).
A two-sample tree-based test for hierarchically organized genomic signals.
Journal of the Royal Statistical Society, Series C, 73(3):774–795.

Pramanik, D., Shelake, R. M., Kim, M. J., and Kim, J.-Y. (2021).
CRISPR-mediated engineering across the central dogma in plant biology for basic research and crop improvement.
Molecular Plant, 14(1):127–150.

Randriamihamison, N., Vialaneix, N., and Neuvial, P. (2021).
Applicability and interpretability of Ward's hierarchical agglomerative clustering with or without contiguity constraints.
Journal of Classification, 38:363–389.

Servant, N. (2017).
Analyse de données de conformation chromosomique et application au cancer.
Phd thesis, Université Paris 6.

Soler-Vila, P., Cuscó, P., Farabella, I., Di Stefano, M., and Marti-Renon, M. A. (2020).

Hierarchical chromatin organization detected by TADpole.
Nucleic Acids Research, 45(7):e39.

Spielmann, M., Lupiáñez, D. G., and Mundlos, S. (2018).
Structural variation in the 3D genome.
Nature Reviews Genetics, 19(7):453–467.

Thirion, B., Varoquaux, G., Dohmatob, E., and Poline, J.-B. (2014).
Which fMRI clustering gives good brain parcellations?
Frontiers in Neuroscience, 8:167.

Zhang, H., Lam, J., Zhang, D., Lan, Y., Vermunt, M. W., Keller, C. A., Giardine, B., Hardison, R. C., and Blobel, G. A. (2021a).
CTCF and transcription influence chromatin structure re-configuration after mitosis.
Nature Communications, 12:5157.

Zhang, Y. W., Wang, M. B., and Li, S. C. (2021b).
SuperTAD: robust detection of hierarchical topologically associated domains with optimized structural information.
Genome Biology, 22:45.