HICREAM : Méthode d'analyse différentielle de données Hi-C par inférence post hoc et visualisation interactive

Elise Jorge[1,2], Toby Dylan Hocking[3], Pierre Neuvial[4], Nathalie Vialaneix[2] & Sylvain Foissac[1]

[1] *GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet-Tolosan, France,* {*elise.jorge, sylvain.foissac*}*@inrae.fr*

[2] *Université Fédérale de Toulouse, INRAE, MIAT, 31326 Castanet-Tolosan, France, nathalie.vialaneix@inrae.fr*

[3] *LASSO lab, Département d'informatique Université de Sherbrooke, Sherbrooke, QC, J1K-2R1, Canada, toby.dylan.hocking@usherbrooke.ca*

[4] *Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, CNRS, UPS, F-31062 Toulouse Cedex 9, France, pierre.neuvial@math.univ-toulouse.fr*

Résumé. La conformation tridimensionnelle du génome a un rôle important dans de nombreux processus biologiques. L'expérience Hi-C, largement utilisée pour caractériser l'organisation 3D du génome, estime les proximités spatiales entre positions génomiques par un nombre d'interactions. L'analyse différentielle de données Hi-C permet alors l'identification de régions génomiques réorganisées entre deux conditions biologiques d'intérêt. Ici, nous présentons **hicream**, un cadre d'analyse exploratoire permettant d'obtenir des régions génomiques différentielles de forme arbitraire et fournissant des garanties statistiques sur ces régions grâce à l'inférence post hoc. Nous proposons également une visualisation dynamique des résultats permettant leur évaluation et leur exploration. La méthode **hicream** est implémentée en R et disponible sur le CRAN https://CRAN.R-project.org/package=hicream.

Mots-clés. données Hi-C, génomique 3D, inférence post hoc, test statistique

Abstract. The three-dimensional conformation of the genome plays an important role in many biological processes. The Hi-C experiment is widely used to characterize the 3D organization of the genome by estimating spatial proximities between genomic positions from interaction frequencies. Differential analysis of Hi-C data aims at identifying reorganized genomic regions by comparing two biological conditions of interest. Here, we present **hicream**, an exploratory analysis framework that allows to obtain differential genomic regions of arbitrary shape and provides statistical guarantees on these regions using post hoc inference. We also provide a dynamic visualization of the results, outlining differential regions and enabling their exploration. **hicream** is implemented in R and available as a package on CRAN https://CRAN.R-project.org/package=hicream.

Keywords. Hi-C data, 3D genomics, post hoc inference, statistical testing

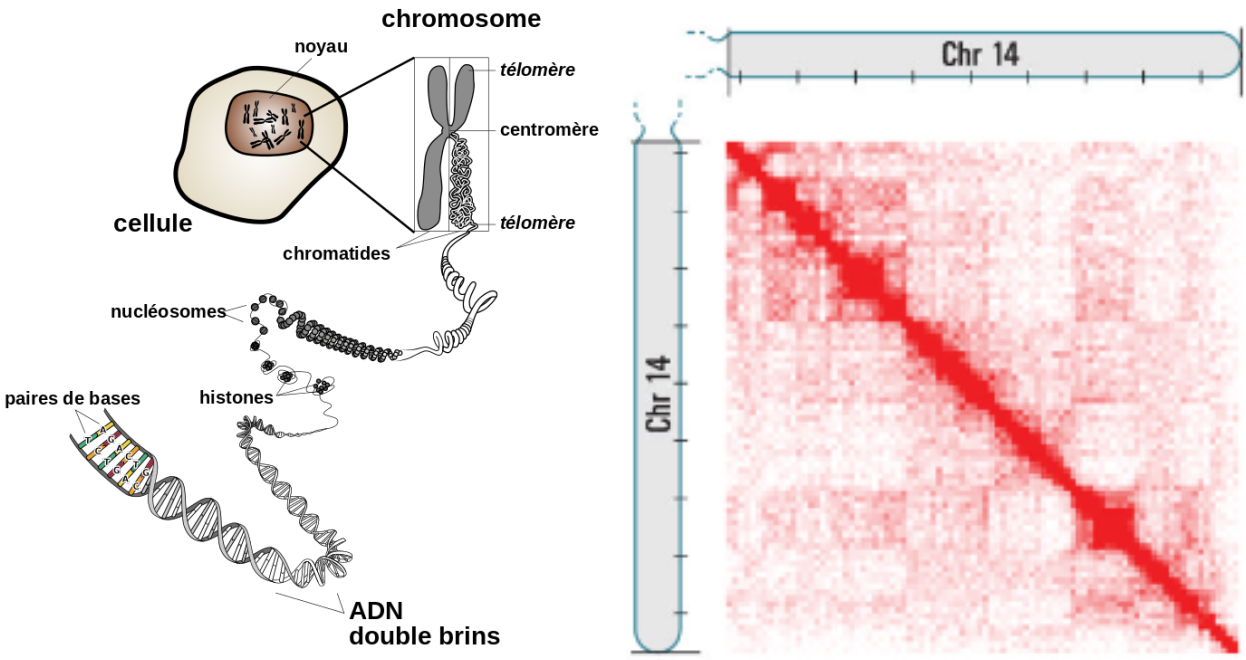

FIGURE 1 – Gauche : Schéma de la compaction de l'ADN en chromosome (« Chromosome fr » par Phrood commonswiki, Wikimedia Commons). Droite : Matrice Hi-C du chromosome 14 de [1].

1 Introduction

Conformation tridimensionelle du génome. Comme observé sur la figure 1, des contraintes spatiales au sein du noyau d'une cellule conduisent à la forte compaction de l'ADN. Ainsi, il peut exister une proximité spatiale entre des positions génomiques linéairement éloignées sur le chromosome. Cette conformation tridimensionnelle de l'ADN joue un rôle dans de nombreux processus biologiques tels que la régulation de l'expression des gènes. Des remaniements de cette structure tridimensionnelle peuvent avoir des conséquences importantes comme l'apparition de pathologies neurologiques [2] ou de malformations [3].

Données Hi-C. La méthode Hi-C [1] permet la caractérisation de l'organisation tridimensionelle du génome en estimant des distances entre positions génomiques par une mesure du nombre d'interactions. Ainsi, le résultat d'une expérience Hi-C peut être résumé dans une matrice symétrique, creuse, à entrées positives et discrètes, où sont renseignés le nombre d'interactions enregistrées durant l'expérience pour toutes les paires de positions génomiques considérées.

Analyse différentielle de données Hi-C. L'analyse différentielle de données Hi-C consiste en la comparaison de deux groupes de matrices obtenues dans deux conditions biologiques d'intérêt, $\mathcal{C}_1$ et $\mathcal{C}_2$. L'objectif est d'identifier des zones des matrices qui correspondent à des régions génomiques ayant subi une réorganisation entre les deux conditions.

Formellement, on compare deux groupes de $r = r_1 + r_2$ matrices Hi-C de taille $p \times p$ où $M^{g,l}$ correspond au l-ème réplicat de la « condition » $g \in \{1, 2\}$. Pour chaque pixel (i, j) d'une matrice où $i, j \in \{1, \cdots, p\}$ et $j \leq i$ – par symétrie, on ne regarde que la demie-matrice – on note $m_{ij}^{g,l}$ le nombre d'interactions enregistrées entre les « bins » (positions génomiques) i et j.

L'analyse différentielle des données Hi-C est généralement organisée en deux grands types de méthodes : l'analyse différentielle basée sur les pixels et l'analyse différentielle basée sur les structures.

Le premier type de méthode consiste à tester l'hypothèse $H_0^{(i,j)}$: « *Le nombre moyen d'interactions entre les paires de positions génomiques i et j n'est pas différent entre les deux conditions* » et ainsi obtenir une p-valeur pour chaque pixel (i, j). Ce type de méthode fournit des résultats limités en terme d'interprétation biologique avec des pixels différentiels potentiellement éparpillés dans toute la matrice.

Le second type de méthode consiste à chercher des régions différentielles qui correspondent à un type de structure défini a priori. Dans les données Hi-C, il est par exemple possible d'identifier des structures comme des TADs (*Topologically Associating Domains*), qui sont des régions génomiques au sein desquelles les positions intéragissent fortement entre elles et représentées par des triangles le long de la diagonale de la matrice. En plus de reposer sur des définitions de structures qui ne sont pas consensuelles, ces méthodes limitent l'exploration en contraignant la détection de régions différentielles à des types/formes pré-définis et ne fournissent généralement pas de garanties statistiques.

Afin de parvenir à identifier des régions génomiques différentielles de forme arbitraire, on pourrait vouloir utiliser les résultats de l'analyse différentielle au niveau des pixels et sélectionner les plus significatifs afin de former des sous-ensembles de pixels différentiels. Cependant, en réalisant un grand nombre de tests – ici, on réalise théoriquement jusqu'à $p(p+1)/2$ tests – les p-valeurs doivent être ajustées pour la multiplicité. Pour ce type de méthode, c'est souvent le contrôle global du FDR qui est obtenu en utilisant la correction de Benjamini-Hochberg (BH) [4]. Or, le FDR représentant l'espérance de la proportion globale de faux positifs parmi les hypothèses rejetées, il apparaît que le contrôle global du FDR sur l'ensemble des pixels considérés n'implique par le contrôle de cette quantité sur un sous-ensemble de pixels sélectionnés [5]. Ainsi, il n'est donc pas possible de fournir de garanties sur la présence de faux positifs dans une région correspondant à un sous-ensemble de p-valeurs ajustées seulement en utilisant la méthode BH.

2 Description du cadre proposé par la méthode hicream

Nous présentons ici **hicream**, un cadre d'analyse différentielle de données Hi-C qui s'articule en trois étapes :

1. Analyse différentielle basée sur les pixels, fournissant une p-valeur $\mathbf{P}_{ij}$ pour chaque pixel (i, j).
2. Définition de régions génomiques candidates par une classification des pixels $\mathcal{C} = \{C_1, \cdots, C_K\}$.
3. Inférence post hoc fournissant une borne inférieure sur la proportion de vrais positifs $\gamma_\alpha(C_k)$ pour chaque classe C_k.

Analyse différentielle basée sur les pixels. Dans un benchmark [6] de méthodes d'analyse différentielle au niveau des pixels, nous avons montré que **diffHic** [7] est une des méthodes fournissant les meilleures garanties statistiques en terme de contrôle de l'erreur de type I (contrôle

des faux positifs) et de l'erreur de type II (puissance de détection des pixels différentiels). Dans la suite, nous proposons d'utiliser les p-valeurs fournies par cette méthode.

Définition de régions candidates par classification. Afin d'obtenir des régions génomiques d'intérêt sans avoir d'a priori sur les données, nous proposons d'utiliser une classification ascendante hiérarchique prenant en compte une contrainte de proximité spatiale. Nous définissons $\mathbf{X}$, matrice à m lignes et deux colonnes, où m est le nombre de pixels contenant un comptage non-nul dans au moins un des réplicats. Soit $q := (i, j)$ un pixel, alors $\mathbf{X}_{q,\cdot} = (\log \mathbf{P}^{\text{adj}}_{ij}, \log \text{FC}_{ij})$ où les $\mathbf{P}^{\text{adj}}_{ij}$ sont obtenues après correction de Benjamini-Hochberg des p-valeurs obtenues par **diffHic**.

Nous définissons également $\mathcal{G}$, graphe dont les sommets représentent les m pixels des données et où les arêtes relient les pixels adjacents en deux dimensions sur la carte Hi-C.

Nous utilisons l'implémentation de la classification ascendante hiérarchique de **scikit-learn** avec le critère de Ward [8] sur la matrice $\mathbf{X}$ avec la contrainte induite par $\mathcal{G}$. Le dendrogramme obtenu est coupé à une hauteur h choisie en cherchant un "coude" dans l'allure de l'évolution du critère de Ward en fonction de l'étape de la classification ascendante hiérarchique. Cette étape permet l'obtention d'une unique classification des données en régions candidates : $\mathcal{C} = \{C_1, \cdots, C_K\}$.

Notons que le cadre d'analyse de la méthode **hicream** permet d'évaluer des régions génomiques candidates quelle que soit la manière dont elles ont été obtenues.

Inférence post hoc : définition et utilisation. Notre objectif est ici de quantifier le signal présent dans chacune des classes de pixels de $\mathcal{C} = \{C_1, \cdots, C_K\}$, indépendamment du choix de ces classes. Plus formellement, après avoir obtenu une p-valeur $\mathbf{P}_{ij}$ par pixel (i, j) avec **diffHic**, nous cherchons à quantifier pour chaque classe C_k le $\text{TDP}(C_k) = 1 - |C_k \cap \mathcal{H}_0|/|C_k|$, c'est-à-dire, la proportion de vrais positifs dans C_k. Pour cela, nous nous appuyons sur les méthodes post hoc [5] qui fournissent une garantie sur le nombre de faux positifs dans une classe C_k, $|C_k \cap \mathcal{H}_0|$, indépendamment du choix de cette classe.

Formellement, on appelle *borne post hoc* [5] une fonction V_α telle que :

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

Si (1) est vérifiée, alors pour tout C_k la quantité $\gamma_\alpha(C_k) = 1 - V_\alpha(C_k)/|C_k|$ permet de minorer $\text{TDP}(C_k)$:

$$\mathbb{P}\left(\forall C_k, \text{TDP}(C_k) \geq 1 - \frac{V_\alpha(C_k)}{C_k}\right) \geq 1 - \alpha\,. \tag{2}$$

Comme $\gamma_\alpha(C_k)$ minore la proportion de vrais positifs dans une classe d'intérêt C_k, indépendamment du choix de C_k, cette mesure peut être utilisée pour comparer des classes de p-valeurs arbitrairement sélectionnées.

Si l'on a m hypothèses nulles testées, on définit la borne de Simes :

$$V_\alpha(C_k) = \min_{1 \leq n \leq |C_k|} \left[\sum_{(i,j) \in C_k} \mathbb{1}_{\left\{\mathbf{P}_{ij} > \frac{\alpha n}{m}\right\}} + n - 1 \right].$$

Sous des hypothèses d'indépendance des p-valeurs ou de dépendance positive (PRDS) [9], on peut montrer [10] que la borne V^{Simes} satisfait l'équation (1) (et est donc une borne post hoc). L'hypothèse PRDS est considérée comme réaliste pour les applications génomiques [11]. En particulier, c'est sous cette hypothèse que le contrôle du FDR par la procédure BH est valable. Nous proposons donc d'utiliser la borne V_α^{Simes} dans le cadre de l'analyse différentielle de données Hi-C afin de calculer, pour toute classe C_k, la borne associée $\gamma_\alpha(C_k)$.

3 Application et visualisation : changement de conformation durant la différentiation cellulaire

Nous avons implémenté la méthode décrite dans la section précédente dans le package R **hicream** [12].

Les outils de visualisation de données Hi-C ne sont pas bien adaptés à la visualisation des résultats d'**hicream** car ils ne permettent pas de représenter des régions génomiques correspondant à des classes de forme arbitraire. Nous proposons une visualisation dynamique produite grâce au package R **animint2** [13, 14] qui permet de représenter chaque classe détourée et colorée selon sa proportion minimale de vrais positifs $\gamma_\alpha(\cdot)$.

Nous avons testé l'approche **hicream** sur des données Hi-C issues de lignées cellulaires murines [15] pour deux conditions biologiques correspondant à des stades de différenciation cellulaire différents de cellules neuronales (ES : cellules souches embryonnaires et CN : neurones corticaux). Sur la figure 2, nous représentons les résultats obtenus pour le chromosome 18. Cette visualisation permet d'observer des zones précises de la matrice, sélectionnées sur la partie *Genomic interaction summary*. Le zoom sur la région affichée dans *Genomic interaction zoom* permet d'observer les classes détourées en vert et, pour chacune, la borne inférieure $\gamma_\alpha(\cdot)$ calculée sur le TDP. On peut ainsi observer des classes possédant un TDP élevé, proche de 1, indiquant une région différentielle identifiée par la méthode.

4 Perspectives

Le cadre d'analyse différentielle de données Hi-C proposé dans **hicream** présente l'avantage d'être entièrement *data-driven* et de fournir des régions différentielles de forme arbitraire sur lesquelles nous possédons des garanties statistiques, permettant une approche exploratoire de ce type d'analyse. La pertinence des régions génomiques différentielles identifiées par cette méthode a pu être validée grâce à des données biologiques. Par ailleurs, des travaux en cours visent à appliquer la méthode **hicream** à l'étude comparative de la conformation tridimensionnelle de génomes d'animaux d'élevage. Dans la suite, il sera possible d'envisager d'obtenir des bornes post hoc plus précises en prenant en compte la dépendance spatiale entre pixels.

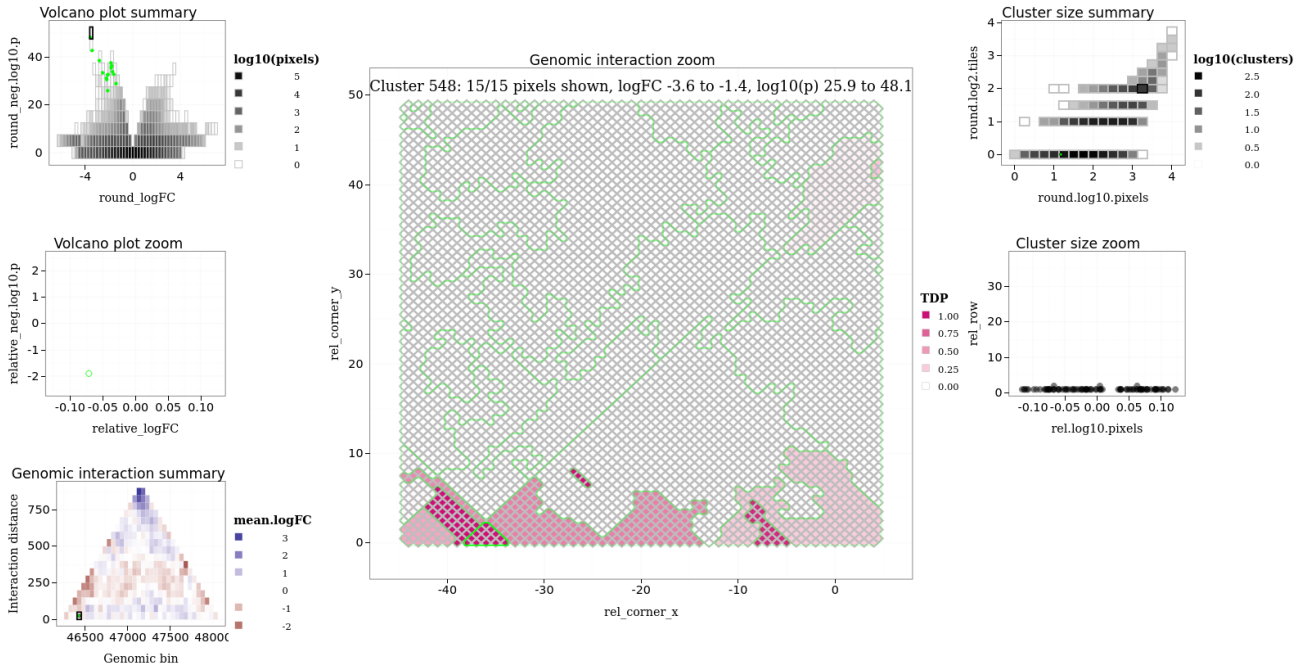

FIGURE 2 – Extrait de la représentation interactive des résultats de la méthode **hicream** pour le chromosome 18 des données de différentiation neuronale. Les différents cadres interagissent entre eux. La sélection d'une sous-région de la matrice dans *Genomic interaction summary* permet d'afficher un zoom sur *Genomic interaction zoom* où sont représentées les classes entourées en vert et la borne $\gamma_\alpha(\cdot)$ sur le TDP calculée pour chacune d'entre elles.

Bibliographie

[1] Erez Lieberman-Aiden et al. "Comprehensive mapping of long-range interactions reveals folding principles of the human genome". In: *Science* 326.5950 (2009), pp. 289–293. DOI: 10.1126/science.1181369.

[2] Malte Spielmann, Darío G Lupiáñez, and Stefan Mundlos. "Structural variation in the 3D genome". In: *Nature Reviews Genetics* 19.7 (2018), pp. 453–467. DOI: 10.1038/s41576-018-0007-0.

[3] Darío G Lupiáñez et al. "Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions". In: *Cell* 161.5 (2015), pp. 1012–1025. DOI: 10.1016/j.cell.2015.04.004.

[4] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

[5] Jelle J Goeman and Aldo Solari. "Multiple testing for exploratory research". In: *Statistical Science* 26.4 (2011), pp. 584–597. DOI: 10.1214/11-STS356.

[6] Elise Jorge et al. "A comprehensive review and benchmark of differential analysis tools for Hi-C data". In: *Briefings in Bioinformatics* 26.2 (2025), bbaf074. DOI: 10.1093/bib/bbaf074.

[7] Aaron T.L. Lun and Gordon K. Smyth. "diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data". In: *BMC Bioinformatics* 16 (2015), p. 258. DOI: 10.1186/s12859-015-0683-0.

[8] Joe H Ward Jr. "Hierarchical grouping to optimize an objective function". In: *Journal of the American statistical association* 58.301 (1963), pp. 236–244.

[9] Yoav Benjamini and Daniel Yekutieli. "The control of the false discovery rate in multiple testing under dependency". In: *Annals of Statistics* 29.4 (2001), pp. 1165–1188. DOI: 10.1214/aos/1013699998.

[10] Gilles Blanchard, Pierre Neuvial, and Etienne Roquain. "Post hoc confidence bounds on false positives using reference families". In: *The Annals of Statistics* 48.3 (June 2020), pp. 1281–1303. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/19-AOS1847.

[11] Jelle J Goeman and Aldo Solari. "Multiple hypothesis testing in genomics". In: *Statistics in medicine* 33.11 (2014), pp. 1946–1978. DOI: 10.1002/sim.6082.

[12] Elise Jorge et al. *hicream: HIC diffeREntial Analysis Method*. R package version 0.0.2 (2025-11-19), published on CRAN. Maintainer. Source code at https://forge.inrae.fr/scales/hicream. 2025. URL: http://CRAN.R-project.org/package=hicream.

[13] Carson Sievert et al. "Extending ggplot2 for Linked and Animated Web Graphics". In: *Journal of Computational and Graphical Statistics* 28.2 (2019), pp. 299–308. DOI: 10.1080/10618600.2018.1513367.

[14] Toby Hocking et al. *animint2: Animated Interactive Grammar of Graphics*. R package version 2025.10.17 (2025-10-22), published on CRAN. Maintainer. Source code at https://animint.github.io/animint2/. 2025. URL: https://CRAN.R-project.org/package=animint2.

[15] Boyan Bonev et al. "Multiscale 3D genome rewiring during mouse neural development". In: *Cell* 171.3 (2017), 557–572.e24. DOI: 10.1016/j.cell.2017.09.043.