Inférence post hoc pour l'analyse différentielle de données Hi-C

Élise Jorge[1,2], Pierre Neuvial[3], 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] *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. Les données Hi-C fournissent une information sur l'organisation tridimensionnelle du génome à partir de mesures d'interactions entre positions génomiques le long de la chromatine. Cette structure en trois dimensions a un rôle important dans la régulation de l'expression des gènes. L'objectif de l'analyse différentielle est d'identifier, à partir de réplicats obtenus dans deux conditions biologiques différentes, des régions génomiques qui présentent des différences significatives de structure entre les deux conditions. Ici, nous proposons de nous appuyer sur des outils d'inférence post hoc couplés à de l'analyse différentielle restreinte aux pixels. Il est ainsi possible de quantifier la présence d'interactions différentielles dans des sous-ensembles de pixels arbitrairement choisis permettant ainsi d'identifier les régions génomiques les plus différentielles.

Mots-clés. données Hi-C, génomique 3D, inférence post hoc, tests multiples

Abstract. Hi-C data provide insights into the three-dimensional organization of the genome by measuring interactions between genomic positions along the chromatin. This three-dimensional structure plays a crucial role in regulating gene expression. Differential analysis aims to identify genomic regions that display significant differences in structure between two different biological conditions. Here, we propose to use a post hoc inference strategy on results obtained from pixel-level differential analysis. This makes it possible to quantify signal in arbitrary clusters of pixels and thus to identify differential genomic regions.

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

1 Introduction

Structure de l'ADN et données Hi-C. La chromatine est compactée au sein du chromosome selon une structure hiérarchique, comme illustré sur la figure 1 (gauche). Les données Hi-C sont des données de séquençage haut-débit qui permettent d'obtenir des informations sur l'organisation tridimensionnelle du génome dans la cellule, en mesurant la fréquence d'interactions spatiales entre régions génomiques. L'étude de ces données a permis de montrer

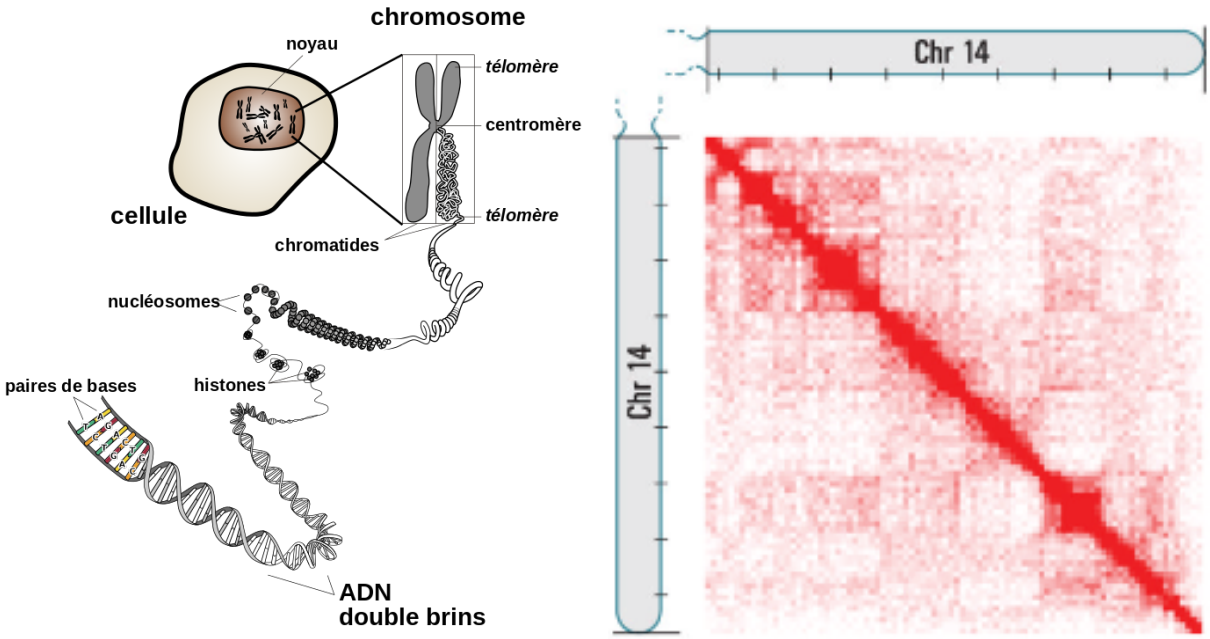

FIGURE 1 – Gauche : Schéma de la compaction de l'ADN en chromosome (« Chromosome fr » par Phrood commonswiki, Wikimedia Commons). Droite : Matrice Hi-C du chromosome 14 de [1].

qu'il existait, le long de la chromatine, des régions génomiques appelées TADs (*Topologically Associating Domains*) au sein desquels les interactions sont fréquentes.

L'apparition de modifications dans cette structure de compaction, par exemple la disparition d'une frontière entre deux TADs impliquant leur fusion, peut avoir un impact majeur sur l'expression des gènes dans la zone considérée. Ces modifications peuvent provoquer des pathologies neurologiques [2] ou des malformations [3].

D'une manière plus formelle, les données Hi-C se présentent sous la forme d'une matrice carrée symétrique dont l'entrée (i, j) – que l'on appellera « pixel » – correspond au nombre d'interactions observées dans l'expérience Hi-C entre les positions génomiques i et j. La figure 1 (droite) représente une telle matrice, dans laquelle l'intensité de couleur est proportionnelle à la valeur du nombre d'interactions du pixel considéré.

Analyse différentielle de données Hi-C. On s'intéresse ici à un problème d'analyse différentielle entre des ensembles de matrices Hi-C obtenues dans deux conditions différentes, $\mathcal{C}_1$ et $\mathcal{C}_2$. L'objectif est d'identifier des régions génomiques qui présentent des différences significatives d'interactions entre ces deux conditions.

Formellement, on possède $r = r_1 + r_2$ matrices de taille $p \times p$ où $M_k^{\mathcal{C}_1}$ $(k = 1, \cdots, r_1)$ (resp. $M_l^{\mathcal{C}_2}$ $(l = 1, \cdots, r_2)$) correspond à la matrice obtenue pour le k-ème (resp. l-ème) réplicat de la condition $\mathcal{C}_1$ (resp. $\mathcal{C}_2$).

L'analyse différentielle peut être réalisée au niveau des pixels, en faisant 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 » avec $i, j \in \{1, \cdots, p\}$ et $j \leq i$ (par symétrie, on ne regarde que la partie triangulaire supérieure de la matrice). Ainsi, le test de ces hypothèses permet d'obtenir une unique p-valeur pour chaque pixel (i, j).

Dans un benchmark [4] de méthodes d'analyse différentielle au niveau des pixels, nous avons montré que **diffHic** [5] est une des méthodes fournissant les meilleures garanties statistiques. Cependant, l'interprétation des résultats de ce type de méthode est limitée d'un point de vue biologique car l'information obtenue est ponctuelle, très dispersée sur la matrice et ne peut pas être facilement interprétée en « régions de compaction / décompaction » dans la matrice. De plus, les résultats par pixel ne sont pas facilement et directement généralisables à des ensembles de p-valeurs. Ainsi, nous nous intéressons dans la suite à une manière « d'agréger » l'information fournie par ces résultats afin d'identifier des régions génomiques différentielles.

2 L'inférence post hoc pour l'analyse différentielle

Dans cette partie, nous nous intéressons à la problématique d'extension des résultats de l'analyse différentielle ponctuelle donnant une p-valeur par pixel à une analyse différentielle sur des ensembles de pixels. On montre notamment qu'utiliser l'inférence post hoc peut permettre de quantifier le signal présent dans des sous-ensembles de p-valeurs.

Limites du contrôle du False Discovery Rate. Un grand nombre de tests étant réalisés simultanément – ici, on réalise théoriquement jusqu'à $p(p+1)/2$ tests – on doit ajuster les p-valeurs obtenues pour corriger pour la multiplicité des tests. Les p-valeurs fournies par la méthode **diffHic** sont corrigées à l'aide de la méthode de Benjamini-Hochberg (BH) [6], assurant un contrôle du FDR sur l'ensemble des pixels testés. Si l'on note $\mathcal{H}_0$ l'ensemble des hypothèses nulles et R l'ensemble des hypothèses rejetées par la méthode, le FDR s'écrit comme $\mathbb{E}\left(\mathrm{FDP}(R)\right)$, où $\mathrm{FDP}(R) = \frac{|R \cap \mathcal{H}_0|}{|R| \vee 1}$ est la proportion de faux positifs parmi l'ensemble des hypothèses rejetées par la méthode. Alors, il apparaît que le contrôle global du FDR sur l'ensemble des pixels considérés n'implique par le contrôle de cette quantité sous un sous-ensemble de pixels sélectionnés [7]. Ainsi, il n'est donc pas possible de fournir de garanties sur la présence de faux positifs dans un sous-ensemble de p-valeurs ajustées seulement en utilisant la méthode BH.

Inférence post hoc : objectif et définition. Ici, notre objectif est de donner une mesure de la quantité de signal dans des sous-ensembles de p-valeurs *arbitrairement sélectionnés*. Les méthodes post hoc [7] fournissent précisément une telle garantie. Pour tout ensemble S de p-valeurs, elles fournissent une borne supérieure sur $|S \cap \mathcal{H}_0|$, le nombre de faux positifs dans S. Formellement, on appelle *borne post hoc* [7] une fonction V_α telle que :

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

Si (1) est vérifiée, alors pour tout S la quantité $\gamma_\alpha(S) = 1 - V_\alpha(S)/|S|$ permet de minorer la proportion de vrais positifs $\mathrm{TDP}(S) = 1 - |S \cap \mathcal{H}_0|/S$:

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

Comme $\gamma_\alpha(S)$ minore la proportion de vrais positifs dans un sous-ensemble d'intérêt S, indépendamment du choix de S, cette mesure peut être utilisée pour comparer des sous-ensembles de p-valeurs arbitrairement sélectionnés.

Utilisation de la borne post hoc de Simes. Si l'on note $p_1, \ldots, p_m$ les p-valeurs correspondant aux m hypothèses nulles, on définit la borne de Simes :

$$V_\alpha^{\mathrm{Simes}}(S) = \min_{1 \leq k \leq m} \left[\sum_{i \in S} \mathbb{1}_{\{p_i > \frac{\alpha k}{m}\}} + k - 1 \right].$$

Sous des hypothèses d'indépendance des p-valeurs ou de dépendance positive (PRDS) [8], on peut montrer [9] 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 [10]. 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 $V_\alpha^{\mathrm{Simes}}$ dans le cadre de l'analyse différentielle de données Hi-C.

3 Application : changement de conformation durant la différentiation cellulaire

Nous avons implémenté la méthode décrite en section 2, et l'avons testée sur des données Hi-C issues de lignées cellulaires murines [11] pour deux conditions biologiques correspondant à des stades de différenciation cellulaire différents de cellules neuronales (ES : cellules souches embryonnaires et CN : neurones corticaux).

Ici, les ensembles arbitraires ont été définis à partir des données en utilisant une classification ascendante hiérarchique avec contrainte de voisinage [12]. Soient $i, j \in \{1, \cdots, p\}$ avec $j \geq i$. Dans la classification, chaque pixel (i, j) est représenté par $\mathbf{m}_{i,j} := (m_{i,j}^{\mathcal{C}_1,1}, \cdots, m_{i,j}^{\mathcal{C}_1,r_1}, m_{i,j}^{\mathcal{C}_2,1}, \cdots, m_{i,j}^{\mathcal{C}_2,r_1})$ correspondant aux comptages de l'interaction entre les positions génomiques i et j dans tous les réplicats des deux conditions. Nous utilisons l'implémentation **scikit-learn** avec le critère de Ward [13]. 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. Les bornes post hoc sont calculées sur les classes ainsi obtenues.

Sur la figure 2, nous représentons les résultats obtenus pour le chromosome 18. La partie inférieure de la figure correspond aux p-valeurs ajustées de l'analyse différentielle par pixel

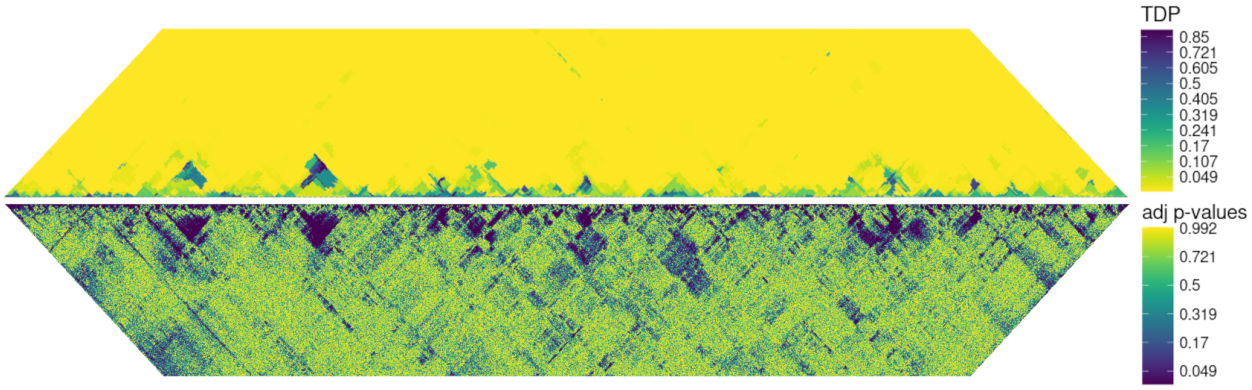

FIGURE 2 – En haut, diagonale de la (demie) matrice des résultats de l'inférence post hoc où chaque classe de pixels est représentée par sa proportion minimale de vrais positifs calculée par l'inférence post hoc. En bas, diagonale de la (demie) matrice des p-valeurs ajustées obtenues par la méthode **diffHic**.

obtenue par la méthode **diffHic**. On distingue des régions dans lesquelles se trouve un nombre important de p-valeurs ajustées contigues significatives mais ces zones sont mélées de p-valeurs ajustées plus fortes. Inversement, des p-valeurs ajustées faibles se retrouvent de manière éparpillée dans des zones isolées de la matrice.

La partie supérieure de la figure représente la proportion minimale de vrais positifs des classes obtenues par classification ascendante hiérarchique. On observe une correspondance entre des régions avec une forte densité de p-valeurs ajustées faibles (partie haute) et des classes possédant une proportion de vrais positifs élevée.

4 Perspectives

Ainsi, l'utilisation de l'inférence post hoc couplée à une approche totalement automatisée de classification ascendante hiérarchique a permis de passer d'une information ponctuelle (les p-valeurs ajustées fournies par **diffHic**) et biologiquement limitée à une information sur des régions génomiques différentielles.

Dans la suite, on s'intéressera à la validation biologique des résultats obtenus, en identifiant par exemple les gènes présents dans les classes dont les proportions de vrais positifs sont les plus élevées.

L'approche que nous avons décrite est en cours d'évaluation sur d'autres jeux de données comme les données [14] correspondant à des cellules murines post-mitotiques. Dans ce jeu de données, on retrouve une condition contrôle (CTCF+) et une condition dans laquelle une déplétion de la protéine CTCF a été provoquée (CTCF-). Dans la condition de traitement (CTCF-), une modification de structure est attendue de par la présence favorisée de la protéine CTCF dans des régions telles que les frontières de TADs. Nous souhaitons utiliser des informations biologiques telles que la présence de pics CTCF afin d'obtenir une validation des classes identifiées par la méthode comme différentielles.

Remerciements

Ce travail est soutenu par le groupe de travail ChrocoNET financé par le métaprogramme INRAE DIGIT-BIO. La thèse d'Élise Jorge est financée par INRAE.

Bibliographie

[1] Erez Lieberman-Aiden, Nynke L. Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M.A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

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

[3] 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.

[4] Élise Jorge, Sylvain Foissac, Pierre Neuvial, Matthias Zytnicki, and Nathalie Vialaneix. A comprehensive review and benchmark of differential analysis tools for Hi-C data. *Briefings in Bioinformatics*, 2025. Forthcoming.

[5] 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.

[6] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[7] Jelle J Goeman and Aldo Solari. Multiple testing for exploratory research. *Statistical Science*, 26(4):584 – 597, 2011.

[8] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.

[9] 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.

[10] Jelle J Goeman and Aldo Solari. Multiple hypothesis testing in genomics. *Statistics in medicine*, 33(11):1946–1978, 2014.

[11] Boyan Bonev, Netta Mendelson Cohen, Quentin Szabo, Lauriane Fritsch, Giorgio L Papadopoulos, Yaniv Lubling, Xiaole Xu, Xiaodan Lv, Jean-Philippe Hugnot, Amos Tanay, and Giacomo Cavalli. Multiscale 3D genome rewiring during mouse neural development. *Cell*, 171(3):557–572.e24, 2017.

[12] 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.

[13] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[14] Haoyue Zhang, Jessica Lam, Di Zhang, Yemin Lan, Marit W Vermunt, Cheryl A Keller, Belinda Giardine, Ross C Hardison, and Gerd A Blobel. Ctcf and transcription influence chromatin structure re-configuration after mitosis. *Nature communications*, 12(1):5157, 2021.