Analyse comparative de données de génomique 3D

Sujet de thèse 2023-2026
Sylvain Foissac (GenPhySE), Nathalie Vialaneix (MIAT), INRAE Toulouse-Auzeville &
Pierre Neuvial (Institut de Mathématiques de Toulouse)

Contexte biologique. La conformation tridimensionnelle du génome a un impact majeur sur son fonctionnement, avec des implications importantes dans le développement fœtal, la différentiation cellulaire ou le développement de maladies comme le cancer[1]. Or, les progrès récents de la biologie moléculaire ont permis de changer la façon d'étudier l'organisation spatiale des chromosomes et donc de mieux comprendre les liens entre structure et fonction du génome. En particulier, la technologie dite Hi-C, pour « High-throughput chromosome conformation capture »[2], produit des données de séquençage d'ADN qui visent à estimer la proximité spatiale entre régions génomiques en mesurant la fréquence d'interactions physiques entre ces régions. L'ensemble des mesures associées à toutes les paires de régions forme ainsi une matrice d'interactions qui caractérise la structure 3D du génome dans les cellules d'un échantillon biologique donné.

Questions scientifiques. Ce projet est motivé par la problématique de la comparaison de ces matrices, afin d'identifier les différences significatives de conformation génomique entre groupes d'échantillons associés à des phénotypes distincts. La plupart des méthodes d'analyse différentielle qui abordent actuellement cette question[3] procèdent en identifiant des positions ponctuelles différentielles dans la matrice. Comme discuté dans [Foissac *et al.*, 2023 ; travail en cours], cette approche est pertinente pour identifier des variations de structures de type boucle d'ADN par exemple, typiquement entre régulateur et promoteur. En revanche, ces approches ne rendent pas bien compte de changements de structures plus larges de type TADs par exemple (*Topologically Associating Domains*), qui sont pourtant des éléments fonctionnels critiques[4]. C'est cette limite que nous souhaitons aborder dans cette thèse avec une approche plus générale reposant sur une modélisation originale des matrices en arbres.
De manière plus précise, l'objectif de cette thèse consiste à s'appuyer sur la méthode développée lors d'une précédente thèse[5] (sous la direction des mêmes co-encadrants en partie) pour proposer une approche permettant de réaliser ces analyses de manière automatisée à l'échelle du génome. En effet, la méthode actuelle permet de réaliser un test statistique qui compare la structure de deux familles d'arbres. Elle a déjà été utilisée dans le cadre de l'analyse de données Hi-C lors d'une

1 Lupiáñez, Kraft, Heinrich *et al.* (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5): 1012-1085.
Won, de la Torre-Ubieta, Stein *et al.* (2016) Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature*, 538(7626): 523-527.
Zheng and Xie (2019) The role of 3D genome organization in development and cell differentiation. *Nature Reviews Molecular Cell Biology*, 20(9): 535-550.

2 Lieberman-Aiden, Van Berkum, Williams *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950): 289-293.

3 Lun and Smyth (2015) diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics*, 16: 258.
Stansfield, Cresswell, Vladimorov, and Dozmorov (2018) HiCcompare: an R package for joint normalization and comparison of Hi-C data. *BMC Bioinformatics*, 19: 279.
Djekidel, Chen, and Zhang (2018) FIND: difFerential chromatine INteractions Detection using a spatial Poisson process. *Genome Research*, 28: 412-422.
Cook, Hristov, Le Roch, Vert, and Noble (2020) Measuring significant changes in chromatine conformation with ACCOST. *Nucleic Acids Research*, 48(5): 2303-2311.

4 Dixon, Selvaraj, Yue *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485: 376-380.

5 Neuvial, Randriamihamison, Chavent, Foissac, Vialaneix (2022) Testing differences in structure between families of trees. *Submitted for publication.*

preuve de concept, en utilisant une modélisation préalable de la structure hiérarchique des matrices par CAH (Classification Ascendante Hiérarchique) contrainte. Toutefois, plusieurs verrous méthodologiques et théoriques restent à lever pour la rendre utilisable à l'échelle du génome :

- déterminer comment cibler les régions à tester (le test de l'ensemble des sous-matrices d'une matrice Hi-C obtenue à l'échelle du génome ayant un coût de calcul prohibitif). En particulier, nous envisageons la mise au point d'un algorithme de parcours hiérarchique guidé de la matrice entière ;
- contrôler, au niveau de la hiérarchie des régions testées, la validité statistique des tests effectués : en effet, les tests réalisés sur la hiérarchie sont fortement redondants car imbriqués, voire potentiellement chevauchants.

Pour réaliser ces deux objectifs, une des premières pistes identifiées consiste à adapter des travaux sur le contrôle hiérarchique du FDR[6] permettant de les appliquer en conservant un maximum de puissance dans le contexte particulier de notre application.
L'approche développée sera utilisée dans le cadre de projets en cours, portant par exemple sur l'annotation de génomes animaux[7] ou l'étude du développement musculaire porcin en lien avec la mortalité périnatale afin d'améliorer les conditions de vie d'animaux d'élevage[8].
Le sujet est donc à l'interface entre apprentissage statistique (non supervisé, *data mining*), bioinformatique et biologie moléculaire, avec des aspects computationnels critiques pour le passage à l'échelle sur ces données collectées à l'échelle du génome. L'objectif du projet sera aussi de rendre accessible l'outil développé dans un package logiciel publié en open source.

Enjeux sociétaux et finalisés. L'objectif finalisé de la thèse est en lien avec les données de projets en cours :

- Ces dernières années ont vu la collecte de plus en plus importantes de données omiques diverses sur les espèces d'intérêt agricole (project FAANG https://www.faang.org/). L'objectif est de mieux appréhender les mécanismes moléculaires impliqués dans les phénotypes visibles à l'échelle de l'individu (croissance, résistance aux maladies, robustesse, …) et d'ouvrir de nouvelles pistes pour l'amélioration de ces espèces. Mais cet objectif doit être soutenu par le développement de méthodes d'analyse adaptées à la diversité et la particularité des données générées. Le projet de thèse aborde une partie de cette question et, en particulier, utilisera les données Hi-C produites dans le cadre du projet FR-AgENCODE et du projet H2020 GENE-SWitCH (www.gene-switch.eu) comme illustration de la pertinence de la méthode développée ;
- Dans les élevages porcins, le progrès génétique des dernières années s'est accompagné d'une augmentation substantielle de la mortalité des porcelets. La maturité du porcelet, définie comme l'état de plein développement permettant la survie à la naissance, est un déterminant important de la mortalité précoce mais ses mécanismes sont encore largement méconnus. Le projet de thèse utilisera donc aussi les données de (Marti-Marimon et al, 2021)[8] pour approfondir les résultats déjà obtenus sous un angle plus structurel.

Enfin, d'un point de vue méthodologique, la thèse permettra de produire un package **R** ou un pipeline python complet et finalisé permettant l'utilisation de la méthode à l'échelle du génome.

Financement de la thèse. La moitié du financement de la thèse est assuré par un engagement des départements INRAE. L'autre moitié est en cours d'instruction pour un financement par la Région Occitanie. En cas de non financement de la seconde moitié, les encadrants peuvent préparer la

6 Yekutieli (2008) Hierarchical false discovery rate-controlling methodology. *Journal of the American Statistical Association*, 103(481): 309-316.

7 Foissac, Djebali, Munyard, Vialaneix *et al.* (2019) Multi-species annotation of transcriptome and chromatin structure in domesticated animals. *BMC Biology*, 17: 108.

8 Marti-Marimon, Vialaneix, Lahbib-Mansais, Zytnicki *et al.* (2021) Major reorganization of chromosome conformation during muscle development in pig. *Frontiers in Genetics*, 12: 748239.

candidate potentielle[9] au concours de l'école doctorale SEVAB https://ed-sevab.univ-toulouse.fr/ ou bien MITT https://ed-mitt.univ-toulouse.fr.

Profil recherché. La candidate sera issue d'un parcours en bioinformatique ou mathématiques appliquées avec un goût marqué pour les applications et la programmation. Les encadrants sont en mesure de s'adapter (et d'adapter le sujet de thèse) à la diversité de ces profils (pouvant aller d'une appétence plus grande pour la biologie à un goût pour les développements théoriques en statistique).

Modalités de candidature. Envoyer un CV et une lettre de motivation à sylvain.foissac@inrae.fr et nathalie.vialaneix@inrae.fr.

9 Annonce écrite au féminin générique : l'emploi du féminin n'indique donc pas une préférence de genre.