DUPRAZ–BARDOU
Roman

ENSAE 1re année
Stage d'ouverture au monde professionnel
Année scolaire 2023 - 2024

Défaire les nœuds de l'ADN : Analyse différentielle des matrices Hi-C

MIAT (Inrae)
Auzeville-Tolosane

Maitres de stage : Nathalie VIALANEIX
et Elise JORGE
10/06/2024 - 26/07/2024

Sommaire

Remerciements

Je tiens tout d'abord à remercier l'ENSAE Paris et plus particulièrement tous les acteurs travaillant au sein de l'école qui ont rendu ce stage possible.

Je remercie l'ensemble des personnes de Inrae que j'ai pu côtoyer et en particulier le laboratoire MIAT pour son accueil.

Enfin, je tiens à remercier tout particulièrement Elise et Nathalie, mes deux maîtres de stage pour m'avoir fait découvrir le monde de la recherche et pour avoir grandement contribué à la réussite de ce stage.

Introduction

Ayant un projet professionnel assez précis, j'ai vu dans ce stage d'ouverture une possibilité de découvrir un milieu qui m'intéresse depuis longtemps : celui de la recherche en mathématiques. Je le connaissais déjà par certains proches qui y travaillent mais il me paraissait nécessaire de m'y confronter pendant presque deux mois, afin de confirmer ou infirmer mes préjugés et mes attentes. C'est avant tout l'enseignement qui m'intéresse mais au vu de la difficulté de faire des stages courts dans ce domaine, j'ai préféré m'orienter vers la recherche qui est une étape presque nécessaire et une partie intégrante de l'enseignement dans le supérieur.

Souhaitant effectuer mon stage près de Toulouse pour des raisons pratiques, j'ai cherché du côté de l'Université Paul Sabatier et notamment de l'Institut de Mathématiques de Toulouse (IMT). C'est comme cela que mon C.V. est parvenu à Pierre Neuvial, responsable de l'équipe Statistique et Optimisation de l'IMT, qui a transmis mon dossier à Nathalie Vialaneix et Elise Jorge. Nathalie Vialaneix est directrice de recherche à INRAE et directrice d'unité adjointe du laboratoire MIAT. Elise Jorge est en première année de thèse à INRAE, au sein du laboratoire GenPhySE et est co-encadrée, en partie, par Nathalie Vialaneix et Pierre Neuvial.

Par ailleurs, Pierre Neuvial est un alumni ENSAE (2003) et il encadrait au même moment un stage d'un élève de deuxième année à l'ENSAE (Nils Peyrousset). Cela montre la cohérence de mon stage et de mon projet professionnel avec les enseignements de l'ENSAE.

Avant de commencer mon stage, j'ai eu deux entretiens en visioconférence, un le 28 mars et le deuxième le 16 mai. Le premier, avec Nathalie Vialaneix, consistait en une discussion où je présentais mon parcours et mes acquis de premère année à l'ENSAE, ainsi que ma motivation pour ce stage. En retour, Nathalie Vialaneix m'a présenté son parcours ainsi que certains de ses travaux. Le deuxième entretien, lors duquel Elise Jorge était également présente, a consisté à régler les derniers préparatifs de mon stage ainsi que mon accueil à INRAE. Les deux entretiens étaient formels mais beaucoup moins que peuvent l'être des entretiens avec un responsable des ressources humaines. J'ai été convaincu par l'idée de suivre le déroulé d'une thèse en statistique (une partie tout du moins) et d'en apprendre plus sur ses attendus et son fonctionnement.

Mon stage s'est déroulé du 10 juin au 26 juillet et avait pour objet d'appliquer la méthode développée par Elise durant sa thèse à de nouvelles données. En effet, en partant de ses travaux et de l'état de l'art sur le sujet, je devais comprendre les enjeux et processus à la fois statistiques et biologiques puis analyser et normaliser des données brutes pour enfin produire des graphiques pouvant être utilisés dans le travail d'Elise.

1 Présentation de l'organisme d'accueil et de l'environnement de travail

1.1 Organisme d'accueil

L'Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement (INRAE) est un organisme public de recherche qui est installé sur l'ensemble du territoire français. Plus précisement, c'est un *établissement public à caractère scientifique et technologique* (EPST) au même titre que le CNRS ou l'INED par exemple. En plus d'être divisé en centres régionaux, INRAE est aussi divisé en différents départements de recherche, pour chacun de ses domaines d'études et d'applications. Lors de mon stage, j'ai été accueilli par l'unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT) rattachée au département *Mathématiques et numérique* (MATHNUM). Comme beaucoup d'unités de recherche, le MIAT est divisé en équipes, aux objectifs et champs d'applications variés. Cela permet une meilleure collaboration et communication entre les membres d'une même équipe ou d'équipes différentes. En effet, diviser un laboratoire en équipes permet de cartographier les domaines d'applications des différents chercheurs mais aussi de faciliter les échanges et projets de recherches (comme des thèses par exemple).

FIGURE 1 – Logo INRAE

FIGURE 2 – Logo MIAT

Tout comme Nathalie Vialaneix, j'étais rattaché à l'équipe Statistique et Algorithmique pour la Biologie (SaAB) qui analyse des données biologiques et développe des outils statistiques et informatiques permettant d'aider les chercheurs en biologie. Personnellement, je trouve que cela illustre parfaitement le rôle de la recherche en mathématiques appliquées : il faut réussir à adapter des outils mathématiques et informatiques (parfois très théoriques) à des enjeux et problématiques plus concrets (ici biologiques). Cela nécessite donc une compréhension très fine des processus mathématiques mais aussi une grande ouverture scientifique pour comprendre le domaine d'application (alors même que ce n'est pas toujours notre domaine de prédilection). La versatilité est l'une des choses que j'apprécie particulièrement dans la recherche en mathématiques appliquées car cela permet de s'orienter vers une multitude de métiers variés.

J'ai aussi été encadré par Elise Jorge, doctorante au sein de l'unité Génétique Physiologie

et Systèmes d'Elevage (GenPhySE). Sa thèse est co-encadrée par Nathalie Vialaneix, Sylvain Foissac (GenPhySE) ainsi que Pierre Neuvial (CNRS, Institut de Mathématiques de Toulouse). Cette thèse est interdisciplinaire et mobilise plusieurs unités de recherche. D'un point de vue personnel, j'ai trouvé ce point très intéressant car cela permet d'échanger avec des personnes aux parcours très variés, aussi bien en termes d'études que d'expériences professionnelles.

1.2 Environnement professionnel

1.2.1 Organisation du travail

Lors de mon stage, je disposais d'une certaine autonomie. C'est un point que j'apprécie tout particulièrement dans le métier de chercheur et qui permet de s'organiser assez librement, du moment qu'on anticipe assez. Je pouvais commencer mes journées plus tôt et les finir plus tard si nécessaire ou inversement. Je pense que c'est une chose nécessaire pour les chercheurs car ils peuvent moduler leur volume de travail en fonction de leurs échéances mais aussi de leurs contraintes personnelles. Cela conduit à un plus grand épanouissement mais peut aussi permettre, à mes yeux, une meilleure productivité en fonction de notre état d'esprit.

Néanmoins, si j'avais un soucis, je n'étais jamais vraiment seul. Je pouvais communiquer avec mes encadrantes grâce à un service de messagerie instantanée (Mattermost). Cela me permettait de poser des questions techniques ou concernant la méthodologie mais aussi de fixer des réunions concernant la thèse d'Elise Jorge et/ou mon stage. Je m'en suis servi quasiment tous les jours de mon stage et il permet à chacun de communiquer en dehors de ses contraintes temporelles et spatiales (séminaires, missions, télé-travail, congés...). De plus, lorsque des problèmes techniques le nécessitaient ou que des personnes ne pouvaient pas se déplacer physiquement pour des réunions, nous avions recours au logiciel zoom pour faire des visioconférences.

Outre cela, toutes les semaines, j'assistais à des réunions de suivi de thèse entre Elise et ses encadrants durant lesquelles elle présentait ses avancées techniques, ses interrogations et ses objectifs. C'était un moment privilégié pour elle, qui lui permettait de cadrer son travail mais aussi d'envisager de nouveaux angles de recherche. De mon côté, cela me permettait de voir quelles étaient les étapes et les échéances d'une thèse de manière plus concrète.

De plus, je partageais mon bureau avec Éric Casellas (Ingénieur d'études au MIAT) qui a su m'aider pour certains des problèmes techniques que j'ai rencontrés. Il y avait aussi une documentation extrêmement riche, écrite par les agents du MIAT et d'INRAE qui permet de se former facilement à certains outils techniques. Je l'ai personnellement beaucoup utilisée pour me former à l'utilisation avancée de Git et Linux, notamment pour l'installation de programmes.

1.2.2 Outils et méthodes de travail

Au cours de ces sept semaines, j'ai aussi dû me familiariser avec des outils techniques spécifiques au monde de la recherche mais aussi au MIAT. Tout d'abord, les logiciels dits *Open Source* occupaient un rôle central à commencer par le système d'exploitation qui était Linux. Il change radicalement de Windows et MacOS, notamment pour l'installation de logiciels mais il présente de nombreux avantages. Tout d'abord, la distribution Debian de Linux est entièrement gratuite, et son code source est libre d'accès, d'utilisation et de modification. Cela est extrêmement utile pour l'adapter à une utilisation spécifique pour ensuite le partager entre utilisateurs. C'est une des raisons qui le rend très utile dans la recherche. J'ai tout de suite fait le lien avec le concept d'*Open Data* qui avait été abordé dans le cours d'introduction à l'éthique et au droit des données de Madame Tubaro. Que ce soit pour l'*Open Data* ou les logiciels *Open Source* l'objectif est de démocratiser leur utilisation dans l'intérêt public et général. Le mouvement *Open Source* est un mouvement se basant d'ailleurs sur des principes de liberté et d'égalité. Ainsi, les logiciels *Open Source* s'adaptent facilement aux contraintes et aux objectifs des chercheurs, permettent de mutualiser plus efficacement des connaissances et présentent aussi des avantages individuels. Moyennant un temps de formation et d'appropriation parfois plus conséquent que les logiciels *Closed Source* ou propriétaires, ils permettent une plus grande liberté dans leur utilisation. De plus, même s'ils sont parfois compliqués d'utilisation, leur architecture est souvent assez similaire. En effet, pour qu'un logiciel soit considéré comme *Open Source* il faut que leur programmation respecte des règles assez contraignantes (notamment en termes de commentaires) mais permettant au plus grand nombre de le comprendre et de pouvoir l'utiliser et le modifier à sa guise.

Une deuxième spécifité technique de ce stage était le versionnement. Le versionnement, qui consiste à gérer les différentes versions d'un fichier, est très pratique pour travailler à plusieurs sur un projet sans en perdre le fil. Le logiciel utilisé au MIAT était Git qui permet de décentraliser ses travaux pour les rendre accessibles à d'autres collaborateurs. Je m'en servais notamment pour partager l'avancée de mon travail mais aussi pour récupérer des codes fournis par mes encadrantes et nécessaires pour mon stage. Cela servait aussi pour partager les sources LaTeX des supports de présentation ou bien de mon rapport de stage. Néanmoins, j'étais déjà familier avec l'environnement de Git grâce au projet de programmation du second semestre de première année de l'ENSAE qui avait entièrement été fait sur VS Code intégrant un client Git. J'ai ainsi pu renforcer mes compétences dans un environnement qui ne m'était pas totalement inconnu.

Enfin, la bioinformatique étant une discipline qui nécessite une puissance de calcul impor-

tante, le MIAT administre un cluster de calculs et de stockage accessible à une communauté en bioinformatique de plus d'un millier d'utilisateurs : GenoToul-Bioinfo. Il permet de réaliser des *jobs*, qui sont un ensemble de tâches, à distance et de manière bien plus efficace que ne le ferait un ordinateur de travail. Les données et les résultats sont stockés par le cluster et il n'est ainsi pas nécessaire de garder son ordinateur allumé pendant la réalisation du *job*. Le cluster est géré par un ordonnanceur. Un ordonnanceur est un logiciel qui lance l'exécution des *jobs*, leur attribue un ordre de priorité en fonction des capacités du cluster (nombre et puissance des processeurs, mémoire...) et les interrompt en cas de problème technique. L'ordonnanceur utilisé par GenoToul-Bioinfo est SLURM, qui est un logiciel *Open Source* utilisé par la plupart des clusters. Son utilisation n'est pas toujours aisée mais je pense qu'avoir appris à utiliser un tel outil me sera forcément bénéfique plus tard. Les métiers vers lesquels l'ENSAE oriente sont tournés vers l'utilisation et le traitement de données, cela nécessite souvent d'utiliser des clusters de calcul pour accélérer le traitement voire même le rendre possible quand le stockage ou la mémoire vive (RAM) en local ne sont pas suffisants. De plus, grâce au cours sur l'impact du numérique, j'étais sensibilisé aux enjeux liés à l'utilisation intensive des clusters de calcul, ce qui m'a poussé à en avoir un usage raisonné.

2 Description de la mission

2.1 Données Hi-C

Comme expliqué précédemment, mon travail s'est inscrit dans la continuité de la thèse d'Elise Jorge intitulée *Analyse comparative de données de génomique 3D*. L'entièreté des êtres vivants sont constitués d'une à plusieurs milliers de milliards de cellules. Ces cellules contiennent, dans leur noyau, l'information génétique dont les chromosomes font partie. Ils sont eux-mêmes composés d'un fil, appelé chromatine. Elle est composée de paires de bases azotées (Adénine et Thymine ou Guanine et Cytosine) qui code l'expression génétique de la cellule (sa fonction). La chromatine est extrêmement compacte dans un chromosome ce qui résulte en une conformation spatiale particulière. Celle-ci influence le bon ou mauvais fonctionnement de la cellule. Notamment, il y a certaines zones denses en chromatine et isolées du reste du chromosome qui sont appelées des *Domaines Topologiquement Associés* ou *Topological Associated Domain* (TAD). Ils évoluent en même temps que la cellule se spécialise (différenciation cellulaire) et leur étude permet donc de connaître le stade biologique d'une cellule. De plus, leur modification peut entraîner des malformations ou maladies génétiques assez graves. Il est donc important d'étudier la conformation spatiale des chromosomes et notamment au travers des TADs. La figure 3 décrit l'organisation spatiale de l'ADN et y sont représentés les cellules, leur chromosomes et les TADs de ses derniers.

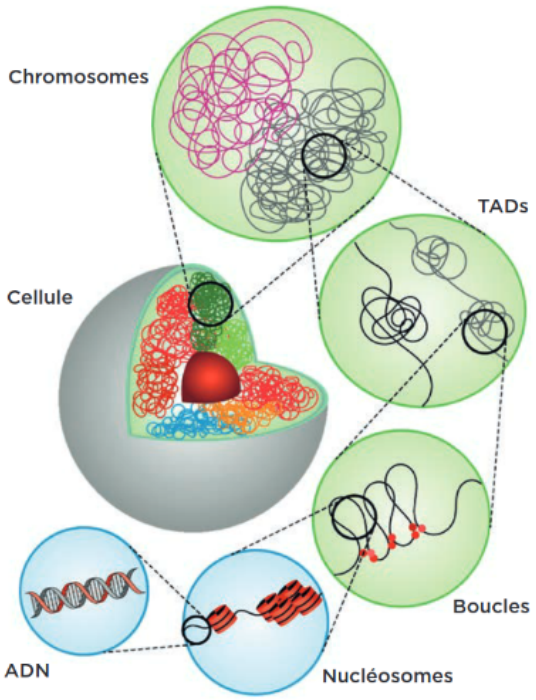

FIGURE 3 – Organisation spatiale de l'ADN, Biologie, *Comprendre l'organisation spatiale de l'ADN à l'aide de la statistique*, (p.164 - 168) (P. Neuvial, N. Vialaneix, S. Foissac)

La méthode *HIgh-throughput Chromatin Conformation Capture* (Hi-C) est une approche expérimentale basée sur le séquençage haut débit qui permet de décrire la conformation spatiale de la chromatine dans le chromosome. De manière simplifiée, l'approche consiste à découper la chromatine en petites zones génomiques et à compter le nombre de contacts observés dans un grand nombre de cellules entre chacune des paires de zones. Il est supposé que plus le nombre de contacts entre deux zones génomiques (appelées *bins*) est élevé, plus elles sont proches dans le chromosome. Au final, on obtient une matrice dite de comptage où le coefficient $(i; j)$ est le nombre de contacts entre le *bin* i et le *bin* j. On se rend compte directement que la matrice obtenue par méthode Hi-C est à coefficient entiers positifs et qu'elle est symétrique. On réduit donc souvent son analyse à sa moitié triangulaire supérieure. De plus, comme on peut le remarquer sur la figure 4, les nombres de comptage les plus importants sont proches de la diagonale. Cela signifie que les *bins* dont les numéros sont proches ont un nombre important de contact. Enfin, la résolution des données représente la taille des *bins* en kilobase (ici 10 kb, 50kb et 250 kb).

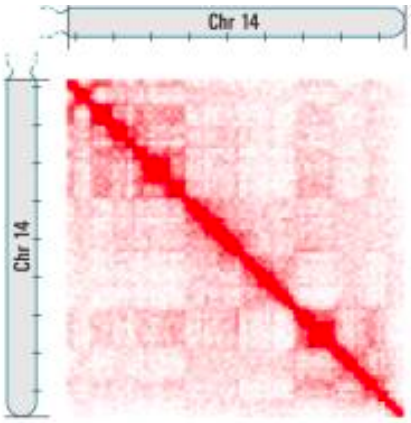

FIGURE 4 – Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950), 289-293 (Lieberman-Aiden et *al.*)

Après avoir obtenu ces données, il est pertinent d'essayer de trouver les positions des TADs. Ils se traduise par des triangles à valeurs plus importantes représentés sur les figures 5 et 6. En effet, il y a beaucoup de contacts entre les parties proches et peu avec le reste du génome (frontières). Comme expliqué précédemment, ils ont un lien fort avec l'expression cellulaire et leur étude présente de nombreux enjeux en génomique.

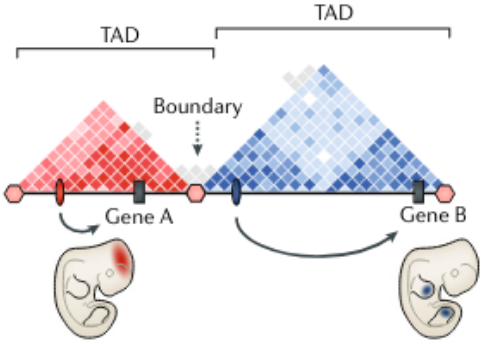

FIGURE 5 – Structural variation in the 3d genome. *Nature Reviews Genetics*, 19(7), 453-467 (Spielmann et *al.*)

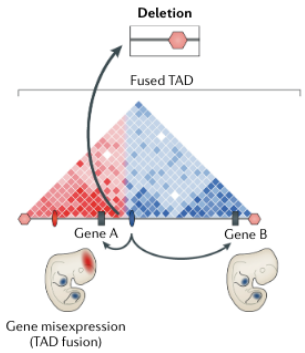

FIGURE 6 – Structural variation in the 3d genome. *Nature Reviews Genetics*, 19(7), 453-467 (Spielmann et *al.*)

2.2 Normalisation des données

Une fois les matrices Hi-C récupérées, l'objectif est de les comparer en fonction du stade biologique. On fait ce qu'on appelle de l'analyse différentielle. Par exemple, on peut comparer des cellules musculaires de foetus de porc dans deux stades biologiques différents (90 et 110 jours de gestation). Ici, sur la figure 8, on a trois réplicats biologiques par stade biologique.

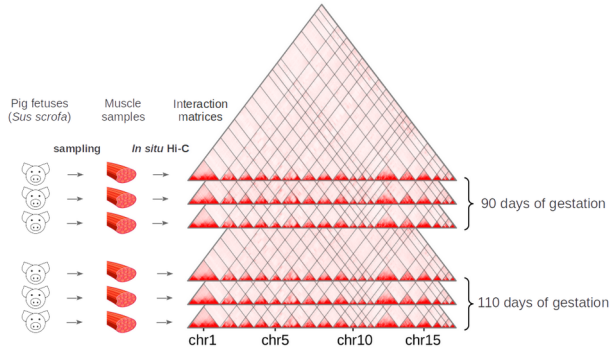

FIGURE 7 – Major reorganization of chromosome conformation during muscle development in pig. *Frontiers in Genetics*, 12, 748239 (Marti-Marimon et *al.*)

Néanmoins, on ne peut pas faire de l'analyse différentielle directement sur nos données brutes en raison de l'existence de biais. Ils proviennent principalement de l'expérience biochimique mise en œuvre pour obtenir les matrices Hi-C ainsi que de la technique de récupération des données par séquençage haut débit. Ma première tâche a donc été de normaliser ces données pour supprimer ce biais. Pour supprimer le biais de ses matrices il existe deux types de normalisation : la normalisation intra-matrice et celle inter-matrice.

La normalisation intra-matrice normalise matrice par matrice, indépendamment des autres. Le but est d'avoir un nombre de contacts par *bins* constant. Pour la matrice cela signifie que la

somme des coefficients de chaque ligne [1] doit être constante.

$$\forall i \in [\![1;n]\!], \sum_{j=1}^{n} A_{ij} = C \qquad où\ C \in \mathbb{R}$$

Par exemple, on peut choisir C de cette manière :

$$\forall (i,j) \in [\![1;n]\!]^2, A'_{ij} = A_{ij} \frac{\sum\limits_{k,l \in [\![1;n]\!]} A_{kl}}{\sum\limits_{k=1}^{n} A_{ik}} \qquad donc\ ici\ C = \sum_{k,l \in [\![1;n]\!]} A_{kl}$$

Cependant, ce type de normalisation n'assure pas de supprimer le biais entre matrices, il supprime seulement le biais de chacune des matrices indépendamment des autres. Or cela est nécessaire pour pouvoir rigoureusement comparer deux groupes de matrices [2] comme dans l'analyse différentielle.

La normalisation inter-matrice normalise un groupe de matrices en entier et permet donc de faire de l'analyse différentielle. Pour cela, il existe plusieurs méthodes mais celle qui m'a été proposée est la normalisation par régression LOESS. Elle fait appel à des notions déjà vu en Introduction à la statistique (régression linéaire par la méthode des moindres carrés) mais aussi à des notions importantes qui seront nécessairement abordées dans la suite de mon cursus (algorithmes KNN, moyenne mobile, régression avec pondération). L'idée est de supprimer le biais existant entre chaque couple de matrices (plus ou moins de contacts en fonction de l'expérience) en procédant donc à une normalisation inter-matrices. Considérons A et A' deux matrices Hi-C. On trace le *MA plot* qui est le nuage de points de différences logarithmiques $M = \log A_{ij} - \log A'_{ij}$ en fonction de la distance génomique $|i-j|$. On estime ensuite la tendance des données par régression LOESS. Le but est que la différence logarithmique, qui représente le logarithme du rapport des coefficients [3], soit nulle en moyenne. Cela supprimerait les écarts du nombre de contacts obtenus d'une matrice Hi-C à l'autre. Comme on le voit sur le graphique de gauche de la figure 8, il existe un biais dans les données *Raw* (brutes). L'idée est donc de modifier les valeurs de sorte que la courbe bleue, obtenue par régression LOESS, soit nulle (ce qui est fait sur le graphique de droite de la figure 8).

1. Ou chaque colonne, ce qui est équivalent car les matrices Hi-C sont symétriques
2. Représentant chacun une condition biologique de la cellule
3. $M = \log A_{ij} - \log A'_{ij} = \log \frac{A_{ij}}{A'_{ij}}$ donc, en moyenne $\log \frac{A_{ij}}{A'_{ij}} \simeq 0 \Leftrightarrow A_{ij} \simeq A'_{ij}$

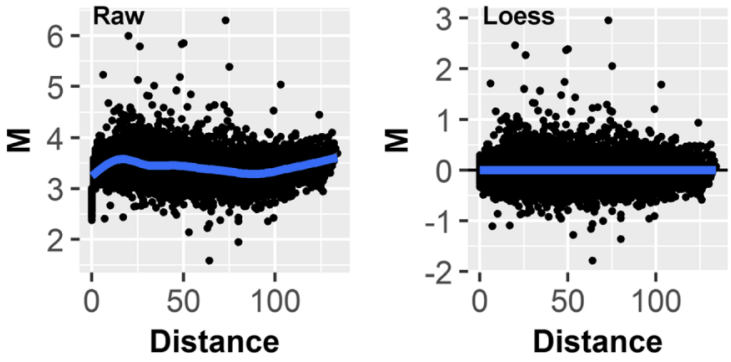

FIGURE 8 – Hiccompare : An r-package for joint normalization and comparison of hi-c datasets. *BMC bioinformatics*, 19, 1-10 (Stansfield et *al.*)

2.3 Application

Les données sur lesquelles j'ai travaillé provenaient de cellules nerveuses de souris. Ces cellules étaient réparties en trois stades cellulaires différents , comme représenté sur la figure 9. Il y a le stade embryonnaire (ES), les cellules progénitrices neurales (NPC) ainsi que les neurones corticaux (CN). J'avais à ma disposition trois résolutions pour chacun des quatre réplicats biologiques, dans chacun des trois stades cellulaires, et pour les 19 chromosomes (soit 684 matrices Hi-C).

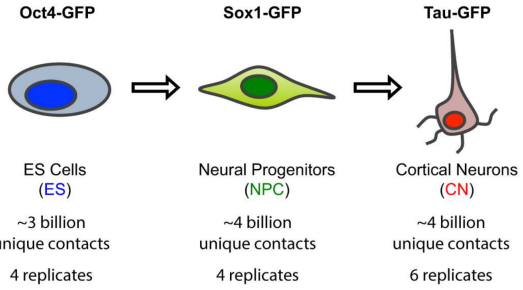

FIGURE 9 – Multiscale 3d genome rewiring during mouse neural development. *Cell*, 171(3), 557-572. (Bonev et *al.*)

Afin de traiter ces données, j'avais à ma disposition des scripts R. Ils étaient écrits pour fonctionner sur d'autres types de données cellulaires (cellules musculaires d'embryon de porc). Je devais ainsi les comprendre, les adapter aux données que j'avais et les faire fonctionner (en local ou sur le cluster). Les fichiers de base étaient des matrices Hi-C par condition et par réplicats biologiques. À partir d'un premier script, je devais découper ces matrices pour en avoir 19, une par chromosome. Ceci étant fait, je pouvais les normaliser à l'aide d'un script que j'ai dû adapter à mes données, en particulier leur format, et sur lequel j'ai effectué des corrections, notamment sur la façon de normaliser. Enfin, un dernier script mis au point par Elise, permettait d'identifier les TAD et les zones à forte densité génomique mais aussi de générer des graphiques des matrices Hi-C avec les TAD mis en évidence. Pour les scripts permettant la normalisation

et la génération des graphiques, j'ai utilisé le cluster, tout d'abord pour rendre possible leur exécution. En effet, pour les fichiers de résolutions 10 kB, qui sont donc les plus volumineux en termes d'espace de stockage, il était impossible de faire fonctionner les scripts en local. De plus, pour pouvoir automatiser ses scripts, les faire fonctionner sur chacun des fichiers et cela même de nuit, le cluster m'a été d'une grande utilité. J'ai donc écrit des jobs, en m'inspirant d'exemples fournis par Nathalie et Elise puis en modifiant les paramètres (temps d'exécution, nombre de processeurs, mémoire allouée).

2.4 Résultats

Finalement, j'ai réussi à produire les résultats escomptés dans le temps qui m'était donné. J'ai tout d'abord normalisé l'entièreté des données qui était à ma disposition. Il est important de notifier que les résultats que j'ai obtenu me paraissait assez suspects. Ne voyant pas d'où venait le problème dans mon code, j'ai donc décidé d'en parler à Nathalie et Elise lors d'une de nos réunions. Et en quelques minutes, elles se sont rendu compte que cela venait d'une erreur dans une bibliothèque utilisée. Ayant déjà rencontré ce problème, elles ont presque immédiatement trouvé la solution, ce qui m'a montré à quel point il est nécessaire de remettre en question ses résultats, mais surtout d'en discuter avec d'autres personnes.

Une fois les données correctement normalisées, je me suis occupé de produire des graphiques illustrant les différences de conformations spatiales des chromosomes en fonction des stades cellulaires. Je me suis appuyé, comme expliqué précédemment, sur les travaux d'Elise afin de produire ces derniers dont vous pouvez voir un exemple avec la figure 10. Les p-valeurs des tests $A_{ij} = A'_{ij}$ sont représentées. Ainsi plus une p-valeur est proche de 0, plus ces zones diffèrent statistiquement entre les deux conditions.

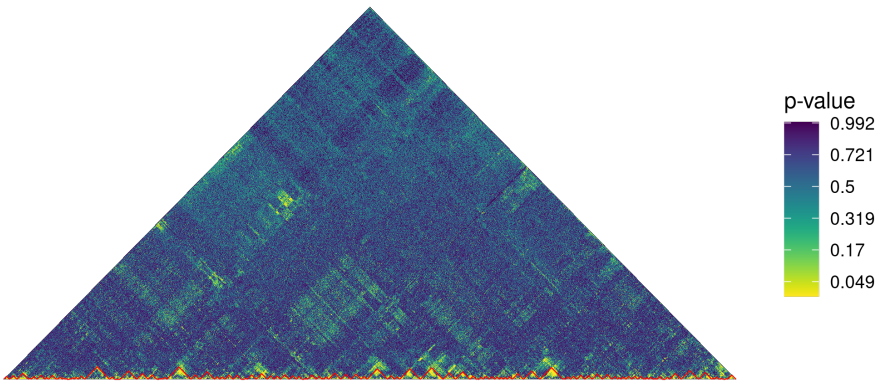

FIGURE 10 – Matrice des p-valeurs obtenues en comparant le chromosome 1 aux stades ES et CN (Résolution de 50 kB)

Pour conclure, j'ai sauvegardé l'intégralité de mon travail sur le cluster GenoToul-Bioinfo

afin qu'il soit accessible à Elise, Nathalie, ainsi qu'à toutes les personnes impliquées dans cette thèse. Bien que cela ne représente qu'une petite portion du travail requis pour une thèse, je pense et j'espère sincèrement qu'il contribuera positivement à la dynamique de celle-ci.

3 Bilan

3.1 Bilan professionnel

D'un point de vue professionnel ce stage m'a enrichi. J'ai découvert l'organisation d'une structure collaborative comprenant plusieurs centaines d'employés (pour le site d'Auzeville) ainsi que celui d'une équipe de recherche et de thèse. J'ai appris l'importance de communiquer de manière efficace pour à la fois ne pas faire perdre de temps à ses collaborateurs mais aussi afin d'avoir des réponses qui nous aident réellement et instantanément. La recherche demande deux qualités que j'ai dû développer : une curiosité inépuisable et un esprit de synthèse affûté. Pour la première, cela signifie qu'il faut chercher à comprendre son travail, à le remettre constamment en question et à explorer différents angles de recherche quand cela est possible. Néanmoins, il faut faire attention à ne pas se perdre et à toujours commencer par les angles d'attaque les plus pertinents et ceux qui ont le plus de chances d'aboutir. Mon stage a duré 7 semaines ce qui est assez court du point de vue de la recherche et j'étais donc assez cadré dans mon travail. Mais même dans mon cas, j'avais un grand nombre de possibilités dans la manière d'appréhender les choses et j'ai donc dû juger les plus appropriées à chaque étape de mon travail. J'ai énormément apprécié cette responsabilité car, bien que limitée, elle me permettait de m'approprier pleinement mes recherches et de leur donner du sens à mes yeux.

Sur le plan technique, j'ai appliqué les connaissances acquises en première année à l'ENSAE. Que ce soit en probabilité, en statistique ou en programmation, l'ensemble de mon travail reposait sur ces bases solides. Néanmoins, j'ai au fur et à mesure découvert le décalage entre la théorie et la pratique. Les cours de l'ENSAE enseignent parfaitement comment écrire et comprendre des programmes mais tout cela dans un cadre précis et défini. Or dans la réalité du monde professionnel, cela ne représente qu'une partie de la programmation. Elle nécessite une aptitude à apprendre rapidement de nouveaux langages de programmation ou à se servir de nouvelles bibliothèques logicielles, mais aussi tous le travail d'installation des logiciels utiles à la recherche en statistique et qui occupe une part non négligeable du métier de chercheur.

3.2 Bilan personnel

D'un point de vue personnel, ce stage a été une réussite. Tout d'abord, j'ai grandement apprécié l'ambiance générale au sein du MIAT. J'ai tout d'abord, le premier vendredi de mon stage, participé à la journée des stagiaires. C'est un évènement annuel, organisé par le MIAT, au cours duquel les stagiaires présentent leur travail et qui se finit par un repas partagé auquel tout le monde est convié, même les anciennes personnes ayant travaillé au MIAT. Cela m'a permis de m'approprier mon sujet très rapidement car pour le présenter, il faut d'abord le comprendre

en profondeur. De plus, il a favorisé mon intégration puisque j'ai aussi pu me présenter ainsi que mon parcours. Outre cela, j'ai participé aux Tolosanes qui est un tournoi de sports entre les différentes unités de INRAE et j'y ai donc représenté le MIAT. Ce fut un moment très convivial durant lequel j'ai pu discuter avec des gens que je ne connaissais pas et passer un moment plus informel avec les membres de mon unité. En définitive, j'ai particulièrement apprécié l'ambiance générale du MIAT. Que ce soit les stagiaires, les doctorants, les chercheurs ou tout autre personne travaillant au MIAT, tout le monde était accessible. J'ai pu discuter avec des personnes ayant des parcours très variés. De plus, il y avait une grande diversité culturelle avec des gens venant de beaucoup de pays et même de continents différents. Cela est un véritable avantage d'avoir été dans un tel environnement, aussi bien pour mon bien-être personnel et pour mon ouverture au monde que pour ma productivité.

Ce stage m'a aussi permis de réfléchir à mon projet professionnel et de le faire évoluer. Trois voies m'intéressaient en entrant à l'ENSAE : la recherche en mathématiques appliquées, l'enseignement et la data science appliquée au sport. Ce stage m'a permis de découvrir la première qui est souvent liée aux deux autres. En effet, avoir fait une thèse aide beaucoup pour pouvoir enseigner dans le supérieur. Quand aux métiers de la data science appliquée au sport, ils sont souvent à la croisée entre l'ingénierie et la recherche. Lors de ce stage, je me suis rendu compte que la recherche n'était peut-être pas le milieu dans lequel je m'épanouirais le plus. Bien que j'apprécie grandement de chercher des solutions à des problèmes en élaborant une méthode particulière, en innovant et en coopérant, cela ne représente qu'une partie du travail de chercheur. Il y a toute une face cachée de bureaucratie et de formalisation de ses recherches qui m'a moins plus. Je ne regrette en aucun cas mon stage et je n'exclus pas la possibilité de faire une thèse mais je ne me vois pas faire de la recherche toute ma carrière.

Conclusion

Avec ce stage, j'ai exploré, à la manière d'un chercheur, mon appétence pour la recherche en statistiques appliquées. Je l'ai entamé avec des idées préconçues sur le monde de la recherche et en suis ressorti avec des certitudes et des enseignements concrets. Ces sept semaines ont été très enrichissantes. J'ai découvert le métier de chercheur au sens large, tout en échangeant avec des personnes aux parcours variés. J'ai à la fois mis en pratique et approfondi mes compétences en statistiques et en programmation acquises à l'ENSAE. Ce n'était pas ma première expérience professionnelle, mais la première véritablement en lien avec mes études. Ce stage n'était donc pas une première ouverture au monde professionnel, mais plutôt une découverte approfondie d'un métier qui peut m'être accessible à la fin de mon cursus. En définitive, cette immersion dans le monde de la recherche a conforté mon projet professionnel et a été une expérience déterminante.