Développement de nouveaux outils pour l'intégration des données du ChIP-Seq et leurs applications pour l'étude du contrôle de la transcription

Thèse

Charles Joly Beauparlant

Doctorat en Physiologie-Endocrinologie
Philosophiæ doctor (Ph.D.)

Québec, Canada

Développement de nouveaux outils pour l'intégration des données du ChIP-Seq et leurs applications pour l'étude du contrôle de la transcription

Thèse

Charles Joly Beauparlant

Sous la direction de:

Arnaud Droit, directeur de recherche
Jacques Corbeil, codirecteur de recherche

Résumé

Les progrès fulgurants des technologies de séquençage permettent de développer des projets de recherche très complexes. De plus, les consortiums internationaux tels qu'ENCODE, Roadmap Epigenomics et Fantom offrent publiquement de vastes jeux de donnés à la communauté scientifique. Ainsi, mon projet de recherche au doctorat a pour but de développer de nouvelles approches bioinformatiques afin d'analyser efficacement les données génomiques de type ChIP-Seq pour cibler les changements dans les patrons d'interactions entre les protéines et l'ADN. De nouveaux outils R tels `ENCODExplorer` et `FantomTSS` ont donc été développés afin de faciliter l'intégration des données publiques. De plus, l'outil `metagene`, développé dans le cadre de mon doctorat, permet de comparer les patrons d'enrichissement des protéines interagissant avec l'ADN. Il extrait efficacement la couverture des régions génomiques, normalise le signal et d'utilise les contrôles pour retirer le bruit de fond. Il produit des graphiques pour comparer visuellement les facteurs et conditions et offre des outils statistiques pour cibler les profils significativement différents. Afin de valider mon approche expérimentale, j'ai analysé une centaine de jeux de données de ChIP-Seq de la lignée GM12878 pour étudier les profils d'enrichissement au niveau des amplificateurs et des promoteurs en fonction de leur activité transcriptionnelle. Cette étude a ciblé deux modes de recrutement distincts, soit *l'effet gradient* et *l'effet seuil*. Face à la complexité et la quantité de données disponibles, il est essentiel de développer de nouvelles approches méthodologiques et statistiques afin d'améliorer notre compréhension des mécanismes biologiques. `ENCODExplorer` et `metagene` sont disponibles sur Bioconductor.

Abstract

Recent progress in sequencing technologies opened the possibility of performing very complex research experiments. Combined with the vast public datasets produced by intenational consortiums such as ENCODE, Roadmap Epigenomics and Fantoms, the amount of data to process can be daunting. The goal of my doctoral project is to develop new bioinformatic approaches to facilitate the integration of ChIP-Seq data for the study of the dynamic of the interactions between proteins and DNA. New tools such as `ENCODExplorer` and `FantomTSS` were developped in R to make the publicly available datasets easier to integrate. Futhermore, the `metagene` package allows the comparison of enrichment patterns of DNA-interacting proteins. This package efficiently extracts read coverage from genomic regions of interest, normalize the signal and uses controls to remove background noise. The main functionnality of the `metagene` package is to visually compare enrichment profiles from multiple groups of genomic regions and to offer statistical tools to caracterize and compare those profiles. To validate my experimental approach, I used over a hundred datasets from the GM12878 cell line produced by the ENCODE consortium to study the enrichment profiles of transcription factors and histones in enhnacer and promoter regions. I was able to define two distinct recruitment patterns: the *gradient effect* and the *threshold effect*. With the ever growing complexity of genomic datasets, it is essential to develop new methodotical approaches to allow a better understanding of the underlying biological processes. `ENCODExplorer` and `metagene` are both available on Bioconductor.

Table des matières

Liste des tableaux

Liste des figures

À Valérie, pour m'avoir aidé à garder le cap et à mes parents pour leur soutien indéfectible.

It's dangerous to go alone!
Take this.

Old man, Legend of Zelda

Remerciements

Je tiens tout d'abord à remercier mon directeur de recherche, Dr Arnaud Droit, qui m'a donné l'opportunité de travailler dans une équipe dynamique à la pointe de la bioinformatique. J'ai commencé mon doctorat avec des bases en informatique et je le termine en tant que bioinformaticien. Les projets de recherche, les collaborations, la participation à de nombreux congrès ainsi que 4 années d'enseignement m'ont permis d'acquérir les compétences essentielles pour être un chercheur productif dans le monde de la bioinformatique. J'ai énormément appris sous sa direction. Mes années en tant qu'étudiant au Doctorat auront été très productives et j'ai confiance que les prochaines années le seront autant.

Merci également à mon co-directeur Dr Jacques Corbeil - professeur titulaire à l'Université Laval - pour m'avoir permis de découvrir de nombreuses approches d'analyses telles que l'assemblage *de novo* et la métagénomique. Sa vision de la recherche est une source d'inspiration.

J'aimerais aussi remercier mes collègues pour les nombreux conseils et pour la motivation. Leur expertise et leur bonne humeur font du laboratoire un endroit où il est plaisant de travailler. Merci !

Merci finalement au Fond de Recherche du Québec - Santé (FRQS) pour la bourse de formation au doctorat qui a permis de financer ma formation.

Avant-propos

Ce document est une thèse avec insertion d'articles. Cette thèse présente l'état de mes travaux de doctorats dont le but principal était le développement d'outils pour l'intégration des données génomiques dans le domaine de la recherche sur la régulation de la transcription, provenant principalement de la technique d'immunoprécipitation de la chromatine suivie de séquençage de nouvelle génération (ChIP-Seq).

Les articles insérés sont les suivants :

— *NGS++ : a library for rapid prototyping of epigenomics software tools.* publié dans la revue *Oxford Bioinformatic* en 2013.
— *metagene Profiles Analyses Reveal Regulatory Element's Factor-Specific Recruitment Patterns* publié dans *Plos Computational Biology* en 2016.

Contributions `NGS++`

Je suis co-premier auteur de l'article portant sur la librairie NGS++. Alexei Nordell Markovits a fait le design de la librairie et a codé la première version de la librairie. Par la suite, nous avons travaillé sur la finalisation de l'interface, la mise en place des tests unitaires, la documentation et l'écriture de l'article. Nicolas Toupin a été impliqué dans le développement des premières versions de la librairie. Le travail a été réalisé sous la supervision de Shengrui Wang, Arnaud Droit et Nicolas Gevry.

Contributions `metagene`

Pour l'article présentant l'outil `metagene`, je suis la personne responsable du développement et de la maintenance de l'outil. Je suis donc le contributeur principal pour le développement du code de l'outil, des tests unitaires et de la documentation. Fabien Lamaze a été impliqué dans le design de l'interface de l'outil, pour la mise en place de

l'approche statistique par booststrap et pour les tests de l'outil sur des données réelles. Astrid Deschênes a participé à la section pour l'affichage graphique, à la documentation et à la mise en place des tests unitaires. Rawane Samb a été impliqué dans l'élaboration de la méthode de comparaison par bootstrap. Pascal Belleau a contribué à l'élaboration des tests statistiques de l'outil. Audrey Lemaçon a développé l'interface graphique de `metagene`, nommé `Imetagene`. L'expertise de Steve Bilodeau et de Fabien Lamaze a été importante pour l'élaboration du protocole expérimental utilisé pour valider la pertinence de `metagene` dans un contexte expérimental réel et pour la rédaction de la section correspondante de l'article. Arnaud Droit a supervisé l'ensemble des étapes de développement et de la publication.

Les travaux présentés dans cette thèse ont été financés par le Fonds de Recherche du Québec - Santé (FRQS) par le biais d'une bourse de formation de doctorat FF1-2D (#28292 ; septembre 2013 à août 2016).

Chapitre 1

Introduction

1.1 Mise en situation

Le visage de la recherche dans le domaine de l'étude de la régulation de la transcription a changé énormément au cours de la dernière décennie. Grâce aux progrès techniques, notamment au niveau des technologies de séquençage, il est maintenant possible de questionner rapidement l'ensemble du génome. Par exemple, on peut mesurer le niveau d'expression de tous les gènes d'un type cellulaire donné en une seule expérience ou bien de cibler toutes les régions où la chromatine est dans un état ouvert. L'analyse des interactions entre les protéines et l'ADN n'a également pas été épargnée. En effet, en couplant la technique d'immunoprécipitation de la chromatine au séquençage de nouvelle génération, on peut maintenant déterminer la position des sites d'interaction d'une protéine d'intérêt avec le génome pour une condition expérimentale donnée. On nomme cette méthode le ChIP-Seq (ImmunoPrécipitation de la Chromatine suivie du séquençage de nouvelle génération).

La quantité de données qu'il est possible de générer nous amène par contre de nouvelles problématiques. Outre la difficulté technique de manipuler de nombreux fichiers volumineux, nous notons aussi des problèmes lorsque vient le temps de comparer plusieurs échantillons. Par exemple, le nombre de sites de liaison pour une protéine donnée déterminé par ChIP-Seq peut varier de quelques milliers à des centaines de milliers. Il faut donc trouver des méthodes pour simplifier la représentation d'un grand nombre de régions. Le défi est donc de trouver des approches pour visualiser efficacement des milliers d'enrichissements en un seul coup d'oeil pour permettre aux chercheurs de cibler rapidement des changements dans les profils d'enrichissement. Il faut aussi être en mesure de déterminer si des différences subtiles entre les profils d'enrichissement sont

significatives. Il importe donc de mettre en place des approches statistiques robustes pour évaluer les similitudes et les différences entre les patrons d'enrichissement.

Un autre problème majeur qui rend délicate l'intégration des jeux de données est la présence de biais dans les données qui peuvent provenir de différentes sources. Parmi ces sources de biais, on note entre autres des points comme l'utilisation de lots de produits différents, des variations dans les protocoles expérimentaux et l'utilisation d'appareils de séquençage différents. Les outils qui cherchent à comparer des résultats provenant de différentes expériences doivent donc tenir compte de ces biais et offrir des solutions permettant de normaliser les données.

Le projet principal de mon doctorat est donc de développer ce type d'outils, qui permettent de travailler facilement avec des jeux de données volumineux et pour obtenir des résultats biologiquement pertinents. Au-delà de l'intégration des données produites par chaque équipe, j'ai aussi mis en place une infrastructure qui facilite la recherche de bases de données publiques offertes par des consortiums internationaux. Cette suite d'outils facilite donc la recherche scientifique des données de génomiques, tout particulièrement pour l'étude de la régulation de la transcription.

1.2 Description des chapitres

Dans le chapitre 2 de cette thèse, je présente les différentes technologies de séquençage en mettant l'accent sur les technologies de deuxième génération, car elles sont à privilégier lors des analyses de ChIP-Seq. Dans cette section, je présente un bref historique des technologies de séquençage automatisé puis je décortique les forces et les limites de chaque approche.

Au chapitre 3, je fais un survol des mécanismes de régulation de la transcription en discutant des promoteurs et des amplificateurs. Je me penche également sur les mécanismes de recrutement des facteurs de transcription dans ces deux types de régions génomiques. Cette section permet d'introduire certains concepts qui seront utiles pour la section 7.8 de l'article sur les analyses réalisées pour démontrer la pertinence de l'outil `metagene`.

Pour le chapitre 4, je présente les bases de données qui ont été utilisées pour l'article de l'outil `metagene`. Je discute tout d'abord du projet ENCODE pour lequel j'ai développé le paquet R `ENCODExplorer`. J'aborde ensuite l'historique des projets du consortium Fantom qui ont mené aux données produites à la 5e phase et qui ont été utilisés dans

le cadre de mes projets.

Le chapitre 5 se penche sur la technologie de ChIP-Seq. Dans ce chapitre, je couvre toutes les étapes de l'approche expérimentale de ce type d'analyse, incluant les critères pour avoir un bon design expérimental. Ensuite, je discute des protocoles d'analyse classique pour ce type de données de manière à mieux situer mes outils au sein des logiciels d'analyses généralement utilisés. Finalement, je discute des différentes approches disponibles pour vérifier la qualité des données produites lors d'une expérience de ChIP-Seq.

C'est dans le chapitre 6 que je présente formellement les outils développés dans le cadre de mon doctorat pour l'étude de la régulation de la transcription et pour faciliter l'utilisation des données produites par les consortiums internationaux tels que Fantom, ENCODE et Roadmaps Epigenomics. Les outils sont `metagene`, `ENCODExplorer`, `FantomTSS.hg19` et `FantomEnhancers.hg19`.

Les chapitres 7 et 8 correspondent à mes articles sur l'outil `metagene` et sur la librairie `NGS++`, respectivement. Au chapitre 9, je décris les autres projets sur lesquels j'ai travaillé lors de mon doctorat. Finalement, je fais un retour sur mes résultats et je présente les voies de recherche que je souhaite explorer pour la poursuite de mes travaux dans le chapitre 10.

Chapitre 2

Le séquençage de nouvelle génération

La recherche en génomique est intimement liée au développement des techniques de séquençage de l'ADN. Chaque nouvelle génération de séquenceurs ouvre la voie à de nouveaux types d'analyse qui n'étaient pas possibles précédemment. Dans cette section, je vais faire un survol des principales technologies de séquençage actuellement disponible (voir la figure 2.1) et je vais discuter de l'approche à privilégier pour l'étude des données d'immunoprécipitation de la chromatine couplée au séquençage de nouvelle génération.

Les appareils de séquençage sont séparés en trois générations : 1^{ière}, 2ième et 3ième génération. Chaque génération se distingue principalement par le nombre de séquences produites et par leur taille. Le terme séquençage de nouvelle génération est souvent utilisé bien qu'il soit moins précis. Il correspond à toutes les technologies qui ont été développées après la 1^{ière} génération d'appareils de séquençage et inclue donc tous les appareils de 2ième et 3ième génération.

2.1 Les appareils de séquençage de première génération

Les appareils de séquençage de première génération qui sont apparus sur le marché en 1987 (Liu et al., 2012) utilisent la technologie de Sanger (Sanger et al., 1977) et ont permis d'augmenter considérablement la qualité et la quantité de séquence qu'il était possible de produire. C'est notamment grâce à cette technologie que la première version complète du génome humain a pu être complétée (Lander et al., 2001).

Cette technologie combine l'utilisation de didésoxyribonucléotides associés à des molécules fluorescentes (Smith et al., 1985) pour distinguer les nucléotides et l'électrophorèse

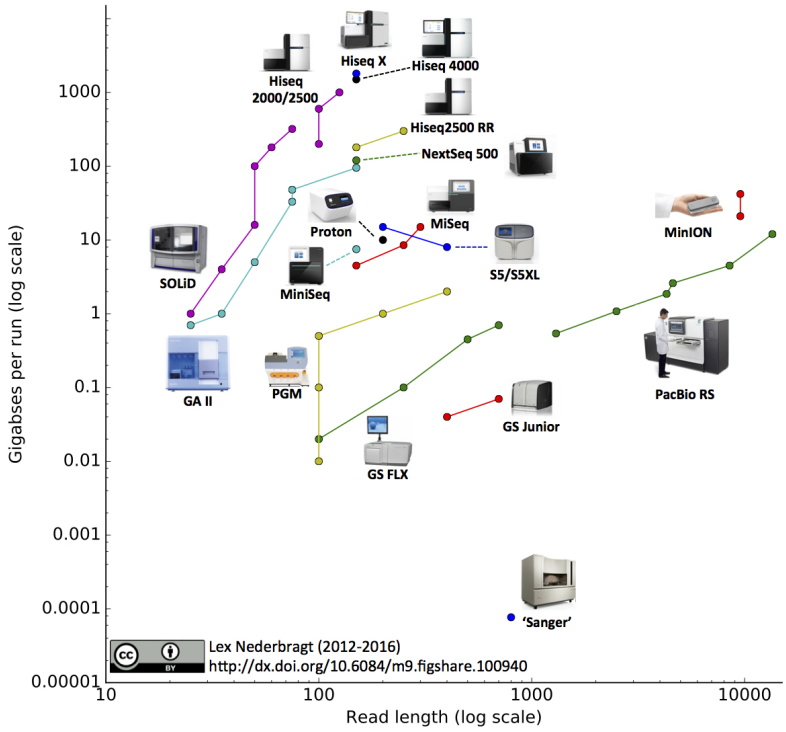


FIGURE 2.1 – Survol des technologies de séquençage. Comparaison de la taille des séquences produites par rapport au nombre de séquence produite basé sur les données fournies par les fabricants. Adapté avec permission de la figure 'Developments in high throughput sequencing' par Lex Nederbragt (https ://dx.doi.org/10.6084/m9.figshare.100940.v9) license CC-BY (https ://creativecommons.org/licenses/by/4.0/legalcode). Téléchargée le 3 août 2016.

capillaire (Jorgenson and Lukacs, 1983) pour séparer les fragments d'ADN selon leur taille (Cohen et al., 1988). Grâce au multiplexage des capillaires (Huang et al., 1992), les appareils modernes peuvent analyser 96 échantillons en parallèle. Par exemple, le 3730xl peut produire des séquences d'une taille variant d'environ 500 à 1000 nucléotides par capillaire pour un total d'un peu moins de 2 millions de nucléotides par jour.

Le séquençage avec cette technologie se produit en 3 étapes (voir figure 2.2) :

1. Amplification de la région génomique d'intérêt.
2. Ajout des molécules fluorescentes.
3. Séquençage par électrophorèse.

Le nombre de réactions en chaîne de la polymérase (PCR) qu'il est possible d'effectuer

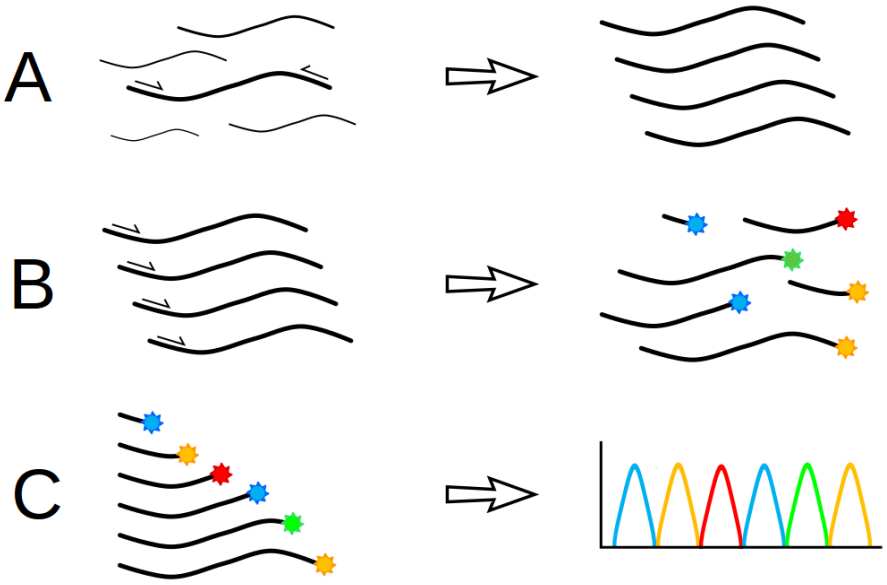


FIGURE 2.2 – Séquençage de première génération. (A) Lors de la première étape, on amplifie la région d'intérêt par réaction en chaîne de la polymérase (PCR) pour s'assurer qu'elle est surreprésentée par rapport aux autres endroits dans le génome et pour en avoir une quantité suffisante. Cette amplification est réalisée en utilisant des amorces se situant de part et d'autre de la région d'intérêt. (B) Le but de la deuxième étape est d'ajouter les nucléotides fluorescents à la fin des séquences pour pouvoir les détecter lors de l'étape de séquençage. La technique utilisée est une amplification PCR dans une seule direction en utilisant un mélange de nucléotides et de didésoxyribonucléotides fluorescents. (C) Pour la troisième étape, il faut séparer les fragments selon leur taille et de lire la fluorescence du dernier nucléotide. À chaque fois qu'un fragment aura terminé sa migration, la molécule fluorescente attachée au didésoxyribonucléotide sera excitée pour produire une lumière qui sera captée par un détecteur. Le résultat final est un spectre qui correspond à la séquence d'ADN ciblée en (A).

en parallèle pour cette technologie est limité à quelques centaines (96 ou 384). De plus, il n'est pas possible d'analyser plus d'un échantillon à la fois dans un capillaire. Il est donc extrêmement laborieux et coûteux de produire les millions de séquences nécessaires pour des études à l'échelle génomique.

2.2 Les appareils de séquençage de deuxième génération

Les technologies de 2ième génération se distinguent d'une part par le nombre de réactions PCR effectuées en parallèle qui est largement supérieur et d'autres parts par les approches utilisées pour l'étape de séquençage qui permettent d'analyser des millions (et parfois même des milliards) de séquences en une seule analyse.

2.2.1 Amplification PCR

La première difficulté rencontrée par ces technologies est que le signal produit par une seule molécule n'est pas suffisamment fort pour être détecté. Il est donc nécessaire d'amplifier les fragments d'ADN étudiés. Il faut non seulement amplifier chaque molécule des milliers de fois, mais il faut s'assurer que chaque réaction contient une seule molécule (amplification clonale).

La première approche développée pour résoudre cette problématique fut la PCR par émulsion. Le principe de cette technique est d'associer chaque fragment à amplifier à une microbille, à raison d'un fragment d'ADN par bille. Ensuite, on isole chaque bille dans une goutte produite par émulsion qui contient tous les réactifs nécessaires pour l'amplification (Dressman et al., 2003, Williams et al. (2006)). En contrôlant correctement le ratio ADN/bille, on s'assure d'avoir un seul fragment par bille qui sera amplifié des milliers de fois (voir la figure 2.3,A).

La seconde approche consiste à associer les fragments d'ADN sur un support solide puis à amplifier chaque fragment localement pour former des *clusters* de molécules clonales (Adessi et al., 2000, Fedurco et al. (2006)). Le principe de cette approche est de fixer préalablement les amorces sur le support solide. Chaque molécule d'ADN devra former un pont pour être amplifiée, ce qui permettra de conserver les molécules amplifiées au même endroit sur la plaque (voir la figure 2.3,A).

2.2.2 Séquençage

La première technologie a avoir été mise sur le marché est le pyroséquençage par la compagnie *454 Life Sciences* en 2005. Cette technologie utilise la technique d'amplification par émulsion présentée dans la section précédente.

Après l'amplification, les billes sont ajoutées à une plaque contenant des millions de

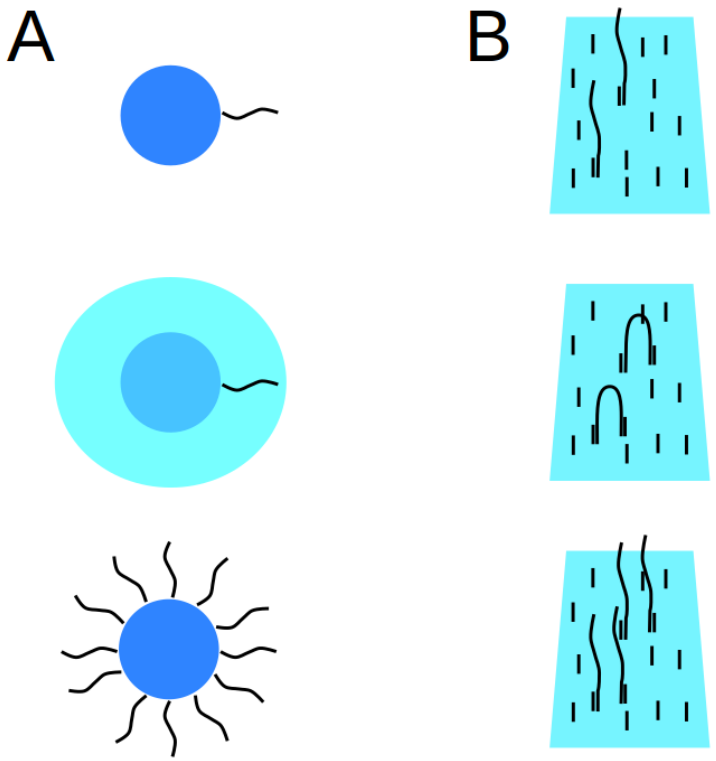


FIGURE 2.3 – Approches d'amplification hautement parallèles. (A) PCR par émulsion. Chaque fragment d'ADN est associé à une microbille. Chaque bille se retrouve ensuite dans une goutte d'émulsion contenant tous les réactifs nécessaires pour l'amplification. Après l'étape d'amplification, chaque bille contiendra des milliers de copies du fragment original. (B) PCR sur support solide. Les fragments d'ADN sont fixés sur une lame. Les amorces nécessaires pour l'amplification sont également fixées sur la lame ce qui oblige les fragments d'ADN à former un pont (*bridge*) et permet de s'assurer que les copies sont toutes situées à proximité les unes des autres.

puits pouvant contenir une seule bille. C'est dans cette plaque qu'aura lieu le pyroséquençage (Ronaghi et al., 1998, Ronaghi et al. (1996)). Cette technique utilise la molécule de phosphate inorganique relâchée lors de la polymérisation pour générer de l'ATP. Cette dernière est ensuite utilisée par une luciférase pour produire de la lumière qui sera détectée par une caméra. Puisque tous les nucléotides produisent le même signal, un seul nucléotide sera utilisé à chaque cycle. En présence d'homopolymères (une suite de nucléotides identiques dans la séquence), l'intensité du signal augmentera de manière proportionnelle au nombre de nucléotides.

Par la suite, la technologie de séquençage par synthèse a été mise au point par Solexa (acquis par Illumina en 2007). Contrairement au pyroséquençage qui utilise des mi-

crobilles, cette technique utilise plutôt un support solide. Pour l'étape de séquençage, cette technologie utilise des nucléotides terminateurs fluorescents réversibles (Bentley et al., 2008). Grâce à ces terminateurs, il est possible d'ajouter un seul nucléotide à chaque cycle. De plus, puisque chaque nucléotide est associé à une molécule fluorescente différente, il est possible d'utiliser tous les nucléotides à chaque cycle. Le nombre de cycles pour une analyse correspondra donc à la taille des séquences. Cette technologie est présentement la technologie de séquençage la plus populaire.

La technologie de séquençage par semi-conducteur a été développée par la compagnie Ion Torrent (Rothberg et al., 2011) et fonctionne de manière similaire à la technologie développée par la compagnie *454 Life Sciences*. En effet, cette technique requiert tout d'abord une amplification de l'ADN par PCR à émulsion. Ensuite, les billes sont mises dans des puits de la même manière que pour le pyroséquençage. Lorsqu'un nucléotide est ajouté lors de la polymérisation, il y a relâchement d'un nombre de protons proportionnel au nombre de nucléotides ajoutés qui va causer un changement de pH. Un capteur placé au fond de chaque puits va mesurer le changement de pH à chaque cycle. L'avantage principal de cette approche est qu'il est beaucoup plus rapide de mesurer un changement de pH qu'une intensité lumineuse. Cette technologie permet de compléter une analyse plus rapidement que les autres technologies de 2ième génération.

La dernière technologie de séquençage de deuxième génération qui a fait son apparition sur le marché en 2007 est la technologie de séquençage par ligation (Mardis, 2008) de la compagnie ABI SoLiD. Comme son nom l'indique, c'est technique se distingue des autres par le fait que le séquençage n'est pas réalisé par polymérisation de l'ADN. Au niveau de l'amplification, cette technologie utilise la même approche que pour le pyroséquençage et le séquençage par semi-conducteurs, c'est-à-dire le PCR à émulsion. L'approche utilisée pour le séquençage est plus complexe que pour les autres technologies et requiert plusieurs cycles de ligation de 5-mers successifs. De plus, les molécules fluorescentes sont associées à des dinucléotides plutôt qu'à un seul nucléotide, ce qui crée le concept d'espace de couleur. L'intérêt de cette approche est qu'elle permet de corriger les séquences erronées, car chaque nucléotide de la séquence est mesuré à deux reprises.

Toutes les technologies ont des forces et des faiblesses. Du point de vue de la bioinformatique, les paramètres les plus importants sont le nombre de séquence produite, la taille des séquences et le profil d'erreur de chaque technologie. En termes de quantité de séquence, la technologie de séquençage par synthèse d'Illumina est de loin la

meilleure. Elle offre aussi l'avantage de ne pas produire d'erreurs systématiques (par opposition à des erreurs aléatoires) qui sont souvent impossibles à corriger. Les technologies qui permettent d'intégrer plus d'un nucléotide par cycle (pyroséquençage et séquençage par semi-conducteur) vont générer des erreurs systématiquement lorsqu'ils se retrouvent face des homopolymères (par exemple la séquence AAAAAA) pour lesquels ils n'arriveront pas à estimer correctement le nombre de répétitions au-delà d'un certain seuil (Metzker, 2010). La technologie d'Illumina est actuellement dominante sur le marché, bien que le séquençage par semi-conducteur pourrait se démarquer dans la niche du séquençage en clinique grâce à sa vitesse et sa simplicité. L'apparition des appareils de troisième génération a sonné le glas pour le pyroséquençage qui se démarquait d'Illumina uniquement par la taille de ses séquences. Depuis 2016, il n'y a donc plus de soutien technique pour cette approche. Finalement, la technologie SOLiD a connu des ratés lors de sa mise en marché et n'a jamais réussi à aller chercher des parts importantes dans le marché des séquenceurs de 2ième génération.

2.3 Les appareils de séquençage de troisième génération

Au cours des dernières années, une nouvelle génération d'appareil de séquençage a fait son apparition : les appareils de 3ième génération. Ces appareils se distinguent de ceux de la 2ième génération par le nombre de séquences produites et par la taille des séquences. Comparativement aux appareils de deuxième génération qui peuvent produire des centaines de millions de séquences en une analyse, les appareils de troisième génération en produisent beaucoup moins (de l'ordre des dizaines de milliers). En contrepartie, la taille des séquences est plus grande d'au moins 2 ordres de grandeur.

Actuellement, on retrouve uniquement une technologie de troisième génération sur le marché : PacBio. La version présentement disponible (RS II) produit environ 55 000 séquences par analyse et la taille moyenne et maximale des séquences est de 8kb et 60kb respectivement. La prochaine version de l'appareil (Sequel) devrait permettre de produire sept fois plus de séquences ayant les mêmes propriétés, selon les spécifications de l'entreprise.

La méthode de PacBio, le séquençage SMRT (*Single Molecule, Real-Time*), permet d'analyser la polymérisation en temps réel (Eid et al., 2009). Cette approche consiste à circulariser les fragments d'ADN et à les associer à une polymérase. Chaque polymérase

est ensuite déposée dans son propre puits sur une plaque. Grâce à l'utilisation de nucléotides associés à des molécules fluorescentes et à des caméras haute performance, la polymérisation est analysée en temps réel. En effet, la présence d'une molécule fluorescente associée à chaque nucléotide ralentit la vitesse de la polymérisation suffisamment pour permettre à une caméra spécialisée de discerner le signal du bruit de fond ambiant.

Oxford Nanopore est une technologie qui pourrait grandement influencer le monde de la recherche, si elle arrive à livrer les résultats promis. D'une part, des séquences de plus de 100kb ont été rapportées par les laboratoires qui ont un accès à la technologie à des fins de développement. Mais surtout, la taille de l'appareil est semblable à celle d'un disque dur portable et son coût est nettement moins élevé que celui des autres appareils. Cette technologie pourrait donc éventuellement permettre à un plus grand nombre d'équipes de recherche de s'équiper de séquenceur sans avoir à acheter des appareils qui peuvent coûter des centaines de milliers de dollars.

Cette technologie utilise des nanopores à travers desquels les fragments d'ADN passeront. À chaque fois qu'un nucléotide passe à travers la membrane en utilisant le nanopore, on peut observer un changement au niveau du potentiel de la membrane. C'est ce changement de potentiel qui est mesuré par l'appareil. Étant donné que l'appareil n'est pas encore officiellement disponible sur le marché, il n'est pas possible de connaître les statistiques sur la taille des séquences attendues, sur le nombre de séquences produites et sur la qualité moyenne des séquences.

Les appareils de troisième génération ont le potentiel de changer grandement plusieurs champs de recherche en génomique, tout particulièrement au niveau du séquençage des génomes. Un autre domaine qui va profiter de ces technologies de séquençage est la transcriptomique. En effet, il est actuellement très difficile de reconstruire les transcrits alternatifs à partir de courtes séquences. Le séquençage complet des isoformes (Iso-seq) pourra donc faciliter grandement l'analyse des transcrits alternatifs en produisant une seule séquence par transcrit, ce qui évitera de devoir utiliser des approches statistiques complexes qui ne garantissent pas de résoudre le problème.

Par contre, dans le cadre l'étude de la régulation de la transcription à l'échelle génomique, notamment pour les analyses de type ChIP-Seq (immunoprécipitation de la chromatine suivie de séquençage de nouvelle-génération), l'utilisation des appareils de troisième génération ne devrait pas être d'une grande influence, car les problématiques rencontrées dans ce domaine ne proviennent pas de la taille des séquences. La raison principale est que le gain dans la précision de l'alignement des séquences plus longues

est généralement trop faible pour contre-balancer la diminution importante du nombre de séquences produites. Il est donc préférable de produire un plus grand nombre de séquence et des réplicats biologiques. Ces aspects seront traités plus en détail dans les prochaines sections.

Ce tableau [1] résume les statistiques des différentes technologies de nouvelle-génération :

TABLEAU 2.1 – Statistiques des principales technologies de séquençage de nouvelle-génération

Technologie	Appareil	Volume (Gb)	Qualité
454	GS FLX+	0.7	NA
454	GS Jr.	0.035	NA
Illumina	HiSeq 4000	1300-1500	>75% des bases >Q30
Illumina	MiSeq	13.2-15	>70% des bases >Q30
SOLiD	5500xl Wildfire	240	65-70% après filtre de qualité
SOLiD	5500 Wildfire	160	65-70% après filtre de qualité
Ion	Ion Proton	15	60-80 millions de séquences après filtre
ThermoFisher	Ion Torrent	10-15	99% de précision mesurée/alignée
PacBio	RS II	0.5-1	>99.999% QV 50
PacBio	Sequel	5-8	>99.999% QV 50
Oxford Nanopore	MinION	10-20	NA

1. Les valeurs ont été obtenues à partir des pages web suivantes : 454, Illumina HiSeq 4000, Illumina MiSeq, SOLiD, Ion Torrent, Ion Proton, RS II, Sequel et MinION. Gb : Gigabases. Q30 : Score phred de 30. Q50 : Score phred de 50.

Chapitre 3

La régulation de la transcription

Il est important que les cellules contrôlent la production des protéines pour leur permettre de bien remplir toutes leurs fonctions et pour répondre correctement aux différents stress auxquels elles sont confrontées. Un déséquilibre au niveau de ce contrôle peut avoir des impacts majeurs sur l'homéostasie de la cellule, des organes et parfois même de l'organisme. Il existe plusieurs niveaux de contrôle des protéines tels que la régulation de la transcription, la régulation de la traduction, les modifications post-traductionnelles et les mécanismes de dégradation des ARNm et des protéines. Le contrôle de la transcription est le mécanisme le plus important pour déterminer l'abondance des protéines dans la cellule (Li and Biggin, 2015).

Il est maintenant bien accepté que des mutations retrouvées dans les régions régulatrices du génome sont associées au développement de plusieurs maladies (Corradin and Scacheri, 2014). Il est donc primordial d'en étudier les mécanismes.

3.1 Survol du contrôle de la transcription

Les éléments régulateurs en *cis* (promoteurs et amplificateurs) jouent un rôle primordial dans le contrôle de la transcription des gènes. En effet, ces éléments sont ciblés par les facteurs en *trans* pour mener à l'expression ges gènes. Bien qu'il existe d'autres classes d'éléments régulateurs tels que les isolateurs (*insulators*) (Gerasimova and Corces, 2001), les inactivateurs (*silencers*) (Ogbourne and Antalis, 1998) et d'autres mécanismes tels que le repliement de la chromatine (Li et al., 2006), je me suis concentré principalement sur les amplificateurs et les promoteurs lors de mes études doctorales. Ces deux classes d'éléments seront donc décrit plus en détail dans ce chapitre.

3.1.1 Promoteurs

Les promoteurs sont des régions régulatrices qui chevauchent les sites d'initiation de la transcription et qui contiennent des sites de liaisons pour les facteurs de transcriptions qui permettent la régulation de l'expression des gènes (Lenhard et al., 2012). C'est donc au niveau du promoteur proximal que le complexe de préinitiation doit être recruté pour permettre la transcription du gène.

Classes de promoteurs

Le modèle classique d'une région promotrice est la présence d'un seul site d'initiation de la transcription clairement définie et d'un enrichissement pour le motif de boîte TATA (Lenhard et al., 2012). Avec l'arrivée des technologies de séquençage de nouvelle génération, il est maintenant possible d'étudier les régions promotrices à l'échelle génomique sans avoir à les connaître a priori. Une des conclusions principales de ces études est que ces promoteurs "classiques" ne représentent qu'une minorité de l'ensemble des promoteurs.

Il y a trois classes de promoteurs qui se distinguent principalement par le contenu en GC, par la présence ou l'absence d'îlots CpG et par la fonction principale du gène (Lenhard et al., 2012). Par exemple, les promoteurs associés à la transcription des gènes qui sont spécifiques à un type cellulaire ont généralement un seul site d'initiation de la transcription clairement défini, un enrichissement pour le motif TATA et contiennent rarement des îlots CpG. À l'inverse, les promoteurs des gènes ubiquitaires n'ont pas un site d'initiation clairement défini, contiennent peu de motifs TATA mais beaucoup d'îlots CpG. Finalement, les gènes impliqués dans le développement contiennent de larges îlots CpG. La présence de région promotrices ayant des caractéristiques différentes suggère qu'il existe différents mécanismes d'activation de la transcription des gènes.

Les expériences qui étudient le niveau d'enrichissement de la polymérase au niveau des promoteurs ont démontré que la polymérase est fréquemment enrichie légèrement en aval des sites d'initiation de la transcription (Nechaev and Adelman, 2011).

Recrutement du complexe de préinitiation

3.1.2 Amplificateurs (*enhancers*)

Les profils d'expressions géniques doivent être correctement régulés lors de la différentiation cellulaire ou en réponse à un stress. Les amplificateurs (*enhancers*) sont des

éléments essentiels pour une activation correcte de l'expression des gènes à travers pour permettre la spécificité des types cellulaires (Andersson et al., 2014). Ils seraient impliqués dans les étapes d'initiation (Kagey et al., 2010), d'élongation (Sawado et al., 2003) et de terminaison (Plank and Dean, 2014) de la transcription. Plus spécifiquement, les amplificateurs seraient directement impliqués dans le recrutement du complexe de pré-initiation au niveau du promoteur, de retirer les marques d'histones répressives et de recruter les enzymes qui régulent les modifications d'histones (Vernimmen and Bickmore, 2015). Les activateurs jouent donc un rôle central dans les mécanismes permettant d'expliquer les patrons d'expression des gènes.

Définition des amplificateurs

Il existe plusieurs définitions des amplificateurs. La définition originale des amplificateurs est que ce sont des éléments distaux qui augmentent les niveaux de transcription indépendamment de leurs orientation, position et distance par rapport au promoteur (Banerji et al., 1981). Cette définition a été mise à jour pour également inclure le rôle des activateurs dans la répression de l'expression des gènes (Barolo and Posakony, 2002). Au niveau moléculaire, on considère que la présence des co-activateurs de la transcription p300-CBP situés à distance des sites d'initiation de la transcription (TSS) peut permettre l'identification des activateurs (Kim et al., 2010), bien que d'autres études proposent que seulement une partie des activateurs soient liés à p300/CBP (Ong and Corces, 2011).

Les amplificateurs ressemblent beaucoup aux promoteurs qui sont pauvres en GC, incluant les motifs de liaison de facteurs de transcription retrouvés (Andersson et al., 2014). De plus, comme les promoteurs, les amplificateurs actifs produisent des ARN (Kim et al., 2010). Il a été démontré que les amplificateurs peuvent être des promoteurs alternatifs (Kowalczyk et al., 2012) et que les promoteurs peuvent agir comme des amplificateurs (Li et al., 2012, Dao et al. (2017)). Certains auteurs suggèrent que ces deux types de régions régulatrices pourraient être des variants d'une même classe d'éléments (Andersson et al., 2015). Les amplificateurs sont plus conservés que ce qui serait attendu par hasard, mais moins que les régions promotrices (De Santa et al., 2010).

Les catégories d'amplificateurs

On peut généralement classer les amplificateurs en quatre catégories : inactifs, actifs, prêt (*poised*) et amorcé (*primed*) (Ernst and Kellis, 2010). Les amplificateurs inactifs se

retrouvent dans des régions hétérochromatiniennes, ne sont pas associés à des marques d'histones ou à des facteurs de transcription (Heinz et al., 2015). Les amplificateurs actifs sont activement impliqués dans la régulation d'un gène. Ils se retrouveraient principalement dans des régions hypométhylées (Plank and Dean, 2014), sensibles à la DNase I (Ong and Corces, 2011) et enrichies en H3K27ac (Vernimmen and Bickmore, 2015). Les amplificateurs amorcés (*primed*) sont dans un état qui leur permet d'être rapidement activés. Ils sont caractérisés par la présence de facteurs pionniers, des variants de nucleosomes H3.3/H2A.Z et d'un enrichissement de la marque d'histone H3K4me1 (Calo and Wysocka, 2013). Finalement, les amplificateurs prêts (*poised*) sont très similaires aux amplificateurs amorcés, mais vont contenir des modifications d'histone répressives (Ernst and Kellis, 2010). Ce type d'amplificateurs se retrouve principalement dans les cellules souches embryonnaires et ils serviraient de marque-page pour les amplificateurs qui seront activés plus tard dans la différentiation (Ernst and Kellis, 2010).

Pour permettre le passage d'un état inactif vers un état actif, les activateurs doivent être accessibles. Les facteurs pionniers seront les premiers à reconnaître les amplificateurs et permettront de préparer les régions amplificatrices. Ce type de facteur est capable de se lier à la chromatine, même dans son état fermé (hétérochromatine) et de favoriser son ouverture (Vernimmen and Bickmore, 2015). Les amplificateurs seront donc dans un état amorcé (*primed*) et pourront être rapidement activés (Kim et al., 2010).

Modèle d'action des amplificateurs

Les amplificateurs pourraient réguler la transcription selon deux modèles qui ne sont pas nécessairement mutuellement exclusifs : l'enhanceosome ou le panneau d'affichage (*billboard*) (Arnosti and Kulkarni, 2005). Selon le modèle de l'enhanceosome, les amplificateurs seraient directement impliqués dans le traitement du signal des facteurs de transcription pour le convertir en activité transcriptionnelle. Dans ce modèle, les amplificateurs seraient donc en mesure de moduler directement les niveaux de transcription des gènes associés. À l'inverse, le modèle du panneau d'affichage propose que les amplificateurs servent à présenter les facteurs de transcription et leur signal serait traité au niveau des promoteurs. Dans ce modèle, chaque amplificateur agirait comme un interrupteur et serait soit actif ou inactif (*on/off*). En réalité, il semblerait qu'on retrouvait une gamme d'intermédiaires entre ces deux modèles plutôt qu'une séparation nette (Spitz and Furlong, 2012).

Pour qu'un amplificateur actif puisse activer la transcription d'un gène, il doit tout

d'abord se trouver à proximité du promoteur. L'ordre exact des étapes d'activation n'est pas encore parfaitement connu. On sait par contre qu'une fois que les amplificateurs sont rendus disponibles par les facteurs pionniers, les autres facteurs de transcriptions pourront être recrutés (Vernimmen and Bickmore, 2015). Ces facteurs sont recrutés par la présence de motifs de liaisons retrouvés dans les amplificateurs (Ong and Corces, 2011). En plus des facteurs de transcription, la machinerie basale de transcription serait également recrutée au niveau des amplificateurs (Plank and Dean, 2014). En effet, environ 70% de la polymérase extragénique se trouverait au niveau des amplificateurs (De Santa et al., 2010). L'effet serait d'une part d'agir comme catalyseur pour accélérer le recrutement d'autres facteurs de transcription en remodelant activement la chromatine dans les régions amplificatrices (De Santa et al., 2010). Les amplificateurs sont nécessaires pour permettre le recrutement des polymérases actives (phosphorylées en sérine 2 et 5) au niveau des promoteurs et pourraient même être directement impliqués dans leur livraison (Vernimmen and Bickmore, 2015).

Recrutement de la machinerie transcriptionnelle

Le recrutement de la machinerie transcriptionnelle aux amplificateurs va également mener à une transcription de ces derniers pour produire des ARN (Kim et al., 2010). Les ARN produits sont majoritairement nucléaires, non méthylés et ont une coiffe en 5' (Andersson et al., 2014). La présence de ces ARN serait corrélée avec le niveau d'activité de l'amplificateur et du niveau transcriptionnel des gènes à proximité (Andersson et al., 2014, Kim et al. (2010)). Il n'est pas encore clair si ces ARN sont fonctionnels ou si c'est la transcription en tant que telle qui est nécessaire. Quelques études ont montré un rôle pour certains transcrits spécifiques, mais une fonction qui serait généralisable à l'ensemble des ARN d'amplificateurs n'a pas encore été démontrée (Vernimmen and Bickmore, 2015). Lorsque la transcription de ces ARN est inhibée, le repliement de la chromatine n'a pas lieu correctement (Ong and Corces, 2011, Plank and Dean (2014)). Bien que le rôle exact de ces transcrits n'est pas encore certain, la corrélation entre leur niveau d'expression et l'activité de l'amplificateur a permis de mieux les caractériser grâce à l'approche de CAGE-Seq (*Cap Analyse of Gene Expression* : analyse transcriptomique des fragments d'ARN avec coiffe en 5') utilisée par le consortium Fantom (Andersson et al., 2014).

3.2 Étude de la régulation de la transcription

Les grands consortiums internationaux, dont certains seront présentés plus en détail dans la prochaine section, ont permis des avancées importantes dans l'étude de la régulation de la transcription. Notamment, le consortium Roadmap Epigenomics a caractériser plusieurs états de la chromatine notamment au niveau des amplificateurs et des promoteurs en intégrant les données de différentes marques d'histones (Mendenhall and Bernstein, 2008). Une meilleure compréhension des liens entre les certaines maladies et l'épigénome ont pu être mis en évidence par les travaux des consortium tels qu'IHEC (*International Human Epigenome Consortium*) (Chun et al., 2016, Mandoli et al. (2016)), ENCODE (Schaub et al., 2012) et Roadmap Epigenomics (Gjoneska et al., 2015, Farh et al. (2015)).

La disponibilité d'un grand nombre de jeux de données offre une excellente opportunité pour l'étude de la transcription. En effet, plusieurs consortiums ont rendu publiques des milliers d'expériences pouvant être utilisées dans le contexte de l'étude de la transcription telle que des données de transcriptomique et d'épigénomique. Par contre, pour utiliser ces données, il est nécessaire de bien comprendre le contexte dans lequel elles ont été produites et d'avoir accès aux bons outils pour rechercher les données les plus pertinentes.

Chapitre 4

Les bases de données

Les bases de données biologiques permettent de stocker de l'information biologique de manière à ce qu'elle puisse être facilement retrouvée par les chercheurs. Il existe présentement plus de 1600 bases de données biologiques (Rigden et al., 2015). Ces bases de données peuvent entre autres contenir des génomes de références, des séquences nucléotidiques et d'acides aminés et des données expérimentales.

Grâce aux améliorations des technologies de séquençage, il est maintenant envisageable de cataloguer l'ensemble des interactions entre les protéines et l'ADN ainsi que l'ensemble des transcrits produits et ce, pour tous les types cellulaires. Les coûts associés à ce type de projet sont encore prohibitifs pour un seul laboratoire de recherche. Les équipes doivent donc se regrouper de manière à obtenir du financement et développer des lignes directrices pour la production des données et leur stockage. Certains groupes se concentrent sur l'étude des transcrits tels que le consortium Fantom.

De nombreux consortiums internationaux existent dans le domaine de l'épigénomique et de l'étude de la régulation de la transcription. Dans le domaine de la recherche en épigénomique, plusieurs regroupement ont vu le jour pour tenter de cataloguer systématiquement les modifications épigénétiques à travers l'ensemble des types cellulaires pour ensuite utiliser cette information pour mieux comprendre des mécanismes biologiques tels que la différenciation cellulaire la prolifération et la sénescence. Un joueur majeur dans ce domaine est le consortium international IHEC (*International Human Epigenome Consortium*) (Stunnenberg et al., 2016) est un regroupement de groupes de recherche tels que le CEEHRC (*Canadian Epigenetics, Environment and Health Research Consortium*) au Canada, NIH (*National Institutes of Health*) Roadmap et ENCODE aux États-Unis, et Blueprint en Europe qui se concentre sur l'étude de la

méthylation de l'ADN, des marques d'histones, des facteurs de transcription et de l'expression des ARN non codants et des microARN.

Les informations produites par ces groupes sont disponibles sous forme de base de données et des portails web sont également disponibles pour faciliter la recherche. Des interfaces accessibles programmatiquement (API : *Application Programming Interface*) sont parfois disponibles pour intégrer ces bases de données à des outils informatiques.

Lors de mon doctorat, j'ai travaillé au développement d'interface entre les bases de données et le langage R. J'ai utilisé principalement la base de données ENCODE, qui contient un grand nombre de jeux de données sur les facteurs de transcription, et la base de données Fantom5, qui a étudié l'expression des sites d'initiation de la transcription au niveau des promoteurs et des amplificateurs. Ce chapitre présentera donc l'historique et le contenu de ces 2 bases de données.

4.1 Épigénomiques

4.1.1 ENCODE

En septembre 2003, le NHGRI (*National Human Genome Research Institute*) a annoncé la création du consortium ENCODE (`https://genome.ucsc.edu/ENCODE/pilot.html`). Ce projet, lancé suite à l'achèvement du séquençage du génome humain, avait pour but de produire une encyclopédie des éléments fonctionnels du génome. Plus spécifiquement, ce consortium a produit et continue de produire des jeux de données génomiques dans un grand nombre de types cellulaires.

Le projet ENCODE se déroule en 3 phases qui permettent de s'assurer que certaines cibles ont bien été atteintes avant de poursuivre la production des données. Chaque phase sera décrite plus en détail dans les sous-sections suivantes.

ENCODE 1 (le projet pilote)

La première phase, aussi connue sous le nom d'ENCODE 1, était le projet pilote. Le but de cette étape était donc de démontrer la faisabilité du projet. Il s'agissait donc d'analyser 1% du génome humain (~30 Mb) et valider l'efficacité de diverses approches expérimentales (Consortium et al., 2004). Plutôt que de cibler une région contiguë de 30 Mb, le consortium a sélectionné 44 régions de plus petite taille (0.5-2 Mb) (Consortium et al., 2004). En plus de 8 groupes originaux, 27 groupes supplémentaires se sont ajoutés

pour participer à ce pilote (Birney et al., 2007). Une partie des régions a été choisie en ciblant la présence d'éléments fonctionnels bien caractérisés ou de la disponibilité de données comparatives alors que d'autres régions ont été choisies aléatoirement parmi des régions pour lesquelles aucun gène connu n'était présent (Consortium et al., 2004). Les analyses n'ont pas été faites sur des tissus primaires, mais plutôt sur des cellules adénocarcinome cervical (HeLa S3) et des lymphocytes B transformés par le virus d'Epstein-Barr (Birney et al., 2007). Les technologies de séquençages n'étaient pas encore très utilisées au début de la phase pilote du projet ENCODE et ce sont donc des technologies de biopuces (*tilling arrays*), de séquençage de 1re génération et de QT-PCR qui ont été utilisées (Birney et al., 2007). Dans les faits, les techniques testés lors de la phase pilote du projet ENCODE n'ont généralement pas été utilisées en production puisque des technologies plus efficaces (utilisant le séquençage de nouvelle génération) ont vu le jour vers la fin du pilote.

ENCODE 2-3

La deuxième phase du projet est souvent nommée ENCODE 2 était plus ambitieuse et avait pour but de produire plus de 1600 jeux de donnés dans 147 types cellulaires (Consortium et al., 2012). Les buts principaux de cette phase du projet ENCODE étaient de déterminer les sites de liaison des facteurs de transcription et des histones, d'analyser l'accessibilité de la chromatine et d'identifier/quantifier les ARN dans les cellules et les compartiments cellulaires (Consortium et al., 2011). Les types cellulaires ont été séparés en trois niveaux de priorité, selon la quantité de données jeux de données produits. Le premier niveau est celui pour lequel le plus grand nombre de données a été produit et comprend les lignées K562 (érythroleucémie), GM12878 (lymphoblastoïde) et H1-hESC (cellule souche embryonnaire). Les lignées HeLa-S3 (cellules de carcinome cervical), HepG2 (hépatoblastome) et HUVEC (cellules endothéliales isolées à partir de la veine de cordon ombilical) sont dans le deuxième niveau. Tous les autres types cellulaires font donc partie du troisième niveau.

La troisième phase d'ENCODE, aussi connu sous le nom d'ENCODE 3, est en cours. Cette phase qui devrait durer 4 ans et se veut la continuité de la phase 2 (`https://www.encodeproject.org/2014-07-14-encode-third-phase/`).

Un portail web est disponible pour faciliter la recherche et le téléchargement d'expériences qui correspondent à certains critères. Il est par exemple possible de sélectionner uniquement les fichiers en format brut (p.e. : fastq) dans un type cellulaire donné (p.e. : GM12878) pour une expérience spécifique (p.e. : ChIP-Seq) tel que représenté dans la

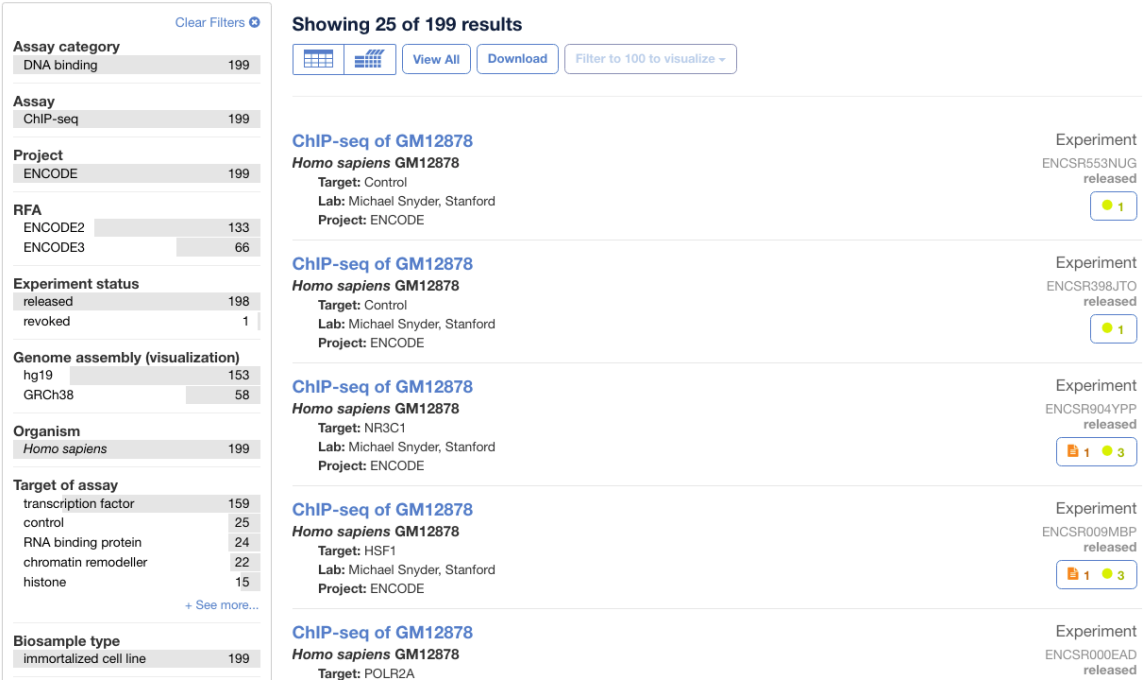


FIGURE 4.1 – Portail web du projet ENCODE. Capture d'écran du portail web d'ENCODE représentant le résultat d'une recherche pour les expériences de ChIP-Seq dans les GM12878 pour lesquelles on retrouve au moins un fichier fastq. Réalisée le 2016-12-09 à l'adresse suivante : https ://www.encodeproject.org/search/ ?searchTerm=fastq+gm12878+chip-seq&type=Experiment

figure 4.1. Jusqu'à tout récemment, il n'était pas possible de télécharger en groupe (*batch*) de nombreuses expériences, il fallait donc les télécharger manuellement une à une. Depuis le mois de mars 2015, il est possible de générer un script permettant de télécharger l'ensemble des fichiers associés à une requête. Le script de téléchargement donne également un lien permettant d'obtenir les métadonnées des fichiers associés à la requête. Par contre, ce document est encore incomplet. Par exemple, il n'est pas possible de déterminer quel est l'échantillon contrôle recommandé pour une expérience donnée. Cette information est souvent essentielle lorsqu'on souhaite reproduire les résultats obtenus ou bien produire des analyses complémentaires.

Finalement, un API REST est disponible pour permettre aux programmeurs de questionner directement la base de données d'ENCODE en utilisant des programmes ou des scripts. Cette interface programmatique entre leur base de données et un langage de programmation est très utile pour automatiser le téléchargement des données et des métadonnées. Il est donc possible d'améliorer la reproductibilité d'une analyse en réduisant au maximum les étapes manuelles qui sont difficiles à documenter et difficiles

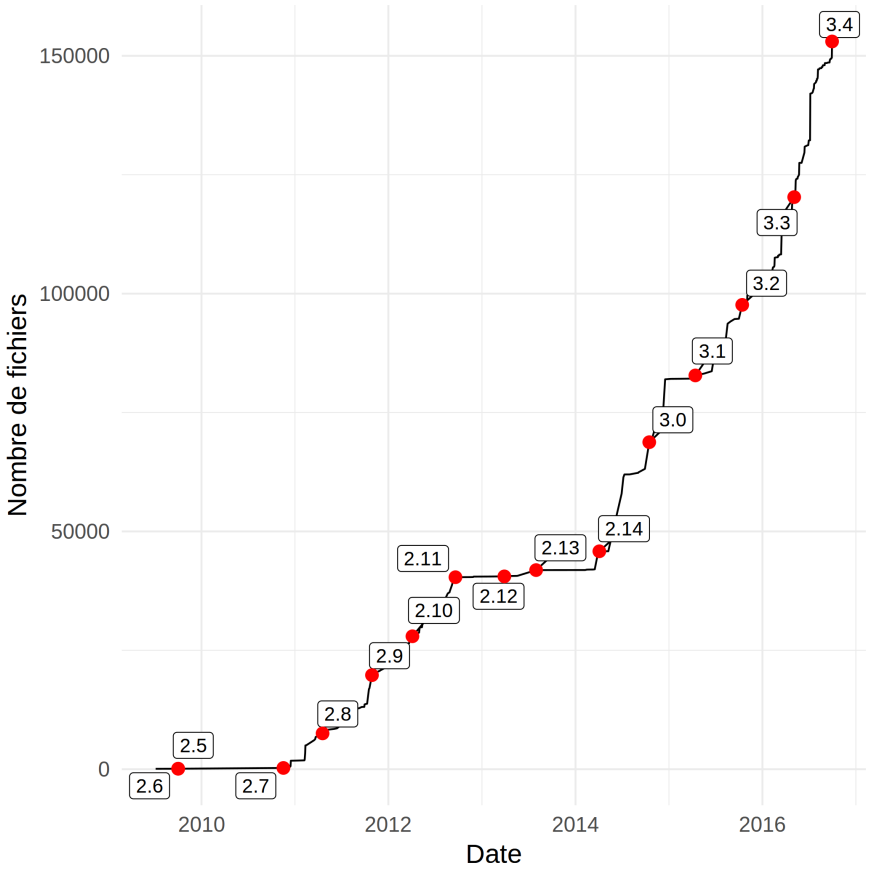


FIGURE 4.2 – Progression des données du projet ENCODE. Cette figure a été produite en utilisant une base de données fournie par l'outil ENCODExplorer v2.0.2. Les chiffres dans les rectangles représentent les versions de Bioconductor, la base de données d'ENCODE n'étant pas versionnée. Le rectangle en rouge correspond à la plage de temps actuellement couvert par l'outil encodexplorer. À noter que 7984 fichiers n'ont pas de date d'ajout et ne sont donc pas inclus dans cette figure.

à reproduire.

4.1.2 ENCODE en chiffre

La base de données d'ENCODE continue de croître à un rythme très rapide. En effet, les données de la troisième phase du projet sont en cours de production. Comme on peut le voir dans la figure 4.2, il y avait 161 004 fichiers disponibles sur la base de données d'ENCODE au moment où la dernière version de Bioconductor (version 3.4) est sortie. Cette figure permet également de constater que la vitesse à laquelle les données sont ajoutées ne semble pas être en voie de ralentir. L'espace disque nécessaire pour stocker l'ensemble des fichiers associés à cette version est d'environ 218 To.

TABLEAU 4.1 – Description des expériences principales produites par le consortium ENCODE.

Type	Compte	Pourcentage
ChIP-seq	5502	55.75
DNase-seq	539	5.46
shRNA RNA-seq	509	5.16
polyA mRNA RNA-seq	467	4.73
RNA-seq	461	4.67
eCLIP	310	3.14
small RNA-seq	211	2.14
RNA microarray	180	1.82
RNA Bind-n-Seq	145	1.47
RAMPAGE	139	1.41
genotyping array	123	1.25
CAGE	121	1.23
WGBS	113	1.14
single cell RNA-seq	112	1.13
RRBS	103	1.04
Autre	834	8.45
Total	9869	100.00

Les données produites par ENCODE sont majoritairement des données de ChIP-Seq (~55%), mais on compte 41 types d'expériences. Les expériences ayant le plus grand nombre de jeux de données sont présentées dans le tableau 4.1.

Le consortium ENCODE a produit des données de ChIP-Seq ciblant 754 facteurs différents dont les principaux sont présentés dans le tableau 4.2.

`ENCODExplorer`

L'outil `ENCODExplorer` a été développé pour compenser l'absence d'outils permettant d'automatiser l'extraction de métadonnées et le téléchargement des séquences. Le but principal était donc de développer un logiciel qui permet d'extraire toutes les informations de la base de données ENCODE en utilisant son API REST et qui offre l'option de les filtrer pour cibler les jeux de données qui correspondent à des critères de recherche spécifiques. C'est dans cette optique que le paquet R `ENCODExplorer` a été mis au point. Un autre aspect limitant de la base de données d'ENCODE est qu'elle n'est pas versionnée, ce qui devient problématique dans le contexte où la reproductibilité des analyses est importante. Il est nécessaire d'avoir accès à une version spécifique de la base de données pour pouvoir relancer reproduire une analyse en utilisant exactement

TABLEAU 4.2 – Description des cibles principales pour le ChIP-Seq produites par le consortium ENCODE.

Cible	Compte	Pourcentage
H3K4me3	374	8.95
H3K4me1	305	7.30
H3K36me3	303	7.25
H3K27me3	302	7.23
H3K9me3	277	6.63
H3K27ac	239	5.72
CTCF	192	4.60
H3K9ac	154	3.69
H3K4me2	112	2.68
POLR2A	90	2.15
EP300	32	0.77
H3K79me2	31	0.74
H2AFZ	25	0.60
MYC	25	0.60
REST	24	0.57
Autre	1693	40.52
Total	4178	100.00

la même version des fichiers que lors de l'analyse originale. Le paquet `ENCODExplorer` est fidèle à la philosophie de Bioconductor à cet égard, car il offre des versions de la base de données qui sont stables pendant chaque cycle de Bioconductor (environ 6 mois). De plus, à chaque modification de l'outil, la version de ce dernier est incrémentée, offrant ainsi la possibilité de télécharger une version précise au besoin.

`ENCODExplorer` est codé en R, car il existe un grand nombre d'outils de pointe spécialisés en génomique dans ce langage de programmation. En effet, grâce à Bioconductor (Gentleman et al., 2004), on retrouve de nombreux paquets de haut niveau implémentés et entretenus par des professionnels tels que `GenomicRanges`, `GenomicAlignments`, et bien d'autres (Lawrence et al., 2013). Ce répertoire de paquets R assure une qualité minimale en s'assurant que les paquets soient révisés avant d'être ajoutés et qu'il respecte plusieurs contraintes tout au long de sa vie : le paquet doit compiler en tout temps, tous les tests unitaires doivent être complétés avec succès et une documentation complète doit être présente. Grâce à ces nombreux critères, Bioconductor a une très grande notoriété dans le monde de la recherche et est un des 2 seuls dépôts de logiciels officiellement endossés par les éditeurs de la revue *Nature Genetics* (Anonymous, 2014). À noter que le second dépôt endossé par les éditeurs de cette revue est CRAN (*Comprehensive R*

Archive Network), le pendant de Biocondutor en R qui contient principalement des paquets statistiques. `ENCODExplorer` est disponible sur Bioconductor depuis avril 2015 et fait partie du top 20% des paquets en terme de téléchargement.

Le paquet `ENCODExplorer` offre essentiellement 3 fonctions : *queryEncode* pour faire des recherches dans la base de données téléchargée, *searchEncode* pour faire une recherche sur le portail web d'ENCODE et *downloadEncode* qui utilise le résultat des 2 autres fonctions pour télécharger les fichiers et valider qu'ils n'ont pas été corrompus lors du téléchargement. Ce paquet cherche donc à offrir une interface très simple entre le langage de programmation R et la base de données d'ENCODE. De plus, les métadonnées de Roadmap Epigenomic sont disponibles sur le portail d'ENCODE depuis octobre 2015 ce qui fait que l'outil `ENCODExplorer` a pu facilement adapté de manière à pour obtenir facilement toutes ces métadonnées.

L'outil `ENCODExplorer` est mature et est disponible sur le site de Bioconductor depuis plus d'un an (`http://bioconductor.org/packages/release/bioc/html/ENCODExplorer.html`). Il a été développé avant que le groupe d'ENCODE mettre en place le *batch download* et la possibilité d'obtenir des métadonnées. L'outil demeure tout de même très pertinent, car à ce jour, il n'est toujours pas possible d'obtenir le design complet d'une expérience d'ENCODE à partir des métadonnées qu'ils fournissent. En effet, l'information sur les contrôles à utiliser n'est toujours pas disponible alors que c'est une information souvent critique pour faire des analyses complémentaires. De plus, `ENCODExplorer` vérifie automatiquement l'intégrité des données téléchargées grâce au md5sum. `ENCODExplorer` permet d'obtenir des données dans un format qui peut être facilement utilisé avec d'autres paquets R spécialisés dans le domaine de la génomique, notamment avec l'outil `metagene` qui sera présenté plus loin dans ce document. Contrairement à la base de données d'ENCODE, `ENCODExplorer`offre un suivi des versions et une interface entièrement en ligne de commande pour faciliter la reproductibilité (une interface graphique est également disponible pour faciliter la tâche des chercheurs qui ne sont pas accoutumés avec le langage R).

4.2 Fantom

Le consortium Fantom est financé par le Riken et a vu le jour en 2000. son but est de déterminer la fonction du transcriptome humain et murin. Le consortium a complété 5 phases de son projet et devrait entamer la 6e phase sous peu. Dans le cadre de cette thèse, ce sont les données de la phase 5 (Fantom5) qui ont été utilisées pour la

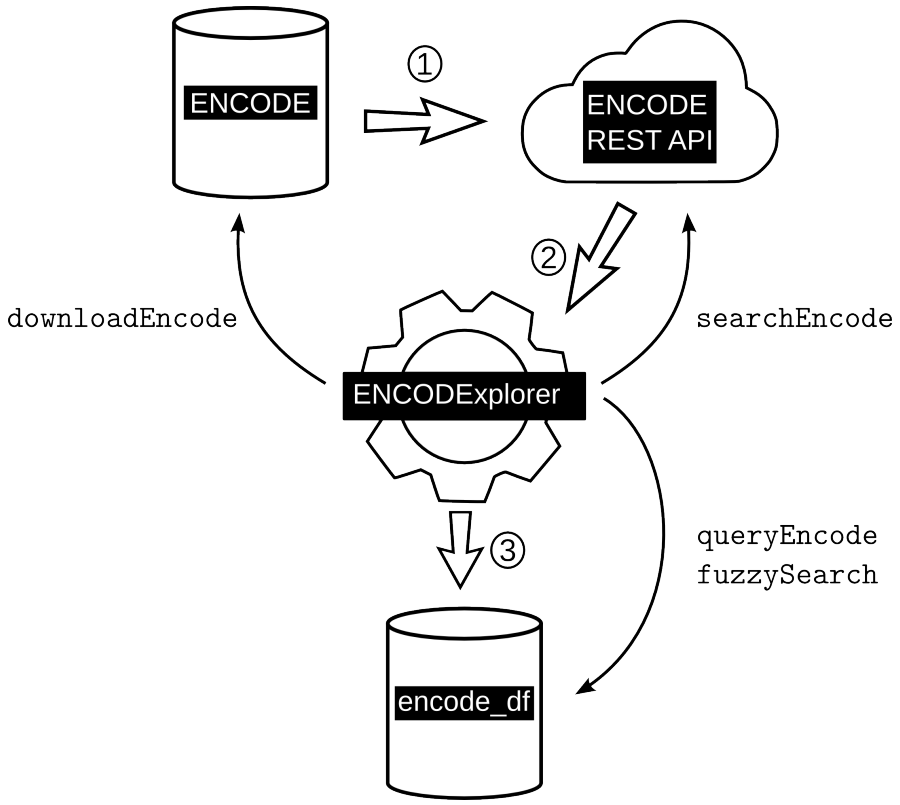


FIGURE 4.3 – Description du paquet R ENCODExplorer. Le consortium ENCODE conserve ces données dans une base de données représentée par un cylindre en haut à gauche. Pour faciliter l'accès aux données, il offre un API REST (représenté par un nuage) qui peut être utilisé pour effectuer des requêtes dans leur base de données. Le logiciel ENCODExplorer va chercher la totalité des tables disponibles dans la base de données d'ENCODE et les sauvegarder sous la forme d'un tableau R (data.frame) représenté par un cylindre au bas de la figure. La fonction searchEncode permet d'interroger directement l'API REST et retourne les mêmes résultats qu'une requête sur le site www.encodeproject.org. Les fonctions queryEncode et fuzzySearch permettent de rechercher de l'information sur la version de la base de données sauvegardée localement pour éviter de surcharger requêtes l'API REST. La fonction downloadEncode permet de télécharger les jeux de données obtenus suite aux requêtes précédentes.

démonstration de la pertinence des outils développés (section 7.8).

4.2.1 phase 1-4

Pour les deux premières phases, le projet Fantom regroupait une centaine de chercheurs (Hayashizaki, 2003) étaient intéressés principalement à l'annotation systématique des ARN complémentaires chez la souris (Kawai et al., 2001). Le projet Fantom1, qui a commencé avant la fin du séquençage du génome humain, avait pour but de séquencer plus de 20 000 clones d'ADNc pour ensuite les annoter par analyse d'ontologie et d'alignement de séquence (Kawai et al., 2001). Une analyse similaire sur plus de 60 000 clones a ensuite été réalisée lors de la deuxième phase de Fantom (Fantom2) (Okazaki et al., 2002).

À partir de la phase 3, le consortium a commencé à utiliser la technique du CAGE (Carninci et al., 2006). Une des conclusions principales de cette phase fut de confirmer la présence d'un grand nombre de transcrits non polyadénylé qui est présent sur la majorité du génome (Carninci et al., 2005). Ils ont aussi démontré que la transcription antisens des ARN codants est un phénomène fréquent chez les mammifères et serait un mécanisme important pour la régulation de la transcription (Katayama et al., 2005). Finalement, ils ont utilisé les résultats pour étudier la structure des régions promotrices à grande échelle pour mieux comprendre les différentes classes de promoteurs (Carninci et al., 2006). Il en ressort qu'une proportion importante des promoteurs n'ont pas un site d'initiation de la transcription précis, mais peuvent plutôt initier la transcription à partir de différents endroits dans une fenêtre d'environ 100 nucléotides.

La phase 4 du projet Fantom a également utilisé la technique de CAGE, mais en utilisant le séquençage de nouvelle génération pour obtenir une meilleure couverture des régions (*deepCAGE*) (Faulkner et al., 2009). Le projet principal était l'étude de la dynamique d'utilisation des sites d'initiation de la transcription dans le contexte de la différenciation cellulaire (Suzuki et al., 2009). Pour cette phase, le consortium a aussi étudié l'expression des rétrotranposons et de l'impact fonctionnel de la transcription de ces éléments. Ils ont démontré l'expression des rétrotransposons a probablement été sous-estimée, car les ARN ainsi produits sont principalement nucléaires (Faulkner et al., 2009). De plus la présence de rétrotransposons activement transcrits à proximité des gènes aurait une influence sur leurs niveaux d'expression.

4.2.2 phase 5

La 5e phase du projet Fantom a été séparée en deux parties. La première partie est la suite logique des deux étapes précédentes où le consortium a utilisé les données de *deepCAGE* pour étudier les sites d'initiation de la transcription à grande échelle. Des données ont donc été générées sur des cellules provenant de 152 tissus humains, 573 cellules primaires et plus de 200 lignées cellulaires (Consortium et al., 2014). La deuxième partie de la phase 5 était l'analyse de jeux de données de série temporelle (Arner et al., 2015).

Pour la première partie, les résultats présentés par le consortium concernaient l'analyse des sites d'initiation de la transcription retrouvés dans les régions promotrices des gènes (Consortium et al., 2014). Dans cette étude, le consortium a démontré que peu de gènes (6-14%) sont véritablement des gènes de maintenance (*housekeeping*), c'est-à-dire qu'ils sont détectés dans au moins 50% des échantillons. Les sites d'initiation de la transcription sont significativement enrichis dans les régions conservées, même si un pourcentage important (43%) des sites retrouvés chez l'humain ne le sont pas chez la souris. Les sites d'initiation de la transcription les plus conservés sont ceux associés aux gènes de maintenance, alors que les gènes spécifiques à un ou quelques types cellulaires sont remplacés plus fréquemment lors de l'évolution. L'autre aspect important de cette partie du projet était l'étude des amplificateurs (Andersson et al., 2014). Grâce à la technique de *deepCAGE*, le consortium a pu analyser le niveau d'expression des amplificateurs à travers tous les échantillons de manière à obtenir une liste de 65 423 éléments. Les caractéristiques de ces éléments ont été décrites plus en détail dans la section traitant de la régulation de la transcription.

Pour la deuxième partie du projet, le consortium s'est concentré sur des analyses de série temporelle pour mesurer et corréler les changements d'expression des amplificateurs et des promoteurs (Arner et al., 2015). Cette étude a permis d'étudier la dynamique de l'activation de la transcription des amplificateurs pour démontrer qu'elle a lieu majoritairement avant l'activation des promoteurs. De plus, ils ont démontré que l'activation d'amplificateurs était un événement ponctuel et que l'expression de ces éléments retourne rapidement à un niveau basal une fois l'expression des gènes cibles activée.

4.2.3 Fantom5 en chiffre

La liste des amplificateurs rendue publique par le consortium Fantom une fois les 2 étapes de la phase 5 complétées contient 65 423 amplificateurs dits permissifs (Arner

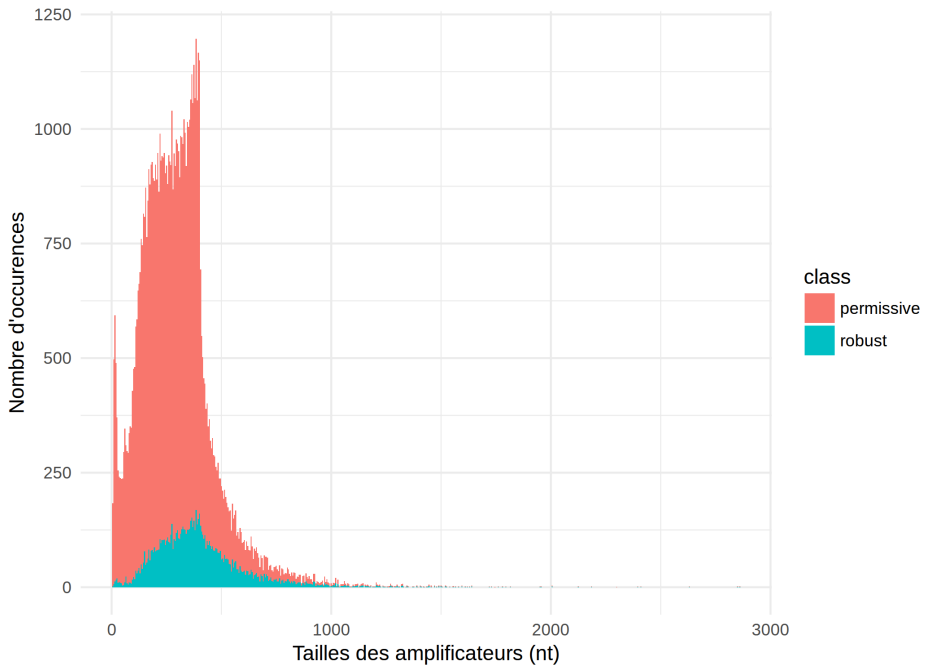


FIGURE 4.4 – Distribution de la taille des amplificateurs. Les informations sur la taille des amplificateurs définis par le consortium Fantom ont été calculées avec R en utilisant le paquet FantomEnhancers.hg19 v0.2.0 (https ://doi.org/10.5281/zenodo.31499).

et al., 2015). Les amplificateurs permissifs correspondent aux amplificateurs qu'il a été possible de mettre en évidence en utilisant les sites d'initiations de la transcription permissifs, c'est-à-dire les sites où une couverture d'au moins 2 séquences a été retrouvée pour au moins 1 nucléotide dans au moins une expérience (Consortium et al., 2014). Il existe également un sous-ensemble de sites d'initiation de la transcription considérés comme étant robustes qui correspond à des sites où une couverture d'au moins 10 séquences a été retrouvée pour au moins 1 nucléotide dans une expérience et dont l'expression normalisée est plus grande que 1 TPM (transcrits par million) (Consortium et al., 2014). 15.6% des amplificateurs de Fantom peuvent être considérés comme robustes (contiennent au moins un site d'initiation de la transcription robuste).

La taille des amplificateurs définis par Fantom varie de 3 nucléotides à 2861 nucléotides, avec une taille moyenne de 282.1 nucléotides. La distribution des tailles d'amplificateur est représentée dans la figure 4.4. On note une différence importante dans la distribution des tailles d'amplificateurs robustes comparativement aux amplificateurs permissifs, principalement au niveau des éléments de petite taille qui sont proportionnellement beaucoup moins fréquents.

Les amplificateurs sont généralement situés à proximité d'un site d'initiation de la

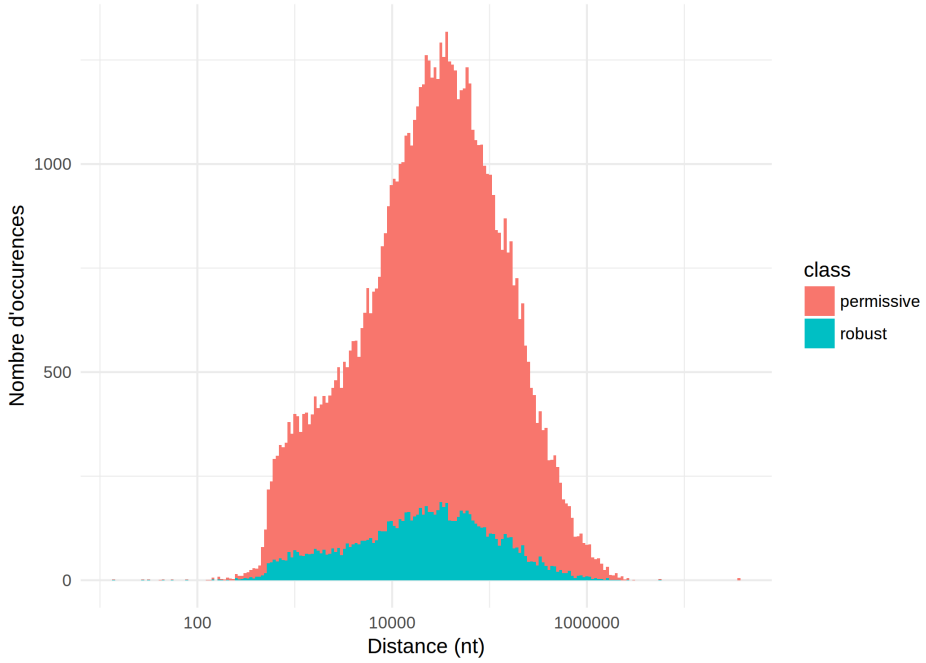


FIGURE 4.5 – Distance entre les amplificateurs et le site d'initiation de la transcription le plus près. Les informations sur la distance des amplificateurs par rapport au gène le plus proche ont été calculé avec R en utilisant le paquet FantomEnhancers.hg19 v0.2.0 (https ://doi.org/10.5281/zenodo.31499) et FantomTSS.hg19 v0.2.0 (https ://doi.org/10.5281/zenodo.31507).

transcription (voir figure 4.5). La distance médiane est de 28 201 nucléotides et la distance moyenne de 85 963. Il avait d'ailleurs été démontré dans les recherches du consortium Fantom que ce sont les promoteurs situés à proximité des amplificateurs actifs qui sont activés le plus rapidement (Arner et al., 2015).

4.2.4 Phase 6

La 6e phase du projet Fantom est en cours (`http://Fantom.gsc.riken.jp/6/`). Il y a peu de détails disponibles sur cette phase du projet. Il a été annoncé que le but principal sera d'étudier les longs ARN non codants (*lncRNA*). Il s'agit de déterminer leurs profils d'expression à travers plusieurs types cellulaires et de déterminer leurs rôles en perturbant leur expression dans différents types cellulaires.

Chapitre 5

Le ChIP-Seq

Pour mieux comprendre les mécanismes qui gouvernent la transcription des gènes, il faut bien comprendre le rôle des protéines qui en sont les effecteurs. Il est donc primordial de cataloguer les sites d'interactions entre les protéines et l'ADN dans le plus de types cellulaires possibles et dans différentes conditions pour être en mesure de mieux comprendre leur fonctionnement.

La méthode de l'immunoprécipitation de la chromatine couplée au séquençage de nouvelle génération (ChIP-Seq) a été publiée en 2007 (Johnson et al., 2007, Barski et al. (2007), Robertson et al. (2007), Mikkelsen et al. (2007)) et permet d'étudier les interactions entre une protéine d'intérêt et l'ADN. Avec l'avènement de cette technique, il est maintenant possible de cataloguer la majorité des sites d'interaction d'une protéine. Par contre, comme il sera décrit plus en détail dans les prochaines sections, cette approche a tout de même des limitations tant au niveau expérimental qu'au niveau des algorithmes disponibles pour effectuer l'analyse des données produites.

Dans cette section, je vais donc présenter la technique d'immunoprécipitation de la chromatine sans laquelle le ChIP-Seq ne serait pas possible. Je vais aussi décrire l'approche par biopuce qui était auparavant utilisée pour l'analyse de l'ADN obtenu par immunoprécipitation. Ensuite, je vais expliquer les différentes étapes expérimentales du ChIP-Seq. Finalement, je vais discuter des étapes d'une analyse bioinformatique typique de ChIP-Seq.

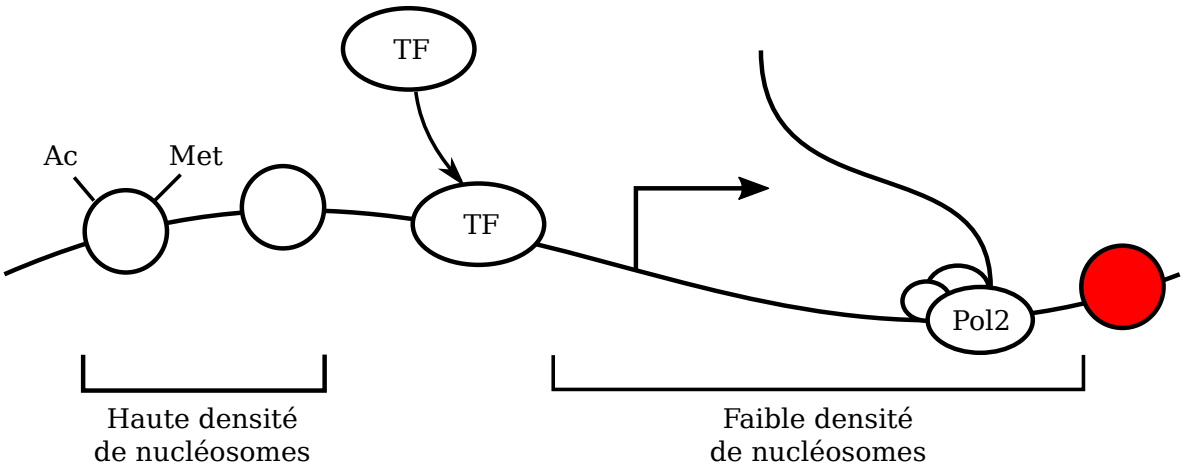


FIGURE 5.1 – Buts du ChIP-Seq. La technologie du ChIP-Seq permet non seulement d'étudier les facteurs de transcriptions (TF), mais également les nucléosomes (cercles), les marques d'histones (Ac et Met), les différentes sous-unités d'histone (cercle rouge et les facteurs impliqués dans la transcription (complexe Pol2).

5.1 Buts

Les analyses de ChIP-Seq ont été développées pour faciliter l'étude des interactions entre les protéines et l'ADN à l'échelle génomique. Il est donc possible de définir la majorité des sites de liaison d'une protéine pour un type cellulaire donné et dans une condition donnée. En utilisant cette technique, il est possible d'étudier non seulement les facteurs de transcription, mais également les histones, les modifications des histones, les nucléosomes ou bien les protéines impliquées dans la transcription des gènes.

5.2 Protocole expérimental

Il est important de bien comprendre les étapes expérimentales qui permettent de produire des données de ChIP-Seq, car plusieurs des limitations qui seront présentées plus loin en découlent directement.

5.2.1 Préparation des cellules

Pour une expérience de ChIP-Seq typique, il faut environ $5x10^7$ cellules (Schmidt et al., 2009, Furey (2012)), ce qui correspond à environ 10-100ng d'ADN et variera selon le nombre de sites de liaison attendu (Park, 2009). La quantité d'ADN nécessaire peut donc être un facteur limitant pour certains types d'expériences (Furey, 2012). Des variations du protocole de ChIP-Seq ont donc été développées pour pallier cette pro-

blématique, notamment le Nano-ChIP-Seq (Adli and Bernstein, 2011). Il est également recommandé de procéder à une étape de lyse cellulaire pour retirer les protéines du cytosol et ainsi améliorer la qualité des résultats (Schmidt et al., 2009).

5.2.2 Pontage covalent

Les cellules sont traitées au formaldéhyde pour créer des pontages covalents réversibles entre les protéines et l'ADN (Schmidt et al., 2009). Le but de cette étape est de fixer solidement les protéines et l'ADN. Cette étape permet donc de capturer des interactions plus faibles ou transitoires (Schmidt et al., 2009). Il est important d'éviter de traiter trop longuement les celles avec du formaldéhyde, car pourrait causer une augmentation importante du bruit de fond, notamment dans les échantillons contrôles (Marinov et al., 2014).

5.2.3 Fragmentation de l'ADN

L'approche la plus fréquemment utilisée pour fragmenter l'ADN est la sonication (Schmidt et al., 2009). Idéalement, la majorité des fragments obtenus devraient avoir une taille entre 200-600 nucléotides (Schmidt et al., 2009, Furey (2012),Park (2009)). Un désavantage de cette approche est que les fragments obtenus sont beaucoup plus longs qu'un site de liaison typique (200 nucléotides versus 6-20 nucléotides) et les librairies obtenues sont souvent contaminées par des fragments qui n'étaient pas liés par la protéine d'intérêt (Furey, 2012). Il est également possible de produire des fragments par digestion enzymatique, soit en utilisant une exonucléase (Rhee and Pugh, 2012) ou une MNase, surtout adaptée pour les analyses des nucléosomes ou des marques d'histones (Schones et al., 2008).

5.2.4 L'immunoprécipitation de la chromatine

Les fragments d'ADN doivent ensuite être enrichis pour conserver uniquement les fragments associés avec la protéine d'intérêt par une étape d'immunuprécipitation de la chromatine. C'est d'ailleurs cette étape essentielle qui donne le nom au ChIP-Seq (ChIP : *Chromatin ImmunoPrecipitation*).

Il est crucial d'utiliser un anticorps de bonne qualité, c'est-à-dire sensible et spécifique, pour minimiser la quantité de bruit de fond (Landt et al., 2012). Il est donc fortement recommandé de valider la qualité des anticorps, même lorsqu'ils sont commerciaux. En effet, le consortium ENCODE a évalué la qualité des anticorps commerciaux et

ils ont trouvé qu'environ 80% des anticorps ciblant des facteurs de transcription ne respectaient pas leurs critères de qualité (Landt et al., 2012).

Outre l'utilisation d'anticorps de mauvaise qualité, l'étape d'immunoprécipitation peut être influencée par les protocoles utilisés pour les étapes en amont (Lefrançois et al., 2010). Par exemple, un traitement au formaldéhyde trop long lors du pontage covalent peut masquer les épitopes reconnus par l'anticorps et une sonication trop longue peut générer des fragments trop courts.

Une alternative à l'immunoprécipitation est l'utilisation d'épitopes tels que HA, Myc, Flag ou V5 (Kidder et al., 2011). Cette approche consiste à produire une protéine de fusion qui contient une de ces étiquettes puis de réaliser l'étape d'immunoprécipitation en utilisant un anticorps la ciblant. Cette approche plus complexe permet généralement de diminuer la quantité de bruit de fond (Lefrançois et al., 2010).

Une fois l'immunoprécipitation de la chromatine complétée, on devrait obtenir des fragments d'ADN fortement enrichis pour la protéine étudiée. Il ne reste ensuite qu'à retirer les protéines avant de passer à l'étape d'analyse principale par biopuces (ChIP-chip) ou par séquençage (ChIP-Seq).

5.2.5 Les biopuces

Les premières analyses à l'échelle génomique des séquences obtenues suite à une immunoprécipitation de la chromatine ont été publiées au début des années 2000 grâce à la technologie des biopuces (Ren et al., 2000). Cette technique consiste à utiliser des biopuces qui contiennent des séquences d'ADN complémentaire aux régions promotrices des gènes connus (Weinmann and Farnham, 2002). L'ADN obtenu suite à l'immunuprécipitation de la chromatine est ensuite couplé à une molécule fluorescente, puis hybridée avec la biopuce (Schena et al., 1995). Puisque les séquences identiques sont regroupées sur la biopuce, on peut déterminer de manière quantitative quelles régions promotrices ont été enrichies en mesurant leur fluorescence.

5.2.6 Le séquençage

À partir des années 2007 (Johnson et al., 2007, Barski et al. (2007),Robertson et al. (2007),Mikkelsen et al. (2007)), il a été possible d'utiliser les nouvelles technologies de séquençages pour analyser les séquences enrichies par immunoprécipitation. La diminution progressive des coûts de séquençage a fait que c'est maintenant l'approche

dominante dans les publications scientifiques. Il y a environ 60% de concordance entre les régions enrichies détectées par l'approche des biopuces comparativement à l'approche par séquençage de nouvelle génération (Ghosh and Qin, 2010). Il est par contre difficile d'analyser conjointement les résultats produits par ces deux approches à cause de l'hétérogénéité des données (Ghosh and Qin, 2010).

Cette approche offre de nombreux avantages comparativement aux analyses par biopuces. L'avantage le plus évident est de ne pas demander une connaissance préalable des sites de liaison potentiels pour le facteur étudié et donc d'offrir une meilleure couverture du génome (Chen et al., 2012). L'approche d'analyse conventionnelle des données de ChIP-Seq requiert tout de même l'utilisation d'un génome de référence, bien qu'une approche d'analyse *de novo* a été récemment proposée (He et al., 2015). Les autres avantages incluent une meilleure plage dynamique (c'est-à-dire une meilleure précision lorsque le signal est très faible ou très fort (Schmidt et al., 2009, Park (2009))), une meilleure résolution (Lefrançois et al., 2010) et un meilleur ratio signal-bruit (Marinov et al., 2014).

L'approche par séquençage amène aussi son lot de problématiques. Par exemple, des biais expérimentaux peuvent être introduits par la méthode utilisée pour la fragmentation de l'ADN, l'amplification PCR, des erreurs au niveau du séquençage, l'effet de lot (*batch,*) la technologie de séquençage choisie ou la qualité des séquences qui n'est pas uniforme (par exemple, les séquences d'Illumina contiennent plus d'erreurs vers la fin) (Lefrançois et al., 2010, Furey (2012),Park (2009)). De nombreux biais peuvent aussi provenir de la structure de la chromatine et du génome étudié tels que la présence de nombreuses régions répétées qui diminuent la mappabilité ou la structure de la chromatine qui rend certaines régions plus difficiles à séquencer (Lefrançois et al., 2010). Le biais le plus important est au niveau de la compaction de la chromatine : les régions où la chromatine est ouverte sont généralement surreprésentées (Chen et al., 2012, Marinov et al. (2014),Auerbach et al. (2009)). Le contenu en GC va aussi influencer le nombre de séquences obtenues : il y a une corrélation entre le pourcentage de GC et le nombre de séquences lorsque le pourcentage de GC se situe entre 20 et 56% (Kuan et al., 2011a). En plus des biais expérimentaux et des biais en lien avec la structure du génome, il existe aussi des biais statistiques : l'enrichissement mesuré dans deux régions situées à proximité n'est pas nécessairement indépendant (Lefrançois et al., 2010). L'utilisation d'un contrôle permettrait de retirer au moins une partie de ces biais (Furey, 2012).

Pour obtenir des résultats de bonne qualité, il est primordial de produire un nombre

TABLEAU 5.1 – Nombre de séquences utilisables recommandées par ENCODE selon le type de région.

Type de profil	ENCODE2	ENCODE3
Ponctuel	10 millions	10 millions
Large spectre	20 millions	20 millions

de séquences suffisant pour le facteur étudié. Un sous-séquençage mènera à un grand nombre de faux négatifs (Marinov et al., 2014). Cette quantité de séquence va dépendre du nombre de régions d'interaction attendu pour chaque protéine et du profil d'enrichissement attendu. Généralement, les facteurs seront considérés comme faisant partie d'une ou des deux catégories suivantes : ponctuel (*narrow*) ou à large spectre (*broad*). Les protéines ayant un patron d'enrichissement de type ponctuel sont généralement des facteurs de transcription et ce type d'enrichissement reflète le fait que ces protéines se lient à l'ADN en reconnaissant un motif spécifique. À l'inverse, les protéines ayant un profil à large spectre sont des protéines qui vont occuper une plus grande région et qui se lient à intervalle régulier. On parle donc généralement des histones et des protéines associées aux histones. Certains types de facteurs auront des profils hybrides. Par exemple, la polymérase peut être retrouvée en pause (profil ponctuel) ou bien en élongation (profil à large spectre) (Kim et al., 2011).

Étant donné que les protéines qui produisent des profils à large spectre vont généralement couvrir une plus grande proportion du génome, il faut produire un plus grand nombre de séquences que pour les protéines à profil ponctuel pour obtenir une couverture similaire à chaque site de liaison. Le consortium ENCODE recommandait donc de produire plus de 10 millions de séquences utilisables (alignées à un seul endroit dans le génome) lorsqu'on travaille avec des protéines ayant un profil ponctuel et plus de 20 millions un profil à large spectre pour un organisme eucaryote pour la phase 2 du projet (Landt et al., 2012). Ces valeurs ont été augmentées à 20 millions de séquences utilisables pour les facteurs ayant un profil ponctuel et 45 millions pour celles ayant un profil à large spectre pour la phase 3 du projet. Ces recommandations n'ont pas fait l'objet d'une publication, mais sont disponibles sur la page du projet ENCODE (`https://www.encodeproject.org/data-standards/chip-seq/`). Les recommandations sont résumées à la table 5.1.

Pour s'assurer d'avoir produit un nombre suffisant de séquence, on peut faire des analyses de saturation. Il s'agit de faire des analyses en série du même échantillon en utilisant des sous-ensembles de différentes tailles provenant des données originales. En

théorie, si le nombre de séquences produites était suffisant, on devrait voir un plateau lorsqu'on produit un graphique représentant le nombre de régions enrichies retrouvées par la taille de l'échantillon utilisée. Expérimentalement, peu de données vont atteindre un plateau (Kharchenko et al., 2008, Rozowsky et al. (2009)). La raison principale est que plus le nombre de séquences est élevé, plus on aura tendance à détecter les régions qui ont un faible ratio d'enrichissement par rapport au contrôle. C'est pour cette raison qu'on peut utiliser une approche qui consiste à déterminer un seuil pour le ratio d'enrichissement pour les analyses de saturation (MSER : *Minimum Saturation Enrichment Ratio*) (Park, 2009). L'utilisation d'un plus grand nombre de séquence permettra donc de retrouver les régions plus faiblement enrichies, mais n'aura pas un impact significatif sur la précision de l'estimation de la position des événements de liaison (Kharchenko et al., 2008).

La taille des séquences est un autre élément important auquel il faut se questionner lors de la production de données de ChIP-Seq. En effet, on veut généralement trouver l'équilibre adéquat entre le nombre de séquences produites, leur taille et le coût du séquençage. Il faut donc éviter de produire trop de séquences et des séquences trop longues si on souhaite produire des designs expérimentaux plus complexes ayant de multiples conditions à comparer. À cet effet, on considère qu'au-delà de 100 nucléotides les gains au niveau du résultat de l'étape d'alignement sont négligeables comparativement au coût associé à leur production (Whiteford et al., 2005). Au niveau du compromis coût/bénéfice, il est mieux de produire un plus grand nombre de séquences ou bien de produire des réplicats biologiques. Pour les mêmes raisons, il est rare de voir des analyses de ChIP-Seq qui utilisent des séquences pairées. Le gain au niveau de l'alignement est marginal, excepté dans les régions répétées où on peut aligner environ 15% de séquences supplémentaires en utilisant des séquences pairées pour une expérience typique de ChIP-Seq (Chen et al., 2012).

La préparation des librairies est l'étape la plus critique du protocole de séquençage. Elle va dépendre d'un grand nombre de critères tels que la quantité de matériel de départ, la durée du traitement au formaldéhyde et de la sonication, la qualité de l'anticorps utilisé et une suramplification PCR (Chen et al., 2012, Bailey et al. (2013)). Une librairie de qualité est généralement complexe, c'est-à-dire qu'elle contient peu de séquences redondantes (Landt et al., 2012). En théorie, si on pouvait augmenter significativement la qualité de l'immunoprécipitation, des expériences réussies pour des facteurs de transcription ayant peu de sites d'interaction avec l'ADN auraient une faible complexité. Mais puisque dans les faits ce sont seulement une minorité des séquences

qui proviennent de fragments associés à la protéine précipitée, une faible complexité sera habituellement associée à une expérience échouée (où l'immunoprécipitation aura produit peu de matériel) (Marinov et al., 2014, Landt et al. (2012)). Il existe certains outils qui permettent de prédire la complexité tels que le paquet `preseq` et l'outil `PCR bottleneck coefficient` d'ENCODE (Bailey et al., 2013).

5.2.7 Design expérimental

Un aspect qu'il ne faut pas négliger lors de la production de données de ChIP-Seq est de s'assurer d'avoir le bon design expérimental pour répondre à l'hypothèse de départ. Un bon design facilitera grandement l'analyse des données de ChIP-Seq. Pour déterminer si une région enrichie détectée dans un échantillon est le fruit du hasard ou si c'est véritablement une région qui peut être systématiquement détectée d'une expérience à l'autre, il est recommandé de produire des réplicats biologiques et des contrôles (Park, 2009, Landt et al. (2012)).

Les réplicats ont pour but de s'assurer que les résultats obtenus sont reproductibles et permettent de quantifier la variation biologique entre les échantillons d'une expérience pour faciliter les analyses d'enrichissement différentiel (Rozowsky et al., 2009). Pour que le réplicat soit valable, il est important que ce soit un réplicat biologique et non pas un réplicat technique. Pour produire ce type de réplicat, il faut que les cellules initialement utilisées proviennent de deux sources indépendantes (cultures cellulaires, regroupement de cellules d'embryon ou échantillons tissulaires) (Landt et al., 2012). L'étude la plus citée pour déterminer le nombre de réplicats biologiques optimal est arrivée à la conclusion qu'il ne semble pas y avoir d'intérêt à produire plus de deux réplicats biologiques (Rozowsky et al., 2009). Par contre, certaines études plus récentes ont proposé qu'il serait nécessaire de produire trois réplicats biologiques ou plus (Lefrançois et al., 2010, Yang et al. (2014)).

En plus des réplicats biologiques, il est généralement recommandé de produire des contrôles pour permettre une meilleure évaluation du bruit de fond et faciliter la détection des régions enrichies. Le bruit de fond est difficile à évaluer, car il n'est pas aléatoire et peut même mener à des profils d'enrichissement similaires à ce qui est obtenu après l'immunoprécipitation de la chromatine (Marinov et al., 2014). Le bruit de fond n'est pas aléatoire puisqu'il est le reflet des biais intrinsèques à la structure de la chromatine et à la méthodologie expérimentale (voir section précédente). Le bruit de fond est en fait le mélange de trois types de signaux : 1) un signal qui correspond à du bruit de

fond typique, sans structure apparente, 2) un fort enrichissement de séquence dans une région spécifique et 3) un enrichissement similaire à un site de liaison légitime.

Trois types de contrôles sont utilisés dans les expériences de ChIP-Seq : l'ADN de départ (*input DNA*), l'IP bidon (*Mock IP*) et l'IP non spécifique (*non-specific IP*) (Landt et al., 2012). Un contrôle de type ADN de départ consiste à utiliser l'ADN obtenu avant l'étape d'immunoprécipitation comme contrôle. Le contrôle aura donc été traité au formaldéhyde et aura été fragmenté par sonication (Lefrançois et al., 2010). Ce type de contrôle se veut donc représentatif de l'état d'ouverture de la chromatine et de son accessibilité pour le séquençage. C'est essentiellement la même procédure expérimentale qu'une expérience de Sono-seq (Auerbach et al., 2009), qui a été élaborée pour détecter les régions où la chromatine est accessible. Les contrôles IP bidon et IP non spécifiques sont réalisés en effectuant l'immunoprécipitation soit sans anticorps ou bien avec un anticorps reconnaissant une protéine ne liant pas l'ADN (habituellement les IgG). Le but de ce type de contrôle est de cibler les biais d'enrichissement qui sont dus à l'étape d'immunoprécipitation. La faiblesse de cette approche est qu'elle produit parfois peu d'ADN, ce qui mène à une librairie de complexité plus faible que les échantillons immunoprécipités et des contrôles utilisant l'ADN de départ (Marinov et al., 2014).

Il y a peu d'études qui ont comparé formellement les différents types de contrôles à large échelle pour déterminer lequel devrait être favorisé. On sait par contre qu'il y a une forte corrélation entre la position des séquences des contrôles avec les séquences obtenues suite à l'immunoprécipitation, entre autres au niveau des sites d'initiation de la transcription actifs (Rozowsky et al., 2009). Les contrôles utilisant l'ADN de départ et l'IP non spécifique sont les plus fréquemment rencontrés (Marinov et al., 2014). En théorie, les contrôles IP non spécifiques seraient les plus appropriés, car ce sont ceux qui suivent de plus près la procédure expérimentale du ChIP-Seq et devraient donc mesurer plus fidèlement les biais spécifiques à ce type d'expérience. Dans les faits, les contrôles qui utilisent un anticorps anti-IgG ont une valeur de complexité plus faible que l'immunoprécipitation (Marinov et al., 2014) alors qu'il est recommandé d'utiliser des contrôles qui ont un nombre de séquences et une complexité similaire à l'échantillon immunoprécipité (Landt et al., 2012). De plus, les contrôles IgG contiennent plus souvent des profils d'enrichissement typiques d'un événement de liaison légitime (Marinov et al., 2014), ce qui suggère qu'ils sont plus susceptibles aux erreurs expérimentales.

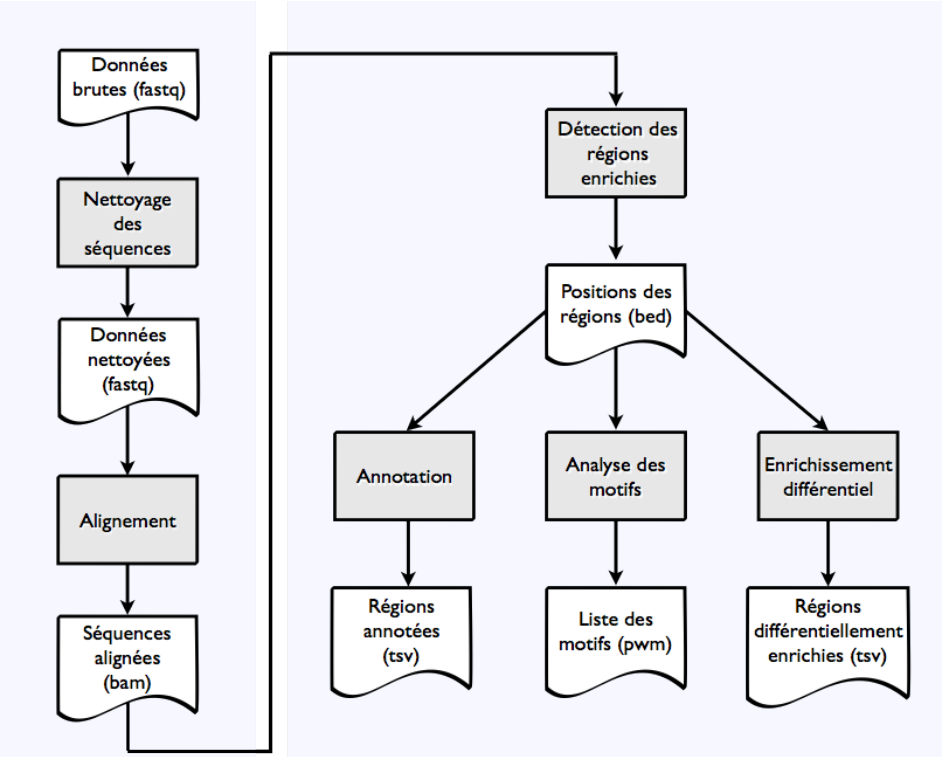


FIGURE 5.2 – Survol du protocole d'analyse standard des expériences de ChIP-Seq. Une analyse typique de ChIP-Seq débutera par le nettoyage des séquences et par leur alignement sur un génome de référence (partie gauche de la figure. L'étape suivante est la détection des régions enrichies où les sites de liaisons légitimes seront décrits. Ces régions seront ensuite annotées pour cibler les gènes à proximité. Une analyse de motifs peut servir de contrôle de qualité ou bien à découvrir de nouveaux motifs de liaison pour la protéine étudiée. Finalement, si le design expérimental le permet, il sera possible de comparer plusieurs expériences pour cibler les régions où la protéine d'intérêt est différentiellement enrichie.

5.3 Les analyses de données de ChIP-Seq

5.3.1 Alignement

Comme son nom l'indique, le but de l'étape d'alignement est de déterminer l'emplacement des séquences générées à l'étape de séquençage sur le génome étudié. Compte tenu du nombre de séquences, il est primordial d'utiliser des algorithmes d'alignement ultrarapides. L'alignement ultrarapide de courte séquence est une méthode heuristique et ne garantit donc pas de trouver le meilleur alignement. Les aligneurs ultrarapides les plus utilisés sont présentement BWA (Li and Durbin, 2009, Li (2013)), Bowtie/Bowtie2 (Langmead et al., 2009, Langmead and Salzberg (2012)), Novoalign (www.novocraft.com). D'une manière similaire aux méthodes par recherche de mots

popularisées par l'outil Blast (Altschul et al., 1997), les méthodes d'alignement ultrarapides vont scinder chaque séquence en sous-séquences et rechercher des matchs parfaits entre ces dernières et le génome de référence. Les régions ainsi ciblées seront par la suite étudiées plus en détail pour confirmer l'alignement et en calculer le score (Ghosh and Qin, 2010).

La performance de l'étape d'alignement des séquences produites lors d'une expérience de ChIP-Seq va dépendre d'une part de la qualité des séquences produites, mais aussi de la qualité du génome de référence. La structure du génome va également avoir une influence importante sur le nombre de séquences alignées. Il est impossible d'aligner à un seul endroit certaines séquences lorsqu'elles se retrouvent dans des régions répétées (qui représentent environ 52% du génome humain (Park, 2009). On parle alors de mappabilité (*mappability*) du génome, c'est-à-dire la fraction du génome qu'il est possible d'analyser avec des séquences d'une taille donnée. Cette valeur va varier d'une espèce à l'autre et de la taille des séquences utilisées.

Il est donc possible d'améliorer la mappabilité d'un génome en utilisant des séquences plus longues et/ou des séquences pairées. Il faut tenir compte coût associé à une augmentation de la taille des séquences et des séquences pairées comparativement au bénéfice obtenu en termes de pourcentage de séquences alignées. Il a été démontré que le gain est minimal pour les analyses de ChIP-Seq au-delà de 100 nucléotides et qu'il n'est pas nécessaire de produire des séquences pairées (Park, 2009). Par exemple, il est possible d'aligner à un seul endroit 79.6% des séquences de 30 nucléotides, le pourcentage passe à 89.3% pour des séquences de 70 nucléotides (Park, 2009, Rozowsky et al. (2009)). De plus, la distribution de la mappabilité n'est pas aléatoire. Par exemple, les régions à proximité des sites d'initiation de la transcription ont une meilleure mappabilité en moyenne comparativement au reste du génome (Rozowsky et al., 2009).

Le pourcentage de séquence alignée peut donner une bonne idée sur la qualité de l'expérience. Lorsque moins de 50% des séquences sont alignées à un seul endroit sur le génome, c'est généralement parce que l'expérience n'a pas fonctionné correctement (Bailey et al., 2013).

5.3.2 Détection de régions enrichies

Un fois les séquences correctement alignées, il faut passer à l'étape de la détection des régions enrichies (*peak calling*). C'est au cours de cette étape qu'on tente de séparer les sites de liaisons légitimes du bruit de fond. Le but est d'une part de définir la liste des

régions génomiques pour lesquelles un véritable événement de liaison entre la protéine d'intérêt et l'ADN et d'autres parts de classer ces régions selon la probabilité que le site soit légitime.

Pour différencier un véritable site de liaison du bruit de fond, les algorithmes vont généralement rechercher un profil d'enrichissement spécifique. Étant donné que les séquences de ChIP-Seq sont rarement pairées, toutes les séquences produites auront la même orientation (5'-3'). Cette propriété fera que la majorité des séquences retrouvées en amont proviendront du même brin d'ADN et que les séquences en aval proviendront principalement de l'autre brin (Kharchenko et al., 2008). En intégrant cette analyse de la forme du profil et de la symétrie des brins, on arrive à mieux discerner les véritables régions enrichies des artéfacts qu'en utilisant uniquement le niveau d'enrichissement par rapport au contrôle (Leleu and Lefebvre, 2010). Il est d'autant plus important de tenir compte du profil de l'enrichissement, car les régions qui contiennent le plus grand nombre de séquences sont souvent des faux positifs (Leleu and Lefebvre, 2010). Les outils utilisés pour la détection des régions enrichies vont donc rechercher des régions pour lesquelles on observe ces deux pics situés à proximité l'un de l'autre et donc l'orientation suit le profil attendu.

La distance entre les deux pics correspond à la taille des fragments sélectionnée avant le séquençage (Park, 2009). Beaucoup d'outils de détection des régions enrichies sont capables d'évaluer la taille des fragments en sélectionnant un sous-ensemble de région (Zhang et al., 2008). Le principe est que chaque région du sous-ensemble de régions sélectionnées doit avoir un nombre de séquence similaire en aval et en amont du site de liaison potentiel. Les algorithmes chercheront donc à définir quelle est la distance qui permet d'expliquer le plus de paires de séquences par une approche de corrélation croisée (*cross-correlation*) (Kharchenko et al., 2008).

Les outils de détection des régions enrichies peuvent habituellement être classées en trois catégories selon l'algorithme utilisé pour la détection des régions : approche non paramétrique, paramétrique ou algorithmes complexes tels que les modèles de Markov cachés (Ghosh and Qin, 2010). Ces approches se distinguent principalement au niveau des techniques statistiques utilisées pour le calcul des scores statistiques tels que le FDR et la valeur p.

C'est en utilisant la comparaison avec le bruit de fond que les algorithmes seront en mesure d'attribuer un score à chaque région et de calculer une valeur p ou q pour chaque région.

Les algorithmes de détection des régions enrichies vont généralement calculer un ou plusieurs scores pour permettre le classement des régions et représenter la probabilité que la région soit valide. En plus du niveau d'enrichissement de l'immunoprécipitation par rapport au contrôle, on retrouve régulièrement un taux de fausse découverte (FDR : *False Discovery Rate*) (Kim et al., 2011, Ghosh and Qin (2010)). Un très grand nombre d'outils sont disponibles pour effectuer cette étape de l'analyse des données de ChIP-Seq (voir la table tab :A.1).

La détection des régions enrichies à large spectre peut soit utiliser un algorithme similaire à ce qui est utilisé pour régions ponctuelles avec quelques modifications mineures à l'algorithme. Par exemple, le logiciel Macs2 utilisera le même algorithme pour cibler les régions, mais effectuera un traitement à la fin de l'analyse pour combiner les régions situées en deçà d'une certaine distance (Zhang et al., 2008). De cette manière, on évite d'obtenir un trop grand nombre de régions qui correspondent finalement au même groupe d'histone. Il existe également d'autres outils qui ont été développés spécifiquement pour la détection de ce type de facteurs tels que PING (Zhang et al., 2012) et RJMCMC (Samb et al., 2015).

La taille des fragments aura également un impact sur la précision des régions retrouvées. Plus les fragments de la librairie sont gros, plus la taille des régions enrichies sera grande (Chen et al., 2012). Il est malgré tout généralement possible de bien positionner les sommets dans chaque région (Chen et al., 2012). Il faut également déterminer si les séquences redondantes doivent être utilisées. Dans la plupart des régions, on retrouve environ 20 à 40% de séquences redondantes (Chen et al., 2012). Ces séquences dupliquées peuvent soit provenir de plusieurs fragments indépendants ou d'une amplification d'un même fragment (Bailey et al., 2013). Dans le premier cas, il serait souhaitable de les conserver alors que dans la deuxième situation, il serait mieux de ne pas les intégrer dans l'analyse. Étant donné qu'il est difficile de distinguer entre les deux sources, il est recommandé de ne pas les utiliser pour la détection des régions enrichies, quitte à les intégrer plus loin lors des analyses quantitatives (Bailey et al., 2013). Lorsque les séquences dupliquées sont retirées, la sensibilité est généralement améliorée (Chen et al., 2012), probablement car les séquences dupliquées légitimes se retrouvent dans des régions fortement enrichies où le signal est déjà très fort.

Au-delà du défi technique de cibler les véritables sites de liaisons reproductibles d'une expérience de ChIP-Seq, séparer les sites de liaison biologiquement pertinents du bruit de fond biologique qui est un problème encore plus difficile. En effet, certains sites de

liaisons considérés comme étant faibles ont été démontrés comme étant biologiquement pertinents (Bailey et al., 2013). Le problème est d'autant plus complexe que les véritables sites de liaisons biologiquement importantes ne sont généralement pas connus. Des études ont démontré que seul un petit pourcentage des sites de liaisons d'une protéine ont un impact significatif sur l'expression des gènes (Farnham, 2009). Par exemple, lorsque l'expression des facteurs de transcription est réprimée, on trouve que l'expression d'environ 1-10% des gènes cibles potentiels est affectée (Farnham, 2009). De plus, il n'y a pas nécessairement un lien direct entre la force du signal et son importance biologique (Bailey et al., 2013, Landt et al. (2012)). Il faut donc utiliser des mesures indirectes pour valider l'importance biologique d'un site telle que la présence d'un site de liaison connu pour la protéine étudiée ou pour ses partenaires d'interaction (Landt et al., 2012).

5.3.3 Analyse des motifs

Une fois les régions enrichies détectées, on s'intéresse souvent à déterminer quels sont les motifs présents dans nos séquences. Un motif est une séquence reconnue par une protéine liant l'ADN et qui permet à cette dernière d'interagir avec l'ADN à des endroits spécifiques. Contrairement aux sites de restriction qui sont des séquences d'ADN très spécifique et où les changements d'un seul nucléotide peuvent avoir un impact majeur sur sa reconnaissance par son enzyme, les motifs de liaison des facteurs de transcription sont beaucoup plus souples (Stormo, 2000).

Comme le montre la figure 5.3, les motifs de liaison ne sont généralement pas des séquences précises, mais plutôt une probabilité de retrouver un nucléotide donné. Une protéine sera donc généralement capable de reconnaître plusieurs séquences similaires sur l'ADN. Cette représentation correspond à ce qui est appelé une matrice de poids position *Position Weight Matrix* (PWM) en anglais.

Il y a trois raisons principales pour lesquelles on pourrait souhaiter faire une analyse des motifs. 1) À titre de contrôle de qualité pour valider que les séquences attendues sont présentent dans les régions enrichies. 2) Pour la recherche de partenaires d'interaction de la protéine étudiée. La présence d'un motif de liaison d'une autre protéine peut potentiellement représenter une liaison indirecte entre la protéine étudiée et l'ADN. 3) Pour caractériser les motifs reconnus par notre protéine. Il y aura donc deux approches principales pour l'analyse des motifs : la recherche de motifs connus et la recherche de motifs *de novo*.

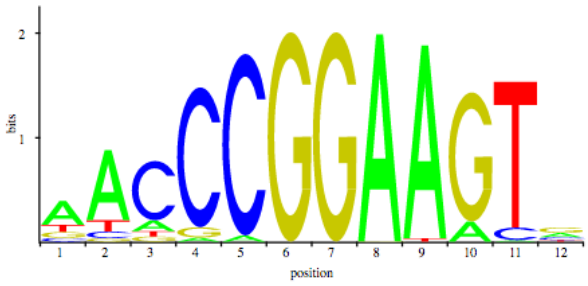


FIGURE 5.3 – Représentation graphique d'un motif de liaison. Logo de séquence représentant visuellement les probabilités de retrouver un nucléotide à une position données dans le motif de liaison de la protéine ELF1. Téléchargé de la base de données JASPAR [@sandelin2004jaspar,http ://jaspar.genereg.net/], le 2 novembre 2016.

L'analyse par recherche de motifs connus dans le contexte du ChIP-Seq se nomme analyse d'enrichissement de motif. Le but est donc de déterminer si des motifs connus sont présents en quantité significativement différente (plus grande ou plus petite) à ce qui serait attendu par hasard (Frith et al., 2004). De cette manière, il est possible de valider la présence d'un motif de liaison attendu.

L'autre approche consiste à faire la recherche de motifs *de novo*. C'est un problème extrêmement difficile qui ne peut être résout de manière exacte dans un délai raisonnable (Lihu and Holban, 2015). Des approches comme MEME (Bailey et al., 2009) et rGADEM (Mercier et al., 2011) vont utiliser des approches itératives pour limiter la recherche à un sous-ensemble pertinent de motifs.

5.3.4 Enrichissement différentiel

Grâce à l'augmentation importante des données produite par le séquençage de nouvelle génération, il est maintenant possible de générer des expériences de plus en plus complexes. C'est pour cette raison que de nombreux outils on fait leur apparition ces dernières années pour faciliter la comparaison de deux ou plusieurs échantillons. Les analyses d'enrichissement différentiel pour les expériences de ChIP-Seq font face à de nombreux défis techniques. Les sites de liaison peuvent se retrouver partout à travers le génome, la force du signal est théoriquement illimitée, la quantité de bruit de fond est très importante et la taille des régions est variable (Steinhauser et al., 2016). De plus, certains outils permettant de détecter l'enrichissement différentiel vont dépendre de l'étape de détection des régions enrichies et vont tenir pour acquis que les régions utilisées en entrée sont significatives et reproductibles (Bailey et al., 2013).

Approches qualitatives

La méthode la plus simple qui permet de comparer des échantillons de ChIP-Seq entre eux est l'approche qualitative. Dans son format le plus simple, il s'agit de représenter sous forme de diagramme de Venn le nombre de sites de liaison partagés. Puisqu'un même site de liaison peut avoir une taille variable d'un échantillon à l'autre selon la distribution des séquences, il est nécessaire de procéder à une étape de normalisation pour ajuster les tailles des régions de manière à ce qu'elles soient comparables dans tous les échantillons. Des outils comme MULTOVL (Aszódi, 2012) permettent de comparer qualitativement des échantillons et offrent des tests statistiques pour comparer les différences d'enrichissement globales. Pour des analyses plus poussées qui permettent de comparer une région à la fois, il faut se tourner vers les algorithmes quantitatifs.

Approches inspirées de la transcriptomique

L'approche principalement utilisée est l'analyse d'enrichissement différentiel, très fortement inspiré par les approches d'expression différentielle développées en transcriptomique tels que edgeR (Robinson et al., 2010) et DESeq (Anders and Huber, 2010). Ces approches appliquées au ChIP-Seq se veulent être une amélioration d'une comparaison naïve par diagramme de Venn où on regarde recherche uniquement les régions qui sont présentent dans une condition et absente dans l'autre. Le principe de ces techniques est de comparer les échantillons région par région pour définir lesquelles ont un enrichissement significativement différent. Ces régions représenteraient donc les endroits où l'intensité de l'interaction entre notre protéine d'intérêt et l'ADN a changé d'une condition à l'autre.

Autres approches

Comme que mentionné plus haut, beaucoup des algorithmes d'analyse d'enrichissement différentiel s'inspirent fortement des approches développées pour l'analyse d'expression différentielle en transcriptomique. L'exemple le plus probant est l'outil DiffBind (Stark and Brown, 2011) qui utilise directement deux des algorithmes les plus populaires pour les analyses de RNA-Seq : EdgeR et DESeq. DiffBind offre un ensemble de fonctions utilitaires qui permet d'adapter les résultats de la détection des régions enrichie dans le format nécessaire pour EdgeR et DESeq. Pour les analyses d'expression différentielle, il est absolument nécessaire d'avoir accès à des réplicats biologiques pour être en mesure de déterminer statistiquement si deux régions sont différentiellement exprimées. Par contre, pour le ChIP-Seq, il arrive encore fréquemment que des expériences soient

réalisées sans produire de réplicats biologiques, ce qui rend impossible l'utilisation des algorithmes basés sur les recherches en transcriptomique.

D'autres approches sont également offertes pour comparer les régions obtenues lors d'analyse de ChIP-Seq. La vaste majorité de ces algorithmes ont comme hypothèse de départ que les deux conditions comparées sont très similaires pour l'ensemble des régions et qu'elles diffèrent seulement pour quelques régions. Cette hypothèse permet de comparer des échantillons sans nécessairement avoir accès à des réplicats biologiques, car on ne recherche pas directement les régions ayant un enrichissement différentiel. On recherche plutôt les régions qui ont un patron d'enrichissement différent de la majorité des régions dans un seul des échantillons.

Les outils qui permettent de cibler les sites différentiellement enrichis donnent des résultats très différents lorsqu'on compare les résultats obtenus sur un même jeu de données (Steinhauser et al., 2016). Outre pour les analyses sur des jeux de données à large spectre, où certains outils performent bien sûr des données simulées, la majorité des outils ont tendance à avoir soit un nombre de faux négatifs très élevés (peu de régions retrouvées) ou soit un nombre de faux positifs très élevé.

5.3.5 Validation de la qualité des données

Il est essentiel de valider la qualité des données brutes et des résultats des étapes principales des analyses. Car si les données utilisées avec un programme sont de mauvaise qualité, les résultats le seront également.

Le consortium ENCODE recommande un certain nombre d'analyses qui permettent de valider la qualité des régions enrichies détectées lors d'une expérience : valider le nombre de régions enrichies retrouvées, calculer la proportion de séquences retrouvées au sein des régions enrichies, vérifier la distribution de la taille des fragments et estimer le coefficient de reproductibilité.

Avant la détection des régions enrichies

Le premier contrôle qualité doit être effectué au niveau des séquences brutes pour déterminer rapidement quelle proportion des séquences a la qualité nécessaire pour la poursuite des analyses. C'est également lors de cette étape qu'on pourra déterminer si on retrouve un nombre anormalement élevé de séquences d'adaptateurs. Grâce aux informations obtenues lors de cette étape, on pourra cibler quels paramètres seront les plus efficaces pour le nettoyage des séquences. Il est recommandé de revérifier la qualité

des séquences une fois le nettoyage terminé pour s'assurer que tous les problèmes ciblés au départ. À la fin de cette étape, on aura donc l'assurance de travailler uniquement avec des séquences qui ont une qualité raisonnable.

Un deuxième point critique dans le pipeline d'analyse est l'étape d'alignement. En effet, même si les séquences originales ont une bonne qualité, il est possible qu'un nombre important de séquence ne soient pas alignés sur notre génome de référence. On estime que plus de 70% de séquences correspondent à un bon alignement alors que moins de 50% de séquences alignées peuvent être problématiques (Bailey et al., 2013). Il est normal qu'une certaine quantité de séquences ne soient pas alignées correctement, même avec des données de très grande qualité, surtout si le génome étudié contient des régions répétées.

Au-delà de la qualité des séquences, on doit également s'assurer que le nombre de séquences uniques alignées sur le génome est suffisant pour le type de facteur étudié tel que présenté dans la section 4.1.3 Séquençage. C'est ainsi qu'on peut s'assurer que la librairie produite est assez complexe pour offrir une vue générale de la majorité des événements de liaison protéine-ADN.

Validation de la qualité d'un seul échantillon

Au-delà du nombre de séquences alignées, il est également important d'évaluer la complexité de l'alignement, c'est-à-dire le nombre de séquences uniques dans le jeu de données. Le consortium ENCODE a proposé d'utilisé la métrique de fraction non-redondante (NRF : *Non-Redudant Fraction*)(Landt et al., 2012), qui est similaire à la métrique de redondance (Heinz et al., 2010).

Pour estimer la proportion de bruit de fond, on peut utiliser la métrique de fréquence des séquences dans les régions enrichies (FRiP : *Frequency of Reads in Peaks*)(Landt et al., 2012). Cette métrique représente simplement le pourcentage des séquences alignées qui se retrouvent dans une région enrichie. Les expériences de ChIP-Seq ont un niveau très élevé de bruit de fond. On considère donc qu'un score FRiP plus grand que 1% (99% de bruit de fond) est normal pour une expérience typique dans laquelle on retrouve une dizaine de milliers de régions enrichies (Landt et al., 2012). Certaines expériences valides peuvent aussi avoir un FRiP inférieur à 1% (Frietze et al., 2010).

Puisque les séquences retrouvées dans un site de liaison légitime ont un profil d'enrichissement bimodal où les séquences sur le brin plus sont situées d'un côté du site et les séquences du brin moins de l'autre côté, on devrait pouvoir cibler les expériences

réussies en recherchant la présence de ce type d'enrichissement. Les analyses de corrélations croisées permettent de déterminer si une proportion importante des séquences ont ce profil d'enrichissement (Kharchenko et al., 2008). Le principe est de calculer la corrélation entre les séquences pour différents décalages. À chaque décalage, les séquences sur le brin plus, les positions de la séquence est incrémenté alors que pour les séquences sur le brin moins, la position est décrémentée. La corrélation maximale d'une bonne expérience devrait donc correspondre à la taille des fragments, où la majorité des séquences du brin plus chevaucherait parfaitement les séquences du brin moins. Une fois les valeurs de corrélations calculées, on peut les visualiser sous forme de graphique. On trouve généralement deux pics : un premier pic qui correspond à la corrélation au niveau des séquences et qui est généralement associé au bruit de fond et un second pic qui correspond au fragment. Idéalement, le deuxième pic devrait avoir une amplitude plus grande ou égale au premier pour une expérience de bonne qualité (Kharchenko et al., 2008, Landt et al. (2012)).

Une autre approche utilisée pour valider la qualité des régions enrichies retrouvées est de s'assurer que les motifs de liaison connus pour la protéine étudiée se retrouvent bel et bien dans les séquences des régions enrichies. Évidemment, cette étape est uniquement possible lorsque des motifs sont connus pour notre protéine d'intérêt. Lorsque c'est le cas, on peut effectuer des tests statistiques qui nous permettront de valider qu'il y a plus d'occurrences des motifs attendus dans nos séquences que ce qui serait attendu aléatoirement.

Validation de la qualité de réplicats biologiques

Lorsque des réplicats biologiques sont disponibles, il est important de valider qu'ils sont bien corrélés ensemble de manière à cibler rapidement si le ChIP-Seq est reproductible. Initialement, des études utilisaient les coefficients de Pearson (par exemple : Frank et al. (2015), Young et al. (2011) et Ho et al. (2011)) ou de Spearman par exemple : Kasowski et al. (2010) et Kim et al. (2010)). Le coefficient de Pearson est considéré comme étant plus sensible pour la comparaison des réplicats de ChIP-Seq (Ho et al., 2011).

Plus récemment, une métrique nommée taux de découverte irréductible (IDR : *Irreducible Discovery Rate*) a été proposée comme étant une amélioration aux mesures de corrélation traditionnelles (Li et al., 2011). Le principe de l'IDR est de classer les régions de chaque réplicat en débutant par les régions ayant le meilleur score. En comparant les rangs correspondant entre les réplicats, on va classifier chaque région comme étant reproductible ou non reproductible. Ensuite, on va choisir un seuil qui va correspondre

au nombre de régions non reproductibles qu'on est prêt à tolérer dans les résultats finaux. Le nombre de régions qui est retourné à un seuil d'IDR donné correspond donc aux nombres de régions classées pour lesquelles on s'attend à retrouver en déca de la valeur du seuil de régions non reproductibles. Le consortium ENCODE recommande d'utiliser des pseudo-réplicats (obtenus en permutant aléatoirement les séquences de chaque réplicat). Le nombre de régions qui sont reproductibles entre les 2 réplicats originaux devrait être au minimum 2 fois plus petit que le nombre retrouvé en utilisant les pseudo-réplicats (Landt et al., 2012).

Chapitre 6

Introduction au logiciel `metagene`

Le logiciel `metagene` offre une alternative aux approches d'enrichissement différentiel. Au lieu d'essayer de comparer une région à la fois, `metagene` cherche plutôt à analyser les patrons d'enrichissement de groupes de régions génomiques. Cet outil cherche donc à comparer des expériences complexes (plusieurs réplicats, contrôles et conditions) de manière à cibler des différences dans les profils d'enrichissement. Ce type d'approche se distingue donc des algorithmes qui cherchent à comparer des paires d'échantillons en comparant les régions une à une.

Pour arriver à analyser convenablement les profils d'enrichissement, il est primordial d'être capable de gérer rapidement un très grand volume de données. En effet, pour être en mesure de comparer quantitativement les profils d'enrichissement, il faut travailler directement avec les données d'alignement pour obtenir le plus de précision possible. Autrement, lorsqu'on travaille directement avec les régions enrichies, chaque région est résumée par une seule valeur qui est souvent soit le score calculé par l'outil de détection des régions enrichies ou simplement le nombre de séquences. L'efficacité des algorithmes utilisés peut donc faire la différence entre une analyse qui va nécessité plusieurs heures de calcul versus une analyse qui sera complétée en quelques minutes.

Dans cette optique, il importe d'utiliser des implémentations des algorithmes d'analyses qui sont optimisées. Pour les paquets R tels que `metagene`, cela se traduit souvent en l'utilisation de librairies codées en langage de programmation plus bas niveau tel que les langages compilés C et C++ (p.e. : que la librairie HTSlib à travers le paquet `Rsamtools`).

J'ai collaboré au développement de librairie `NGS++` codée en C++ et qui offre des fonctions de base très optimisées pour la lecture écriture. Cette librairie sera utilisée

pour améliorer la vitesse de certaines opérations critiques de `metagene` telles que le décompte du nombre de séquence par fichier et le calcul du niveau de couverture. Le chapitre 8 de cette thèse correspond à l'article publié dans Oxford Bioinformatics qui décrit le fonctionnement de cette librairie.

Outre la vitesse d'exécution, qui est un critère majeur pour qu'un outil soit utilisé à grande échelle, il est encore plus important d'avoir une interface adaptée aux besoins et aux attentes de chaque type d'utilisateur. Pour cette raison, les utilisateurs avancés de R ont accès à une interface en ligne de commande qui permet de contrôler avec grande précision chaque étape de l'analyse. Pour les utilisateurs qui ne sont pas familiers avec le langage R, nous offrons également une interface graphique nommée `Imetagene` qui permet de faciliter l'interaction avec le logiciel.

Finalement, le point le plus important est évidemment l'utilisation des algorithmes appropriés pour permettre d'obtenir les résultats les plus fiables possible. Dans le contexte de la comparaison des profils d'enrichissement, on doit tenir compte du bruit de fond qui est endémique dans les données de ChIP-Seq et il faut normaliser les échantillons, car les étapes de la production des données de ChIP-Seq peuvent avoir une grande influence sur les résultats obtenus. Par exemple, la spécificité des anticorps peut être très variable d'un anticorps à l'autre, entre deux lots d'un même anticorps ou même entre deux expériences distinctes réalisées avec le même anticorps.

L'outil `metagene` est donc un paquet R qui permet d'extraire efficacement l'information des données d'alignement, d'utiliser les échantillons contrôles pour évaluer et retirer le bruit de fond et de normaliser les échantillons pour permettre leur comparaison qui performe des tests statistiques sur les distributions des profiles et permet d'afficher les résultats sont forme de graphique de type metagene. Il est disponible sur Bioconductor depuis août 2014 et se retrouve dans le top 20% au niveau du nombre de téléchargements. De plus, des algorithmes permettent d'utiliser un autre paquet R développé dans l'équipe du Dr Arnaud Droit, `similaRpeak`, pour permettre des comparaisons statistiques des profils selon différents aspects de la courbe tels que l'intensité du signal brut ou normalisé et la position maximale de chaque courbe. L'outil `metagene`, son interface graphique `Imetagene` et l'outil `similaRpeak` sont présentés dans le chapitre 7 de cette thèse.

Chapitre 7

L'analyse de profils avec `metagene` révèle des patrons d'enrichissement spécifiques à certains facteurs impliqués dans la régulation de la transcription

7.1 Journal

Cet article a été publié dans la revue *PLOS Computational Biology* le 18 août 2016. Il est distribué sous la license *Creative Common* (CC BY 4.0).

7.2 Résumé

Cet article présente le logiciel `metagene` qui a été développé pour permettre de faciliter l'intégration des données de ChIP-Seq et fournir une représentation graphique favorisant la comparaison rapide d'un grand nombre de régions génomiques. Ce paquet a été élaboré de manière à permettre de comparer des jeux de données de diverses provenances en offrant des options qui permettent de normaliser les données et des fonctions statistiques adaptées à ce type de comparaison. Grâce au paquet `similaRpeak`, `metagene` peut également utiliser différentes métriques pour comparer les profils d'enrichissement en allant au-delà de l'amplitude du signal.

La première partie de l'article présente le logiciel. On y trouve les détails sur son design et son implémentation. On y présente également les différentes fonctionnalités offertes, telles que l'affichage graphique sous forme de metagene ou de *heatmaps* interactives.

Dans la seconde partie de l'article, on présente les analyses de données d'ENCODE utilisées pour démontrer la pertinence d'utiliser l'outil `metagene` dans un contexte biologique réel. Plus de 100 facteurs de transcriptions et histones provenant de ChIP-Seq de cellules GM12878 ont été analysées en utilisant `metagene` pour comparer les mécanismes de recrutement au niveau des amplificateurs et des promoteurs.

Les figures supplémentaires sont également présentées à l'annexe B.

7.3 Contributions

Je suis l'auteur de l'outil `metagene` ainsi que son mainteneur officiel. Pour plus de détails sur la contribution des auteurs, voir la section Avant-propos et la section 7.11.

7.4 Contenu

metagene Profiles Analyses Reveal Regulatory Element's Factor-Specific Recruitment Patterns

Charles Joly Beauparlant[1,2],&, Fabien C. Lamaze[1,3],&, Astrid Deschênes[1], Rawane Samb[1], Audrey Lemaçon[1], Pascal Belleau[1], Steve Bilodeau[1,3,4], Arnaud Droit[1,2],*

[1] Centre de Recherche du CHU de Québec - Université Laval, Québec, Québec, Canada, [2] Département de Médecine Moléculaire, Faculté de médecine, Québec, Canada, [3] Centre de Recherche sur le Cancer de l'Université Laval, Québec, Québec, Canada, [4] Département de Biologie Moléculaire, Biochimie Médicale et Pathologie, Faculté de médecine, Québec, Canada

* `Arnaud.Droit@crchudequebec.ulaval.ca`

& These authors contributed equally to this work.

7.5 Abstract

ChIP-Sequencing (ChIP-Seq) provides a vast amount of information regarding the localization of proteins across the genome. The aggregation of ChIP-Seq enrichment signal

in a metagene plot is an approach commonly used to summarize data complexity and to obtain a high level visual representation of the general occupancy pattern of a protein. Here we present the R package metagene, the graphical interface Imetagene and the companion package similaRpeak. Together, they provide a framework to integrate, summarize and compare the ChIP-Seq enrichment signal from complex experimental designs. Those packages identify and quantify similarities or dissimilarities in patterns between large numbers of ChIP-Seq profiles. We used metagene to investigate the differential occupancy of regulatory factors at noncoding regulatory regions (promoters and enhancers) in relation to transcriptional activity in GM12878 B-lymphocytes. The relationships between occupancy patterns and transcriptional activity suggest two different mechanisms of action for transcriptional control : i) a "gradient effect" where the regulatory factor occupancy levels follow transcription and ii) a "threshold effect" where the regulatory factor occupancy levels max out prior to reaching maximal transcription. metagene, Imetagene and similaRpeak are implemented in R under the Artistic license 2.0 and are available on Bioconductor.

7.6 Introduction

Understanding the global regulation of gene expression programs is an important goal of functional genomics studies. To this end, it is now standard procedure to survey the occupancy of regulatory proteins genome-wide using chromatin immunoprecipitation coupled with massively parallel sequencing (ChIP-Seq) (Johnson et al., 2007). Affordability and accessibility of the technique are now generating more complex experimental designs containing many samples, treatments, controls comparisons and technical replicates. Furthermore, the abundance of public datasets, such as those provided by the ENCODE (Consortium et al., 2012) and Roadmap Epigenomics (Bernstein et al., 2010) consortiums, provides a wealth of information. Unfortunately, the integration of large amounts of ChIP-Seq information remains challenging.

In a typical ChIP-Seq analysis, reads are first aligned using an aligner of choice and peaks are called using peak calling algorithms, such as MACS (Zhang et al., 2008) or PICS (Zhang et al., 2011), to obtain a list of occupied regions. Then, these regions are annotated to genes (Zhu et al., 2010) and/or used to search for DNA binding motifs (Bailey et al., 2006). In addition, tools were developed to quantitatively compare regions from ChIP-Seq experiments in order to define regions with differential binding between conditions (Wu et al., 2015). The algorithms and models used to manage background, to

normalize read counts and to estimate the reads distribution across the genome are the main differences between the different methods. While these tools allow the discovery of regions that are differentially occupied by a factor of interest, they are unable to evaluate differences in the general occupancy patterns of DNA-binding proteins. Furthermore, they rely on the peak calling step which varies greatly based on the algorithm or the parameters used (Szalkowski and Schmid, 2011).

Current approaches to compare and summarize enrichment signals for groups of regions rely on visual representations of the average enrichment at a specific position. These representations are known as metagene plots (also referred to as meta-gene (Shin et al., 2009) or aggregation plots (Kundaje et al., 2012)). To compare multiple samples, many tools implemented reads per million aligned (Kundaje et al., 2012, Shen et al. (2014)) or quantile (Stempor, 2014, Dharmalingam and Carroll (2015)) normalizations. The addition of confidence intervals (represented as ribbons) based on standard errors (of mean or of percentiles) in `ngs.plot` (Shen et al., 2014), on bootstrap approaches in `ChIPseeker` (Yu et al., 2015) or as standard error in `seqPlots` (Stempor, 2014) improved the prediction of the mean. However, while confidence intervals are effective tools to estimate the range within which the true mean is likely to lie, profile comparisons require statistical testing. In addition, valuable information embedded in the enrichment profiles such as the position of the binding event inside the region or the presence of a specific pattern notwithstanding its amplitude is currently ignored. Therefore, representation tools enabling a quantitative assessment and robust statistical comparisons of metagene profiles are needed.

We developed the `metagene` package to quantitatively compare enrichment profiles of group of regions. Specifically, this package is designed to 1) facilitate the integration of signal from many datasets linked by complex experimental designs, 2) statistically compare the enrichment profiles of groups of genomic regions and 3) provide visual representations of the data to facilitate interpretation. Here we used the `metagene` package to investigate how regulatory factors contribute to the transcriptional output of noncoding regulatory regions. Indeed, recruitment of regulatory factors to noncoding regulatory regions, including enhancer and promoter regions, modulates the transcriptional response of each gene. Using the `metagene` and `similaRpeak` package, we identified the similarities and dissimilarities in the recruitment patterns of these factors at enhancer and promoter regions. Our results demonstrate that there are two distinct mechanisms of action for transcriptional regulators. Indeed, we discovered that the level of the regulatory factors either correlates with the transcriptional activity or saturates

prior to maximal transcriptional activity of the regulatory region. We termed those patterns "gradient effect" and "threshold effect".

7.7 Design and Implementation

The `metagene` package builds upon Bioconductor scalable data structures for representing annotated ranges on the genome (Lawrence et al., 2013). Additionally, to efficiently import large datasets, `metagene` supports the most common genomic file formats such as bam, bed and narrowPeak/broadPeak. The number of files used in a single analysis is only limited by the computer memory available. To reduce memory usage, `metagene` produces coverages only for the genomic regions of interest and stores this information in Run-length encoding. It is possible to compare multiple region groups and multiple experiments in a single analysis. To increase the analytical power, `metagene` uses the controls to estimate the signal-to-noise ratio and remove background signal. The datasets are also normalized for an accurate comparison. Furthermore, the directionality of the genomic regions (i.e. the strand) is usable to highlight asymmetric enrichment patterns. In the final graphical output, the metagene plot, each curve summarizes the information of multiple genomic regions (termed region groups) from a single experiment. When used with the `similaRpeak` package, our approach allows the comparison of multiple samples and gives the possibility to statistically compare the results with metrics adapted to different profile features. The `Imetagene` package offers a simple graphical interface to manage complex experimental designs. A workflow of a typical `metagene` analysis is provided in 7.1.

In order to quantitatively compare different experiments, it is crucial to take into account the signal-to-noise ratio and to normalize samples. Indeed, the ChIP-Seq signal is a mixture of legitimate signal and noise. The experimental noise is influenced by biological factors such as the GC content and the chromatin structure (Kuan et al., 2011b) and by technical factors such as the antibody quality, the cell number, the DNA fragmentation and the library construction (Kidder et al., 2011). A common approach to separate true signal from noise is to use controls. Ideally, the controls should be normalized to fit only with the noise component of the chip signal since only this part of the signal will follow the same distribution (Angelini et al., 2015). In order to normalize the controls before subtracting the background, `metagene` uses the Normalization of ChIP-Seq (NCIS) approach (Liang and Keleş, 2012) to calculate the signal to noise ratio. This approach performs well on ChIP-Seq datasets (Angelini et al., 2015) and is

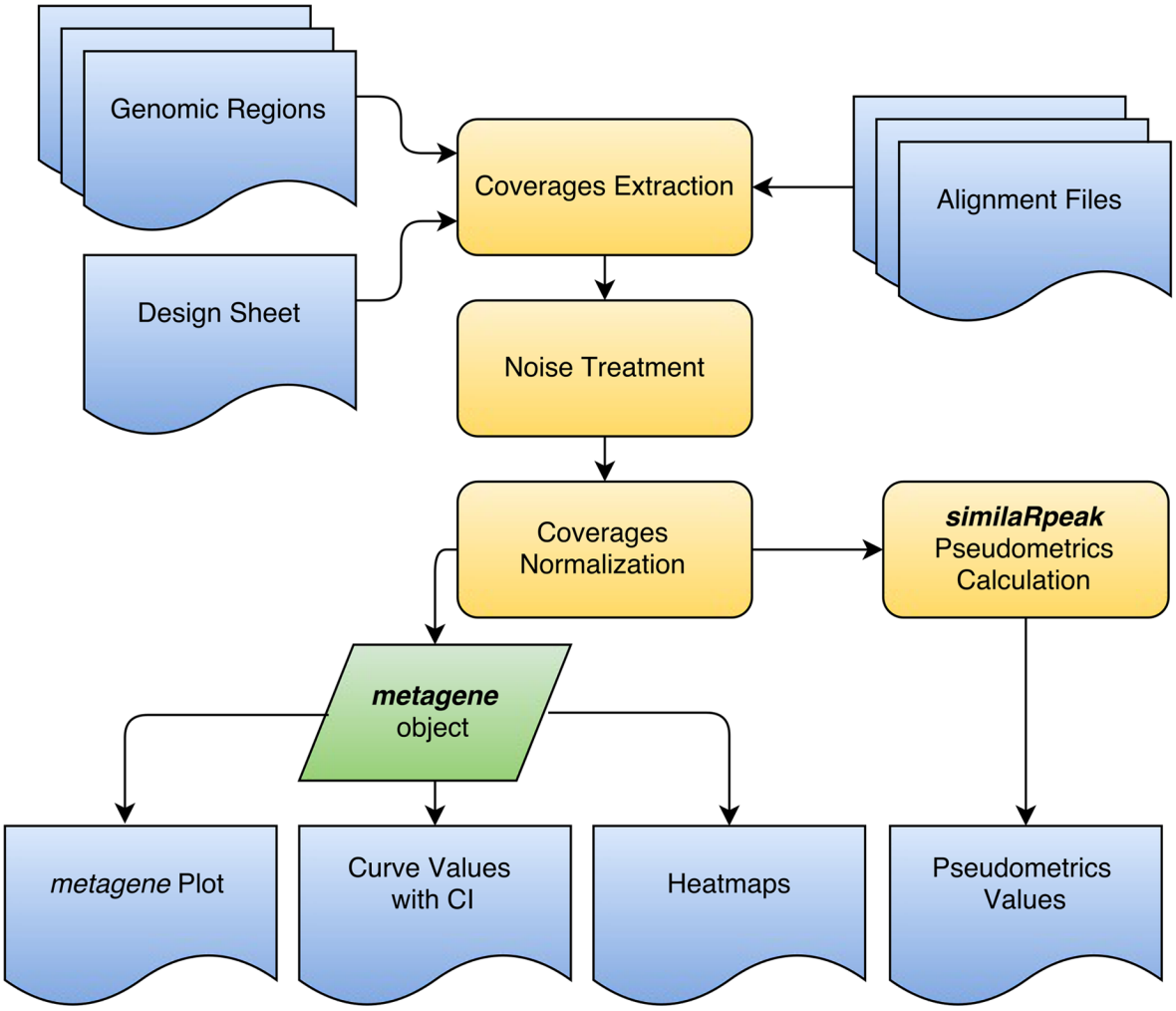


FIGURE 7.1 – metagene workflow. A metagene analysis requires 3 types of inputs : 1) a list of genomic regions (BED or GRanges formats), 2) alignment files (BAM format) and 3) a design sheet (data frame format) explaining the relations between samples. The alignment files are processed to extract the coverages of every genomic regions. Afterward, the background is removed from the coverages and the signal is normalized (reads per millions aligned or RPM) to allow comparison between samples. The main output is the metagene plot. The other outputs are the curve values and confidence intervals (CI) used to produce the plot and an interactive heatmap with Imetagene. The results are compatible with similaRpeak for profile characterization.

readily available in R (7.2A and 7.2B show the effect of noise reduction). If multiple samples are compared together, they should be normalized to take into account the difference in library sizes. This is performed in `metagene` by converting the raw coverage values in read per millions aligned. It is also possible to change the orientation of each genomic region on the negative strand to represent every region in the 5'->3' orientation. The profile of each group defined in the design is calculated using either an average or median profile, as specified by the user. A confidence intervals of the estimators (mean or median) is computed at each base pair using bootstraps (1000 times by default) for each group profile. To reduce the effects of extreme coverage values, a data binning strategy with customizable bin sizes, is applied before bootstrapping. Visually, the confidence interval is represented by a ribbon which includes an editable percentage (default 95%) of the sampled values (see Supplementary Texts for more information on the bootstrap approach implemented in `metagene`). Using the `Imetagene` package, it is also possible to preview the regions as an interactive heatmap (Supplementary Fig. 1).

A unique feature of the `metagene` package is the implementation of a statistical comparison between profiles to detect differential enrichment. The comparison is done through a permutation test using metrics which are specified by the user that is not related to the confidence intervals calculated with bootstrapping . For each round of the permutation test, the metric value is calculated using two profiles obtained by randomly sampling the coverages used to compute the original profiles. The proportion of metric scores above the original score is used to calculate a p-value and determine if two profiles are significantly different (see Supplementary Texts for more details). By enabling the use of a diversity of metrics, the statistical comparison can be tailored to fit custom needs. To facilitate the identification of common patterns between two ChIP-Seq profiles, `similaRpeak` is proposed as a companion package to `metagene`. The `similaRpeak` package implements six pseudometrics specialized in pattern similarity detection (7.2C). The profile submitted to each pseudometric must respect certain editable criterias, specific to each pseudometric, to ensure that the calculation of the pseudometric is only made in presence of informative peaks and to limit the computation of extreme values. A description of each pseudometric is available in S2 Table. Lastly, we developed a graphical user interface powered by Shiny (Chang et al., 2016), `Imetagene`. This graphical interface was developed to facilitate the use of `metagene` without R programming experience. Taken together, this set of software is used to quickly compare multiple region groups to discover enrichment patterns that would otherwise be missed when looking at individual regions.

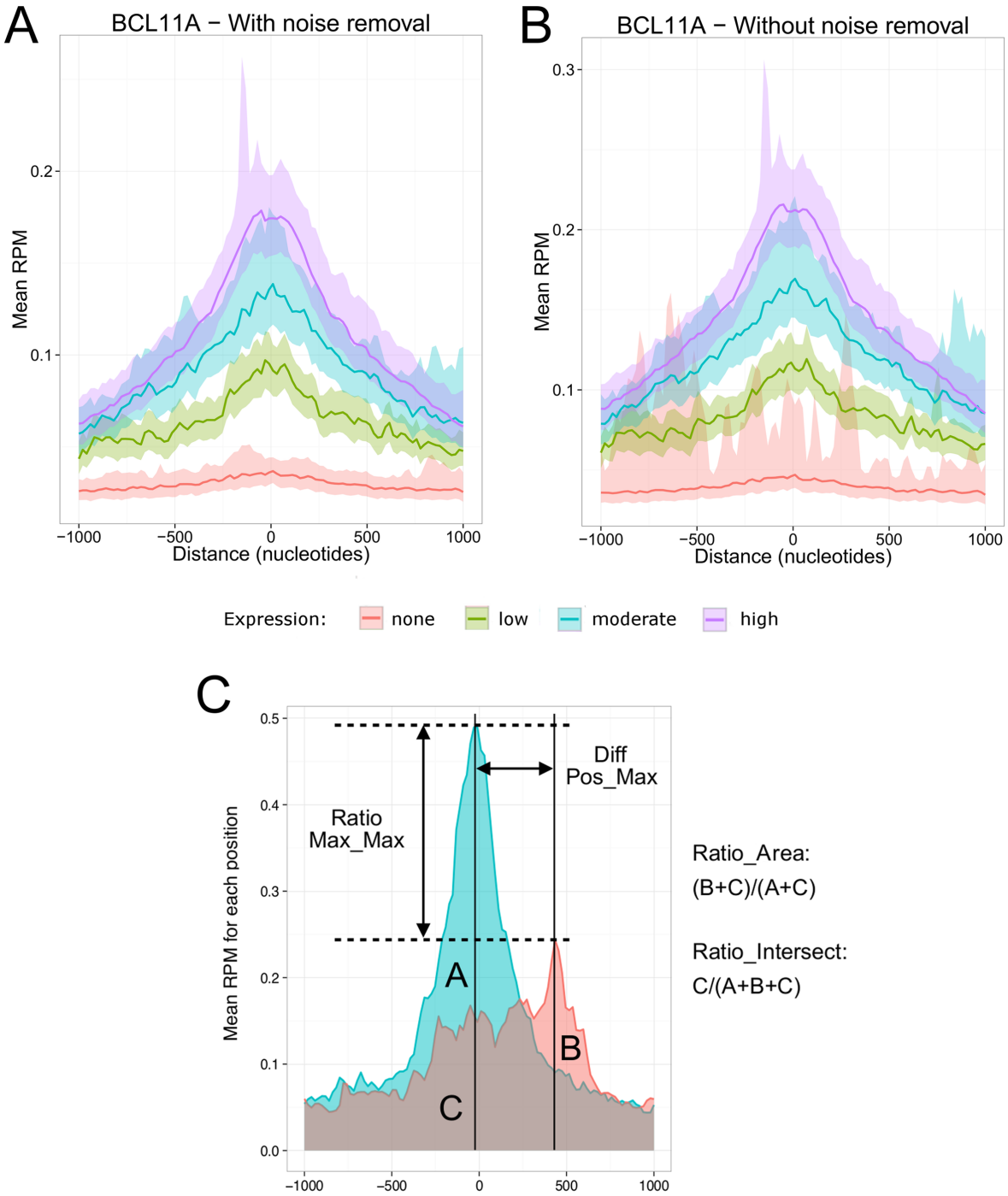


FIGURE 7.2 – Impact of noise removal and description of the pseudometrics. Metagene plots of the BCL11A transcription factor (A) with noise removal using the NCIS algorithm and (B) without noise removal. The x-axis is centered on enhancers and promoters ±1000bp. The y-axis represents the mean occupancy normalized in reads per million (RPM). Each line represents the mean occupancy of the BCL11A replicates. Groups of transcriptional activity of enhancers or promoters are identified by different colors (red = no CAGE signal ; green = low CAGE signal ; blue = moderate CAGE signal ; purple = high CAGE signal ; see Supplementary Texts). Ribbons represent the 95interval of the mean calculated using 1000 bootstraps. (C) Description of some of the pseudometrics implemented in the 'similaRpeak' packages.

7.8 Results

Proper spatiotemporal transcription requires the complex interplay of transcription factors, cofactors and chromatin regulators at noncoding regulatory regions (Ong and Corces, 2011, Vernimmen and Bickmore (2015)). Indeed, enhancer and promoter regions recruit regulatory factors to modulate the recruitment, initiation, pause-release and elongation of the RNA polymerase II (Pol II) (Lenhard et al., 2012, Plank and Dean (2014)). During the transcriptional process, both enhancer and promoter regions are transcribed (Kim et al., 2010, De Santa et al. (2010),Andersson et al. (2014)). Here we use the `metagene` package to correlate the recruitment of regulatory factors at enhancer and promoter regions with their transcriptional output.

7.8.1 Data collection and `metagene` analyses

To define the contribution of transcription factors and cofactors to the transcriptional activity of promoters and enhancers, we gathered the publically available data generated in GM12878 B-lymphocytes (106 available experiment datasets ; 276 alignment files, information in S1 Table). Promoters regions were obtained using the Bioconductor's TxDb.Hsapiens.UCSC.hg19.knownGene package (Lawrence et al., 2013) and enhancers were downloaded from the Fantom5 database (Lizio et al., 2015). Robust enhancer and promoter regions were defined by regions with at least one robust transcription start site (TSS) in the Fantom5 database. Finally, the regions were stratified into four groups based on their cap analysis of gene expression (CAGE) levels (Andersson et al., 2014) : "no expression", "low expression", "moderate expression" and "high expression" (see Supplementary Texts).

7.8.2 Pol II and the general transcription factors levels correlate with transcriptional activity

To validate our transcriptional stratification of enhancers and promoters, we surveyed the occupancy of total Pol II and the general transcription factors (GTFs), in function of the transcriptional activity (Sainsbury et al., 2015, Murakami et al. (2013),Roeder (1998)). As expected, transcriptional levels of enhancer and promoter regions correlated with recruitment of Pol II (7.3A and Supplementary Figs. 2-3), TAF1 (7.3B), and TBP (Supplementary Fig. 4). Histone marks associated with active enhancers (H3K27ac) and with active promoters (H3K4me3) showed a similar pattern (Supplementary Fig. 5). The RATIO INTERSECT pseudometric, which calculates the ratio of the area under

the intersection of two profiles with the total area, was used to compare the coverage between each group (S3 Table). The pseudometric value tends to 1 as the similarity between profiles increases. The statistical analyses confirmed that an increase in transcriptional activity correlates with an increase in the Pol II machinery (permutation p-value 0.001). In addition, the GTFs followed the same correlation with transcriptional activity. These results demonstrate that `metagene` and `similaRpeak` are able to distinguish patterns associated with different levels of transcription activity in a large number of samples by using robust metrics. Together, they offer an excellent tool to investigate the relationship between recruitment of regulatory factors and transcriptional activity.

7.8.3 Differential recruitment of regulatory factors at promoter and enhancer regions

While Pol II and GTFs activities are directly linked to the transcriptional output, the importance of each individual regulatory factor for the transcription process is not well understood. To assess the quantitative recruitment of transcription factors, cofactors and chromatin regulators at cis-regulatory elements as a function of the transcriptional activity, we evaluated the occupancy of regulatory factors, histone modifications and DNAse hypersensitive sites in GM12878 cells. Interestingly, we observed two distinct recruitment patterns at promoter and enhancer regions. Indeed, a "gradient effect" was observed when the occupancy level of a factor correlated with the transcriptional activity (7.3A-B) while a "threshold effect" refers to factors reaching a plateau in their occupancy prior to maximal transcriptional activity (7.3C). We defined a "threshold effect" as a ratio between the intersection area and the total area of the two profiles (RATIO INTERSECT) superior or equal to 0.85 between the high and moderate CAGE signal group. Overall, 44.6% of factors showed a "threshold effect" at enhancer regions while only 19.8% were observed at promoter regions (Supplementary Fig. 6 ; p-value = 0.0048 , Welch's Two Sample t-test). For example, the transcription factor ELF1 levels correlated with the transcriptional activity at promoters regions (RATIO INTERSECT = 0.66), but not at enhancers regions (RATIO INTERSECT = 0.88) (7.3C). A total of 35 regulatory factors including IRF3 and IRF4 (involved in interleukin regulation (Fitzgerald et al., 2003, Rengarajan et al. (2002))) and cofactors like SMC3 and EP300 (Supplementary Figs. 7 and 8) were identified with a similar dichotomy (see S3 Table for a complete list). These results highlight a differential requirement of regulatory factors at enhancer and promoter regions in relation to transcriptional activity.

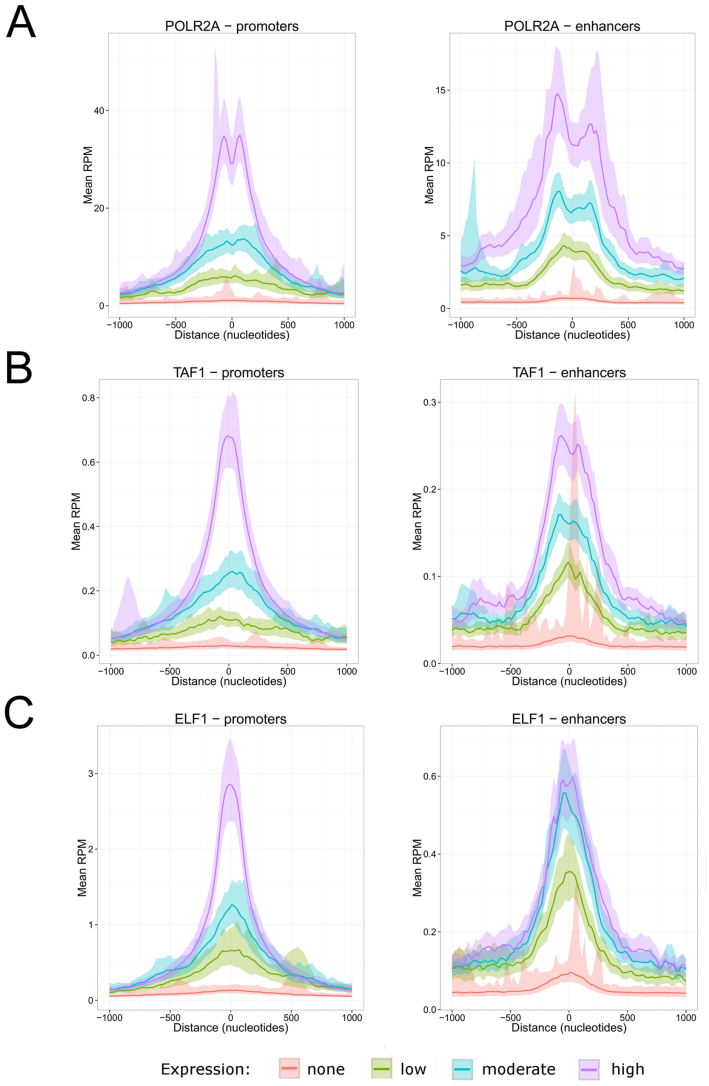


FIGURE 7.3 – Metagene profiles in enhancer and promoter regions. (A) POLR2A, the largest subunit of Pol II. (B) TAF1, a general transcription factor. (C) ELF1, a transcription factor. The x-axis is centered on enhancers and promoters ±1000bp. The y-axis represents the mean occupancy normalized in reads per million (RPM). Each line represents the mean occupancy of the factor replicates. Groups of transcriptional activity of enhancers or promoters are identified by different colors (red = no CAGE signal ; green = low CAGE signal ; blue = moderate CAGE signal ; purple = high CAGE signal). The ribbons represent the 95 percent confidence interval of the mean calculated using 1000 bootstraps.

7.8.4 Threshold versus gradient effects

Differential recruitment of regulatory factors at promoter and enhancer regions raises mechanistic questions. We are proposing different models to explain the "gradient" and "threshold" effects. For the "threshold effect", mostly observed at enhancer regions, the regulatory factors are potentially working as "on/off" switches. In that model, once a predetermined level is achieved for a specific transcription factor or cofactors, the transcriptional contribution is maximized (7.3B, Supplementary Figs. 7-8 and S3 Table). Extrapolation of this model suggests that an accumulation of different regulatory factors is required to achieve maximal transcriptional output at enhancer regions. This idea is corroborated by observations of dozens of transcription factors at enhancers regions in mammalian cells (Arnosti and Kulkarni, 2005). For the "gradient effect" mostly observed at promoter regions, we are considering two models : i) the regulatory factor directly contributes to Pol II transcriptional activity or ii) the "gradient effect" corresponds to the signal accumulation of multiple enhancers connecting to a promoter region through long distance interactions. These models are not mutually exclusive, but the latter is supported by evidence of an average of 4.9 enhancers connecting per promoter (Andersson et al., 2014) in addition to a positive correlation between the number of connections and the transcriptional output (Schoenfelder et al., 2015). Taken together, our results establish different recruitment patterns of regulatory factors at enhancers and promoters.

7.8.5 Other applications of `metagene`

In addition to the current study, the `metagene` package will be usable for multiple applications. For instance, the `metagene` package will be suitable to study differential recruitment in different classes of regulatory elements. For instance, enhancers and promoters regions could be stratified by functional types instead of expressions levels, such as the chromatin states (Kundaje et al., 2015). The enrichment patterns of a transcription factor following drug treatment or an infection could also be analyzed with `metagene` to provide molecular insights into the mechanism of action. Additionally, the dynamic of transcription factors recruitment could be studied using time course datasets. Future studies will reveal new details on the mechanisms of recruitment of regulatory factors and will help in understanding the similarities and dissimilarities between the various classes of regulatory elements.

7.9 Availability and Future Directions

The `metagene` package, the graphical interface `Imetagene`, and the companion package `similaRpeak` are available on Bioconductor with documentation and an example dataset. These packages perform a thorough evaluation of the similarities or dissimilarities of the aggregated signal of region groups. For the current version, the region groups are based on annotations in order to test specific scientific hypotheses. Next, we will be implementing clustering algorithms (as a part of a machine learning strategy) to cluster regions based directly on their occupancy patterns to provide an exploratory approach.

7.10 Supporting information

S1 Fig. **Imetagene interactive heatmap representation.**

After the matrices are computed, the Imetagene package can be used to explore the matrix-associated with each experiment to visualize the coverages of the regions.

`doi:10.1371/journal.pcbi.1004751.s001`

(PDF)

S2 Fig. **Metagene plots of RNA Pol II phosphorylated at serine 2 (POLR2AphosphoS2) in promoters and enhancers.**

The x-axis is centered on enhancers and promoters ±1000bp. The y-axis represents the mean occupancy normalized in reads per million (RPM). Each line represents the mean occupancy of POLR2Aphosphos2. Groups of transcriptional activity of enhancers or promoters are identified by different colors (red = no CAGE signal; green = low CAGE signal; blue = moderate CAGE signal; purple = high CAGE signal; see (S1 Text)[`http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004751#pcbi.1004751.s012`]). Ribbons represent the 95% confidence interval of the mean calculated using 1000 bootstraps.

`doi:10.1371/journal.pcbi.1004751.s002`

(PDF)

S3 Fig. **Metagene plots of RNA Pol II phosphorylated at serine 5 (POLR2AphosphoS5) in promoters and enhancers.**

The x-axis is centered on enhancers and promoters ±1000bp. The y-axis repre-

sents the mean occupancy normalized in reads per million (RPM). Each line represents the mean occupancy of POLR2Aphosphos5. Groups of transcriptional activity of enhancers or promoters are identified by different colors (red = no CAGE signal; green = low CAGE signal; blue = moderate CAGE signal; purple = high CAGE signal; see (S1 Text)[http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004751#pcbi.1004751.s012]). Ribbons represent the 95% confidence interval of the mean calculated using 1000 bootstraps.

doi:10.1371/journal.pcbi.1004751.s003

(PDF)

S4 Fig. **Metagene plots of the general transcription factor TBP at promoters and enhancers.**

The x-axis is centered on enhancers and promoters ±1000bp. The y-axis represents the mean occupancy normalized in reads per million (RPM). Each line represents the mean occupancy of TBP. Groups of transcriptional activity of enhancers or promoters are identified by different colors (red = no CAGE signal; green = low CAGE signal; blue = moderate CAGE signal; purple = high CAGE signal). The ribbons represent the 95% confidence interval of the mean calculated using 1000 bootstraps.

doi:10.1371/journal.pcbi.1004751.s004

(PDF)

S5 Fig. **Metagene plots of H3K27ac at enhancers and H3K4me3 at promoters.**

The x-axis is centered on enhancers and promoters ±1000bp. The y-axis represents the mean occupancy normalized in reads per million (RPM). Each line represents the mean occupancy of the histone mark. Groups of transcriptional activity of enhancers or promoters are identified by different colors (red = no CAGE signal; green = low CAGE signal; blue = moderate CAGE signal; purple = high CAGE signal). The ribbons represent the 95% confidence interval of the mean calculated using 1000 bootstraps.

doi:10.1371/journal.pcbi.1004751.s005

(PDF)

S6 Fig. **Boxplot of RATIO INTERSECT values for 106 experiments in GM12878.**

The RATIO INTERSECT was calculated using the moderate CAGE signal and high CAGE signal groups.

doi:10.1371/journal.pcbi.1004751.s006

(PDF)

S7 Fig. **Metagene plots of the cofactor SMC3 at promoters and enhancers.**

The x-axis is centered on enhancers and promoters ±1000bp. The y-axis represents the mean occupancy normalized in reads per million (RPM). Each line represents the mean occupancy of SMC3. Groups of transcriptional activity of enhancers or promoters are identified by different colors (red = no CAGE signal ; green = low CAGE signal ; blue = moderate CAGE signal ; purple = high CAGE signal). The ribbons represent the 95% confidence interval of the mean calculated using 1000 bootstraps.

doi:10.1371/journal.pcbi.1004751.s007

(PDF)

S8 Fig. **Metagene plots of the cofactor EP300 at promoters and enhancers.**

The x-axis is centered on enhancers and promoters ±1000bp. The y-axis represents the mean occupancy normalized in reads per million (RPM). Each line represents the mean occupancy of EP300. Groups of transcriptional activity of enhancers or promoters are identified by different colors (red = no CAGE signal ; green = low CAGE signal ; blue = moderate CAGE signal ; purple = high CAGE signal). The ribbons represent the 95% confidence interval of the mean calculated using 1000 bootstraps.

doi:10.1371/journal.pcbi.1004751.s008

(PDF)

S1 Table. **Description of the 276 bam files used in this article.**

Experiment accession : unique identifier of the experiment. File accession : unique identifier of the file. Target : the name of the factor that was targeted for immunoprecipitation. Controls : the experiment accession of the recommended controls. Biosample name : the cell type. Assembly : the version of the genome used for the alignment. Href : the URL to download the file.

doi:10.1371/journal.pcbi.1004751.s009

(CSV)

S2 Table. **Description of similaRpeak's pseudometrics.**

Pseudometric : the name of the pseudometric. Definition : the description of the metric. Threshold : criteria that can be set by the user to avoid calculating the value of a pseudometric that would return nonsensical results (division by zero, etc...).

doi:10.1371/journal.pcbi.1004751.s010

(XLSX)

S3 Table. **Classification of GM12878 factors.**

The classification of the 106 regulatory factors in "gradient" or "threshold". Target : the name of the target. Type : enhancer or promoter. RATIO_INTERSECT : the RATIO_INTERSECT score calculated using the moderate and high CAGE signal groups. Class : "gradient" or "threshold".

doi:10.1371/journal.pcbi.1004751.s011

(CSV)

S1 Text. **Data collection : Details of the data collection procedure.**

Bootstrap : Description of the bootstrapping steps. Permutation : Details of the permutation procedure in Metagene and similaRpeak.

doi:10.1371/journal.pcbi.1004751.s012

(PDF)

7.11 Acknowledgments

We thank Frédéric Fournier for his input and his assistance with the preparation of the manuscript. Computations were made on Colosse, the supercomputer from Université Laval, managed by Calcul Québec and Compute Canada. Author Contributions

Conceived and designed the experiments : CJB FCL ADe RS AL PB SB ADr. Analyzed the data : CJB ADe RS AL PB ADr. Wrote the paper : CJB FCL ADe SB ADr. Designed the software : CJB FCL ADe RS AL PB ADr.

Chapitre 8

NGS++ : une librairie pour le développement rapide de prototypes d'outils épigénomiques

8.1 Journal

Cet article a été publié dans la revue *Oxford Bioinformatics* en août 2013. La license de ce journal permet aux auteurs d'utiliser l'article en partie ou en totalité dans le contexte d'une thèse.

8.2 Résumé

Les données brutes produites par les séquenceurs de nouvelle-génération sont extrêmement volumineuses et il peut être difficile de les manipuler efficacement, tout particulièrement dans le cas du développement de nouveaux outils. La librairie NGS++ a donc été développée pour faciliter le prototypage d'outils pour l'analyse de données génomiques. En utilisant des fonctionnalités du langage C++ moderne (C++11), il permet de générer rapidement de nouvelles fonctionnalités, sans compromis au niveau de la vitesse d'exécution.

8.3 Contributions

Je suis coauteur de la librairie NGS++. Ma contribution est dans la mise en place de l'interface finale, la correction de certains bogues, une participation importante dans la

mise en place de la suite de tests unitaires, l'écriture de la documentation, le développement de la page internet (www.ngsplusplus.com) et dans la rédaction de l'article.

Pour plus de détails sur la contribution des autres auteurs, voir la section Avant-propos.

8.4 Contenu

NGS++ : a library for rapid prototyping of epigenomics software tools

Alexei Nordell Markovits[1],&, Charles Joly Beauparlant[2],&, Dominique Toupin[3], Shengrui Wang[3], Arnaud Droit[2],* and Nicolas Gevry[1],*

[1]Department of Biology, Université de Sherbrooke, Sherbrooke, Quebec J1K 2R1, [2]Department of Molecular Medicine, Centre de Recherche du CHU de Quebec, Université Laval, Quebec, Quebec G1V 4G2 and [3]Department of Computer Science, Université de Sherbrooke, Quebec J1K 2R1, Canada

*To whom correspondence should be addressed.

&The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

8.5 Abstract

Motivation : The development of computational tools to enable testing and analysis of high-throughput–sequencing data is essential to modern genomics research. However, although multiple frameworks have been developed to facilitate access to these tools, comparatively little effort has been made at implementing low-level programming libraries to increase the speed and ease of their development.

Results : We propose NGS++, a programming library in C++11 specialized in manipulating both next-generation sequencing (NGS) datasets and genomic information files. This library allows easy integration of new formats and rapid prototyping of new functionalities with a focus on the analysis of genomic regions and features. It offers a powerful, yet versatile and easily extensible interface to read, write and manipulate multiple genomic file formats. By standardizing the internal data structures and presenting a common interface to the data parser, NGS++ offers an effective framework for epigenomics tool development.

Availability : NGS++ was written in C++ using the C++11 standard. It requires minimal efforts to build and is well-documented via a complete docXygen guide, online documentation and tutorials. Source code, tests, code examples and documentation are available via the website at `http://www.ngsplusplus.ca` and the github repository at `https://github.com/NGS-lib/NGSplusplus`.

Contact : `nicolas.gevry@usherbrooke.ca` or `arnaud.droit@crchuq.ulaval.ca`

8.6 Introduction

Previous years have witnessed an explosion in the amount of data produced using next-generation sequencing (NGS) technologies, as exemplified by the ENCODE project (Consortium et al., 2012). However, analysis of these enormous datasets (easily >100 GB) requires the use of a new generation of computational tools. As the quantity of data produced by NGS machines increases, so will the time spent on developing new tools. Although substantial efforts have been made at integrating them into user-friendly frameworks such as Galaxy (Giardine et al., 2005) or GeneSpace (Genome Space)(, 2013), relatively little effort has gone into providing the groundwork needed to increase the productivity of NGS developers, such as libraries and using standardized formats. Improvement in these areas would allow developers to greatly accelerate the speed at which they design and deploy new analysis software.

Although certain tool suites, such as BEDtools (Quinlan and Hall, 2010) and BAMtools (Barnett et al., 2011), offer a library or API to assist developers, these are generally aimed at giving access to the existing tool functionality rather than facilitating development of new ones. As such they are highly specialized. The SeqAn library (Döring et al., 2008) offers functionality for the developmental of future tools, but it specializes in sequence analysis rather than genomic regions and features. Our proposed library, NGS++, aims to fill this gap by offering a powerful set of generic and flexible options to accelerate development and prototyping of epigenomics analysis tools.

8.7 Approach

It is impossible to predict the entirety of future needs for NGS data analysis. As such, NGS++ focuses on being a customizable and generic library that facilitates the prototyping and implementation of new functionalities via a transparent data interface.

In this section, we summarize the three main components of NGS++ : (i) file format management, (ii) data manipulation and (iii) functional operators.

Dealing with the wealth of existing file formats is a time consuming task. NGS++ offers a simple interface to parse and write in many frequently used genomics file formats (BED, GFF/GTF, Sam, Wig, bedGraph) using a generic data structure named Tokens that contain a number of standard features of genomic data entries (eg : Start/End positions, position value and mapping quality). Additionally, the user can define 'on-the-fly' custom formats to deal with the plethora of datasets that do not respect format specifications, and BAM format is supported via integration of the BamTools API. The conversion between most supported formats is a trivial task :

```
uParser parser("filename.sam", "SAM");
uWriter writer("filename.bed", "BED");
while (!parser.eof())
    write.writeToken(parser.getNextEntry());
```

The internal structure of the library is divided into a three-tiers hierarchy separating a genomic dataset by unique scaffolds with each of them containing any number of contigs. This hierarchy is represented in (8.1), and each level offers a number of functions that can be extended via inheritance for specialized data manipulation. Loading datasets is done through integration with our Parser class and allows the user to easily load data :

```
TagExp.loadWithParser(inputStream, "SAM");
RegionExp.loadWithParser(inputStream, "BED");
```

NGS++ offers a powerful set of generic functions to compare, sort, merge and modify the previously loaded data. These include typical operations such as overlapping, merging and comparing that allows the user to easily filter his data as needed. The following would return tags overlapping a BED file :

```
auto fExp=TagExp.getOverlapping(RegionExp);
```

Additionally, the majority of these operators can be used on any given feature of the data objects, including features added by the user via inheritance. This example sorts and counts a subset based on a new object feature :

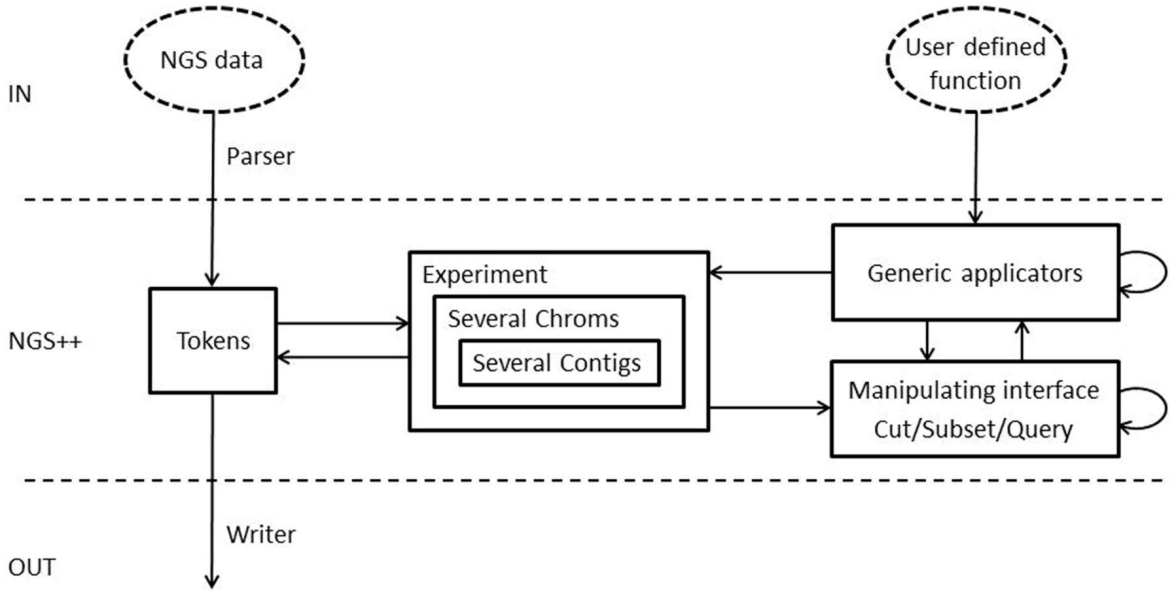


FIGURE 8.1 – Typical workflow of the NGS++ library. Data are read by our Parsing interface then filtered as needed. User-defined functions are executed via our operators, and the transformed data are stored via our Writer interface

```
RegionExp.sortSites([](uRegion item1, uRegion item2){return item1.Score <
item2.Score}, &uRegion::getScore, &uRegion::getScore)
int count = Region.Exp.getSubsetCount(0.5, 0.8);
```

As shown earlier in the text, this is greatly facilitated by the inclusion of anonymous lambda functions in the C++11 standard. Using these flexible operators allows the experienced developer to implement powerful modifications, whereas the default genomic interval operators are easily usable by all.

Finally, to accommodate specific analysis needs, NGS++ offers an interface to run developer defined functions and transformations on the selected data. This allows the developer to concentrate on the function he wishes to implement without having to spend time on the underlying structure that will support it. Borrowing heavily from the functional programming paradigm, this allows for rapid prototyping and implementation. In the following we define and execute a function that allows us to trivially generate a histogram of contig sizes :

```
map<int,int> sizeOfContigs;
uTagsExp.applyOnSites([&](uTag Elem){sizeOfContigs[Elem.getSize()]++;});
```

Additional functional operators exist, allowing a variety of different operations on the dataset. This interface wraps many of the STL algorithms, enabling rapid parallelism via the OpenMP standard (Dagum and Menon, 1998).

8.8 Implementation

NGS++ is written in C++ using the C++11 standard. It offers a complete exception handling interface using the Boost exception class. A complete test suite is implemented using the Google test platform. It has been designed for a Linux environment using a C++11 compatible gcc compiler. Complete user guide, tutorial and discussion are available on the web page. Source code is hosted on GitHub.

8.9 Conclusion

Progress in the development of advanced bioinformatics analysis tools has undoubtedly been hindered by the lack of available programming frameworks. Our library aims to assist in filling this gap for the community of C++ epigenomic developers by giving them access to robust building blocks, thus reducing the time spent on development significantly. Our efforts are now focused on including additional genomic formats and on increasing the breath of our tutorials. Future developments will include the integration of mid-level reusable functions, such as similarity functions and normalization methods. The website provides a list of tutorials and commented working code examples, to assist developers in getting started with the library. Finally, the GitHub should facilitate the integration of suggestions and feedback from the community.

Funding : Natural Sciences and Engineering Research Council of Canada (NSERC) (to S.W.) ; Canadian Institutes of Health Research (CIHR) (to N.G.) ; Ministère du Développement Économique, Innovation et Exportation (MDEIE) (to A.D.). N.G. holds a Chercheur boursier (junior 1) award from the Fond de Recherche en Santé du Québec (FRSQ). A.D. holds a Réseau de médecine génétique appliquée (RMGA) salary award. C.J.B. holds a Centre de Recherche en Endocrinologie et Génomique Humaine (CREMOGH) award.

Conflict of Interest : none declared.

Chapitre 9

Collaborations

Au cours de mon doctorat, j'ai eu l'occasion de collaborer sur plusieurs projets ayant mené à des publications scientifiques. Ces collaborations m'ont permis de me familiariser avec d'autres méthodes en génomique en plus de me permettre d'utiliser mes outils.

9.1 Lemaçon et al. (2017)

Cet article présente le logiciel VEXOR développé par Audrey Lemaçon pour faciliter la recherche de variants causaux au sein de régions génomiques ciblées par des études d'association à l'échelle génomique (*Genome-wide association studies*, *GWAS*). Cet outil est codé en R et offre une interface graphique Shiny. Ma contribution à ce projet a été d'optimiser le code R, de participer à la création des figures.

9.2 Fournier et al. (2014)

Le logiciel rTANDEM est une adaptation de l'outil TANDEM (Craig and Beavis, 2004) à l'environnement de programmation R. Ce logiciel facilite l'identification des peptides à partir des spectres produits par spectrométrie de masse. J'ai été responsable de l'adaptation du code C++ pour permettre l'interface avec le langage R.

9.3 Fournier et al. (2016)

Cet article étudie la régulation de la transcription dans le contexte de cellules cancéreuses. Plus spécifiquement, il recherche les facteurs impliqués dans le recrutement du

complexe mediator et cohésine. J'ai participé aux analyses bioinformatiques de cette recherche.

Chapitre 10

Conclusion

10.1 Résumé

Dans ce document, j'ai présenté un ensemble d'outils que j'ai développés dans l'optique de faciliter l'intégration de jeux de données génomiques, notamment des données produites par ChIP-Seq. En effet, il est maintenant relativement facile de produire des jeux de données complexes, c'est-à-dire ayant plusieurs réplicats, conditions, individus voir mêmes différents types d'expériences. Plusieurs stratégies d'analyse sont disponibles pour analyser des échantillons individuels ou des paires d'échantillons. Par contre, lorsque le nombre de conditions augmente, beaucoup moins d'outils sont disponibles. De plus, il est souvent nécessaire d'inclure des inclure des données provenant de différents types d'expériences si on souhaite obtenir une vue d'ensemble plus complète des mécanismes de la régulation de la transcription.

L'intégration de l'information provenant de jeux de données complexes et hétérogènes est un problème très difficile, car chaque type d'expérience a ses propres caractéristiques et il n'est donc pas possible de fournir une solution unique. Les outils développés durant mon doctorat offrent donc une réponse à certaines facettes de cette problématique. Ils ciblent principalement les données produites par ChIP-Seq, mais ils permettent également d'inclure d'autres types de données. Le principe de l'outil est de regrouper les régions similaires pour représenter et analyser les profils d'enrichissement. Puisque les données de ChIP-Seq sont reconnues pour contenir beaucoup de bruit de fond, il est préférable de combiner les régions pour obtenir une meilleure puissance statistique.

Pour l'article de l'outil `metagene`, nous avons utilisé à la fois des données transcriptomiques et des données épigénomiques pour étudier la dynamique du recrutement des

protéines lors de l'activation d'éléments régulateurs. Cette publication a été possible grâce aux nombreux jeux de données produits par les consortiums ENCODE et Fantom. Malgré l'excellent travail de ces équipes de recherche, nous croyons qu'il est possible d'améliorer l'accès à ces données en offrant des outils appropriés. `ENCODExplorer` permet de rechercher facilement dans la base de données d'ENCODE, offre une interface avec le langage de programmation R et s'assure de l'intégrité des données téléchargées. De plus, grâce aux versions de l'outil et à la possibilité de créer des scripts pour la recherche favorisent la reproductibilité.

10.2 Perspectives

Au cours de cette thèse, j'ai présenté les étapes et les contraintes pour l'élaboration d'une suite d'outil pour faciliter les analyses intégratives d'expériences de ChIP-Seq. Pour mes travaux futurs en lien avec ces outils, je considère qu'il y aura trois avenues principales à considérer : 1) Améliorations techniques et maintenance, 2) ajout d'information biologique et 3) l'étude de la régulation de la transcription.

10.2.1 Améliorations techniques et maintenance des outils

Nouvelle structure de données

C'est en développant le paquet `metagene` que je me suis familiarisé avec le langage de programmation R. C'est donc pour cette raison que cet outil en est déjà à sa deuxième version. En rétrospective, certains choix dans le design initial de l'outil ont mené à un code plus complexe que nécessaire et ont contribué à rendre certaines étapes de calculs moins efficaces. En remplaçant la structure de données principale, qui stocke les comptes normalisés pour passer d'une liste de liste à une structure plus compacte de type `data.table`, il serait possible de simplifier significativement le code et ainsi minimiser le risque d'y retrouver des bogues tout en augmentant la performance.

Optimisation des opérations critiques

L'intégration des paquets `GenomicRanges` et `GenomicAlignments`, produits et maintenus par l'équipe de Bioconductor, a permis d'accélérer grandement l'efficacité de l'outil `metagene`. Le temps nécessaire pour les analyses présentées dans la section 7 sont passés à environ une heure dans la première version de l'outil à moins de 10 minutes avec la version la plus récente. Je souhaite optimiser certaines étapes clés qui représentent une

fraction importante du temps de calcul, notamment le calcul du nombre de séquences, le calcul de la couverture et le bootstrap.

L'intégration de la librairie `NGS++` devrait permettre de coupler certaines opérations et rendre les fonctions bas niveaux plus efficaces en les adaptant spécifiquement pour les problématiques rencontrées par `metagene`. Les opérations statistiques qui ne sont pas liées directement à la génomique pourront être optimisées en les réimplantant en langage C.

Maintenance des outils

Pour l'outil `ENCODExplorer`, la majorité des opérations se déroulent dans un délai raisonnable, excepté la mise à jour de la base de données, mais le facteur limitant est l'API REST fourni par ENCODE. L'interface de l'outil est mature et ne va donc pas demander de changements majeurs. La difficulté principale est au niveau de la maintenance de l'outil, car la base de données ENCODE est régulièrement modifiée pour ajouter de nouvelles informations et dans certains cas, ces changements ne peuvent pas être directement intégrés dans la structure de données d'`ENCODExplorer`. Puisque la troisième phase d'ENCODE est encore en cours, des changements sont encore à prévoir.

Traitement des données de transcriptomique

Une prochaine étape dans développement de `metagene` sera de l'adapter pour pouvoir l'utiliser dans le contexte de la transcriptomique. En modifiant légèrement le code de `metagene` il serait possible l'utiliser pour visualiser des données de RNA-Seq. La différence majeure à tenir compte est l'épissage alternatif. L'approche à privilégier serait donc de concaténer la couverture des exons pour obtenir un seul profil par gène. En regroupant tous les réplicats pour former une seule courbe par gène, il serait possible de vérifier si la couverture est similaire pour tout le gène ou s'il y a des biais d'enrichissement. Il serait également possible de combiner les réplicats de chaque condition pour visualiser les différences d'expression d'un gène entre des groupes qui ont été ciblés comme étant significativement différentes par d'autres outils.

Un prototype est déjà disponible sur la page github de `metagene` (`https://github.com/CharlesJB/metagene/tree/rna`). Une fois l'interface stabilisée et les tests complétés, cette fonctionnalité devrait être publiée dans Bioconductor.

10.2.2 Ajout d'information biologique

Il a été démontré que le pourcentage de CG peut biaiser le séquençage. Il serait donc intéressant de pouvoir afficher cette information grâce aux graphiques produits par `metagene`. On pourrait donc rapidement vérifier si tous les groupes comparés ont une composition en GC similaire. Grâce à l'approche de comparaison des courbes par permutation, l'utilisateur pourrait même tester s'il y a des différences significatives dans la composition en GC.

Pour l'implémentation, le plus simple serait de demander à l'utilisateur de fournir un des paquets Bioconductor de type `BSgenome` qui peut être utilisé directement pour extraire les séquences nucléotidiques et produire les matrices de séquences nécessaires pour calculer le pourcentage de GC à chaque position. L'information pourrait également être fournie sous forme de fichier fasta correspondant au génome de référence ayant la même version que celui utilisé pour les analyses.

D'une manière similaire au pourcentage GC, d'autres informations sur la composition des régions pourraient être intégrées aux analyses. Par exemple, il serait aussi pertinent d'afficher l'information sur les niveaux de conservation moyenne de chaque position dans les groupes de régions. Finalement, une autre source de biais importante est la mappabilité. Si un groupe de régions contient un plus grand nombre de régions de faible mappabilité que les autres, cela pourrait affecter les résultats de leur comparaison. Les scores phastCons de UCSC sont disponibles pour les analyses de conservation et plusieurs outils tels que PICS (Zhang et al., 2011) offrent des scores de mappabilité précalculés.

10.2.3 Étude de la régulation de la transcription

Mes recherches doctorales se sont concentrées principalement au développement d'outils pour l'intégration de jeux de données génomique et principalement pour les expériences de type ChIP-Seq. Lors du développement d'un outil tel que `metagene`, j'ai évidemment eu beaucoup d'idée de projets que je souhaiterais mener pour la suite de ma carrière en recherche. Cette dernière section présente donc quelques-unes de ces idées.

Évaluation des approches de normalisation

Certaines approches de normalisation sont plus pertinentes dans un contexte où on compare différents groupes de régions pour un même échantillon alors que d'autres auront un impact plus important lorsqu'on veut comparer le même groupe de régions

entre plusieurs échantillons. Dans le cas de différents groupes de régions pour un même échantillon, le pourcentage de GC, la mappabilité et le niveau de conservation peuvent avoir un effet significatif, car chaque groupe de région n'aura pas nécessairement les mêmes propriétés. On s'attend par contre à ce que l'impact de l'efficacité de l'anticorps où les autres types de biais expérimentaux soient moins importants, car toutes les mesures de couvertures proviennent de la même expérience. À l'inverse, lorsqu'on compare le même groupe de régions à travers plusieurs expériences, on sait que les mesures telles que le pourcentage de GC, la mappabilité et le niveau de conservation seront les mêmes entre les échantillons comparés. Par contre, chaque échantillon peut avoir des biais expérimentaux qui lui est propre et il peut être nécessaire d'en tenir compte.

L'outil `metagene` offre la possibilité d'analyser l'impact des approches de normalisation sous un angle novateur et pourrait donc servir de base pour un projet qui permettrait de mieux comprendre les forces et les faiblesses de chaque approche présentement disponible, de développer de nouvelles méthodes et de définir des critères qui aideraient à choisir la solution optimale pour chaque situation.

Distinctions et similitudes entre amplificateurs et promoteurs

Plusieurs études ont démontré que les promoteurs peuvent parfois agir comme amplificateurs. De plus, certaines des caractéristiques des amplificateurs actifs, tels que la transcription bidirectionnelle, sont partagées par les promoteurs. Aussi, le consortium Fantom a ajouté un critère arbitraire pour retirer tous les amplificateurs situés à moins de 500 nucléotides d'un promoteur connu de sa liste finale d'amplificateurs. En d'autres termes, plusieurs régions promotrices étaient également considérées comme étant des amplificateurs selon l'algorithme de détection de Fantom. J'aimerais étudier plus en détail les régions promotrices qui ont des caractéristiques d'amplificateurs actifs pour essayer de mieux comprendre la distinction entre ces deux classes d'éléments. Pour ce faire, il faudrait étudier les mêmes sites à travers plusieurs types cellulaires pour mieux comprendre leur comportement et croiser avec des données de transcriptomiques permettant ainsi de mieux caractériser les différents états possibles.

Étude des classes d'amplificateurs et de promoteurs

Comme il a été mentionné brièvement dans la section 3, les promoteurs et les amplificateurs peuvent avoir plusieurs états. La suite d'outils présentée dans cette thèse offre une approche intéressante pour étudier plus en détail la régulation de la transcription

en permettant d'étudier la dynamique qui permet à une région de passer d'un état à un autre. Pour ce faire, le défi principal sera de trouver des jeux de données de ChIP-Seq produits à différents moments après un stimulus ou bien à différents stades de maturation cellulaire. Pour être en mesure de bien valider les résultats, il serait intéressant que les jeux de données soient avec des données de RNA-Seq pour faciliter la validation des résultats.

Étude de la dynamique du recrutement des facteurs de transcription

Il était possible de combiner l'information provenant de plusieurs jeux de données de ChIP-Seq pour tenter d'élucider la distribution spatiale des facteurs de transcription au sein d'une région génomique précise (voir par exemple les travaux de Nagy et al. (2016)). L'outil `metagene` pourrait être utilisé dans un contexte similaire, mais en regroupant plusieurs régions pour rechercher des patrons plus généraux. Le but de ce projet serait donc tout d'abord d'évaluer la faisabilité en reproduisant les résultats d'autres équipes. Ensuite, il s'agit de mettre en place de nouvelles approches pour automatiser les analyses et finalement de valider l'approche à grande échelle en analysant de nouveaux jeux de données.

Structure de la chromatine

Dans l'article de l'outil `metagene`, nous avons observé que certains facteurs ont un profil d'enrichissement de type *gradient* aux promoteurs, mais de type *seuil* aux amplificateurs. Nous avons émis l'hypothèse que l'enrichissement de type *gradient* de ces facteurs au niveau des promoteurs pourrait être expliqué soit par le fait que le recrutement de ces facteurs est bel et bien proportionnel à l'activité transcriptionnelle ou bien qu'on mesure indirectement l'enrichissement de plusieurs éléments amplificateurs qui se retrouvent à proximité suite au repliement de la chromatine.

Je crois qu'il serait très intéressant d'essayer de distinguer entre ces deux possibilités. En utilisant des expériences ayant des jeux de données pour les facteurs impliqués dans le repliement de la chromatine (les sous-unités de la cohésine par exemple) et des jeux de données de repliement chromatinien, on pourrait regrouper les régions selon qu'elles sont impliquées dans un repliement ou non. On pourrait ensuite comparer ces groupes de régions pour étudier la contribution des interactions indirectes au signal mesuré au niveau des promoteurs et des amplificateurs.

10.2.4 Positionnement des outils

Bien que les analyses classiques des données de ChIP-Seq (alignement et détection des régions enrichies) resteront probablement pertinentes pour de nombreuses années, on devrait s'attendre à voir de plus en plus d'outils qui visent l'intégration de plusieurs jeux de données hétérogènes. Un défi important de ce type d'analyse est de normaliser les résultats de manière à les rendre comparables d'une expérience à l'autre. Les programmes développés durant mon doctorat se positionnent de manière intéressante dans ce contexte, car ils permettent non seulement d'accéder rapidement à un grand nombre de jeux de données en se connectant aux bases de données de plusieurs grands consortiums, mais également en offrant une approche rigoureuse pour comparer les profils d'enrichissement. De plus, l'approche proposée peut être généralisée pour d'autres types de données tels que la transcriptomique, ce qui la place en position intéressante pour le développement des nouvelles approches pour les analyses de données de multiples omiques.

10.2.5 Prochaines avancées majeures

La prochaine étape majeure dans le domaine de l'étude de la régulation de la transcription par l'intégration de données de type omiques passera probablement par les données de conformation 3D de la chromatine. En effet, bien qu'il soit intéressant d'étudier les facteurs individuellement, il est aussi important de tenir compte du contexte génomique dans lequel ils se trouvent. Ainsi, il sera possible de "boucler la boucle" et d'évaluer directement l'impact des repliements chromatiniens sur les signaux détectés par les autres approches. Il sera donc important de suivre de près les progrès des approches expérimentales de ce domaine et d'adapter les outils existants en conséquence.

Annexe A

TABLEAU A.1 – Exemples de logiciels disponibles pour la détection des régions enrichies.

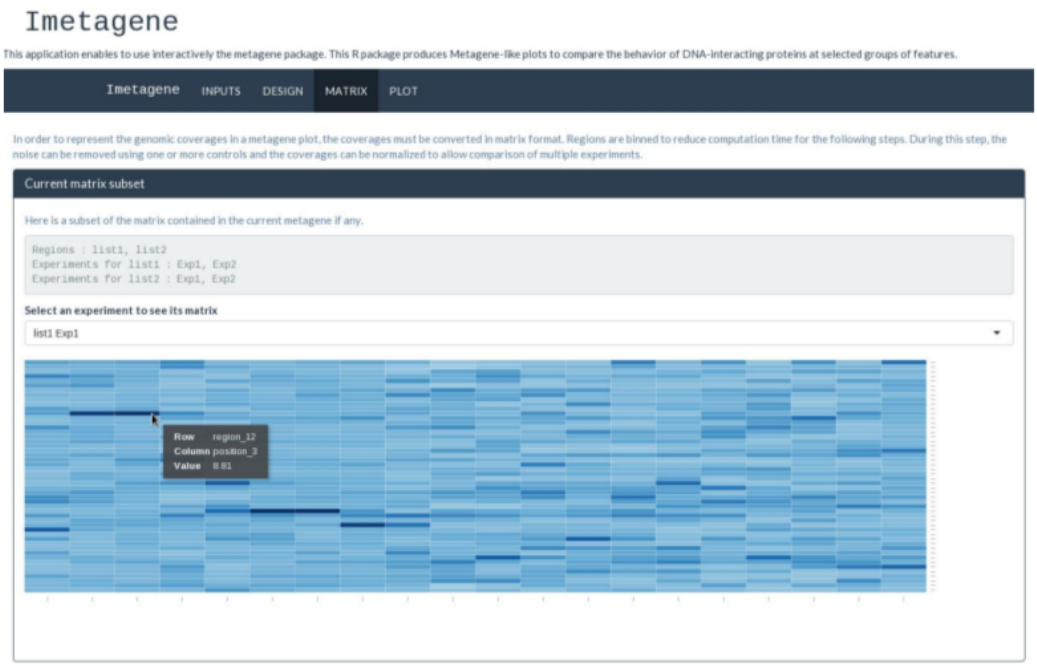
Nom	Référence
AREM	Newkirk et al. (2011)
BayesPeak	Spyrou et al. (2009)
BroadPeak	Wang et al. (2013)
CCAT	Xu et al. (2010)
ChIP-Seq analysis tools	Ambrosini et al. (2016)
ChIPSeq Peak Finder	Chen et al. (2008)
ChIPseqR	Humburg et al. (2011)
CisGenome	Ji et al. (2011)
Cistrome	Liu et al. (2011)
CSAR	Muiño et al. (2011)
CSDeconv	Lun et al. (2009)
DFilter	Kumar et al. (2013)
dPeak	Chung et al. (2013)
DROMPA	Nakato et al. (2013)
E-RANGE	Mortazavi et al. (2008)
EDD	Lund et al. (2014)
epic	Non disponible
F-Seq	Boyle et al. (2008)
FindPeaks 3.1	Fejes et al. (2008)
GEM	Guo et al. (2012)
GeneProf	Halbritter et al. (2012)
GeneTrack	Albert et al. (2008)
GLITR	Tuteja et al. (2009)

Nom	Référence
hiddenDomains	Starmer and Magnuson (2016)
HMCan	Ashoor et al. (2013)
HOMER	Brenner (2012)
Hpeak	Qin et al. (2010)
JAMM	Ibrahim et al. (2015)
Jmosaics	Zeng et al. (2013)
MACS	Zhang et al. (2008)
MICSA	Boeva et al. (2010)
MixChIP	Rautio and Lähdesmäki (2015)
MM-ChIP	Chen et al. (2011)
MOSAiCS	Kuan et al. (2011b)
MSPC	Jalili et al. (2015)
MUSIC	Harmanci et al. (2014)
NEXT-peak	Kim et al. (2013)
OccuPeak	de Boer et al. (2014)
PeakRanger	Feng et al. (2011)
peakrots	Elo et al. (2011)
PeakSeq	Rozowsky et al. (2009)
PeakZilla	Bardet et al. (2013)
PePr	Zhang et al. (2014)
Perm-seq	Zeng et al. (2015)
PICS	Zhang et al. (2011)
PING	Zhang et al. (2012)
polyaPeak	Wu and Ji (2014)
Q	Hansen et al. (2015)
qips	Gogol-Döring and Chen (2010)
QuEST	Valouev et al. (2008)
Ritornello	Stanton et al. (2015)
RSEG	Song and Smith (2011)
seqsite	Wang and Zhang (2011)
SICER	Xu et al. (2014)
SIPeS	Wang et al. (2010)
SISSRs	Jothi et al. (2008)
Sole-Search	Blahnik et al. (2010)

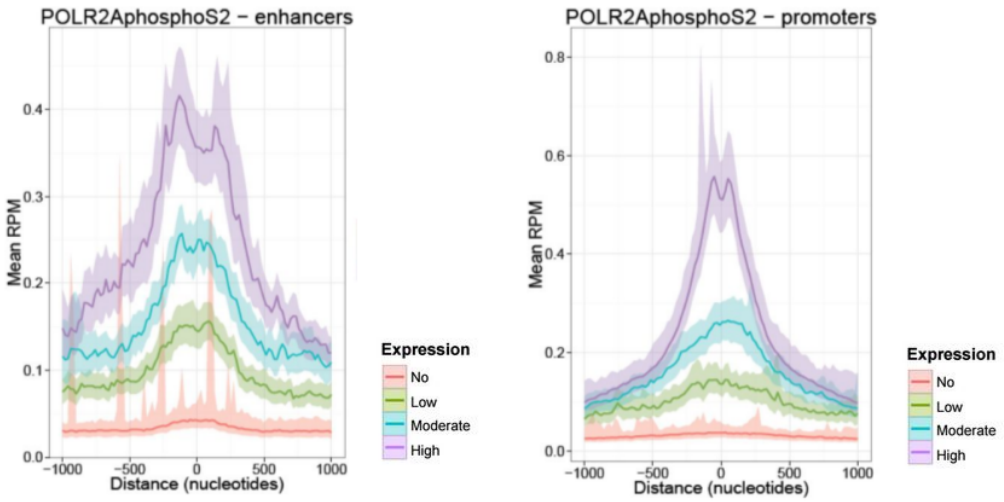
Nom	Référence
spp	Kharchenko et al. (2008)
SWEMBL	Wilder (2010)
T-KDE	Li et al. (2014)
T-PIC	Hower et al. (2011)
USeq	Nix et al. (2008)
W-ChIPeaks	Lan et al. (2011)
Zerone	Cuscó and Filion (2016)
ZINBA	Rashid et al. (2011)

Annexe B

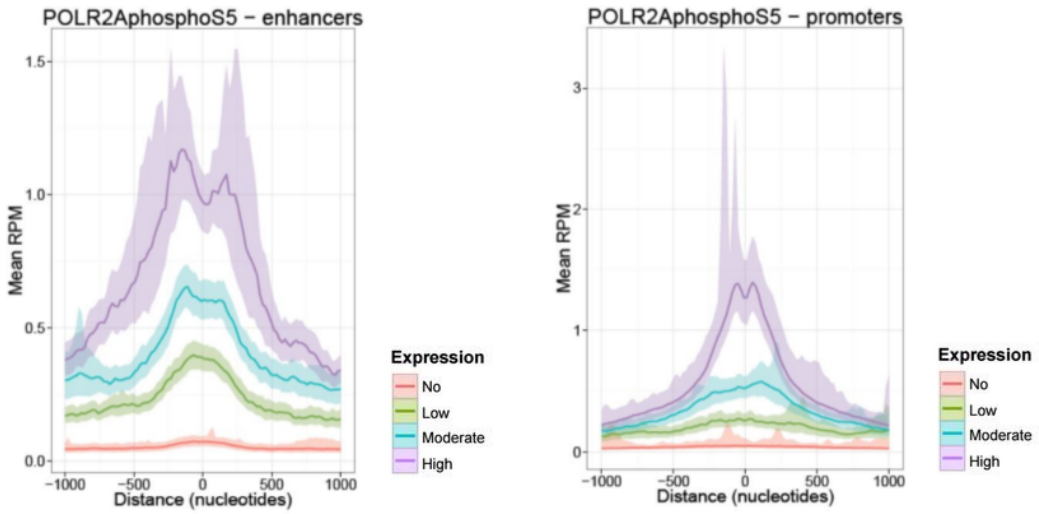
S1 Fig. Imetagene interactive heatmap representation



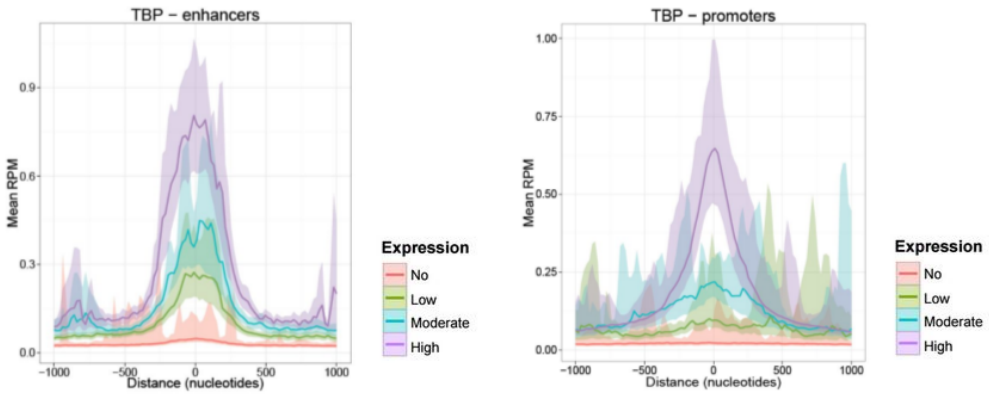
S2 Fig. Metagene plots of POLR2AphosphoS2 in promoters and enhancers



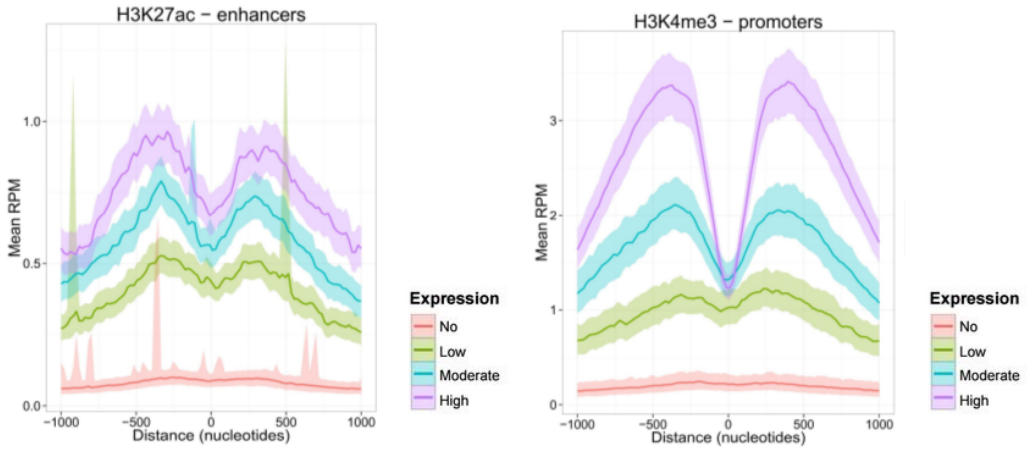
S3 Fig. Metagene plots of POLR2AphosphoS5



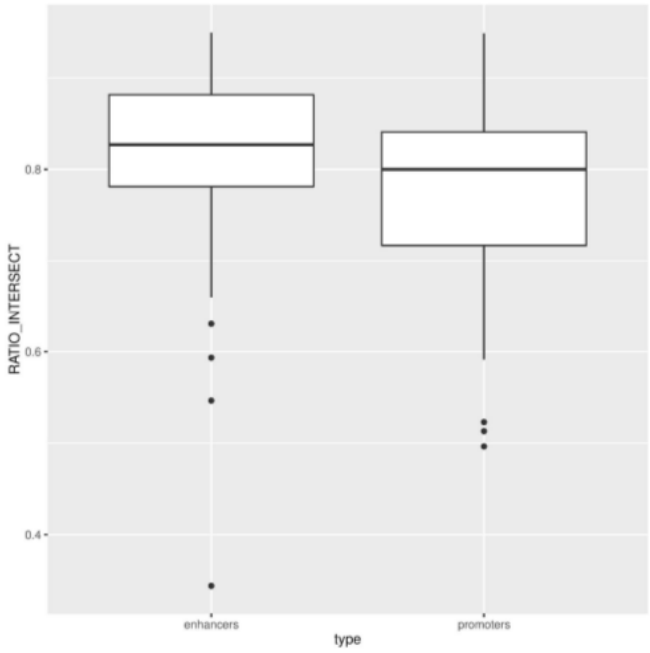
S4 Fig. Metagene plots of the general transcription factor TBP at promoters and enhancers



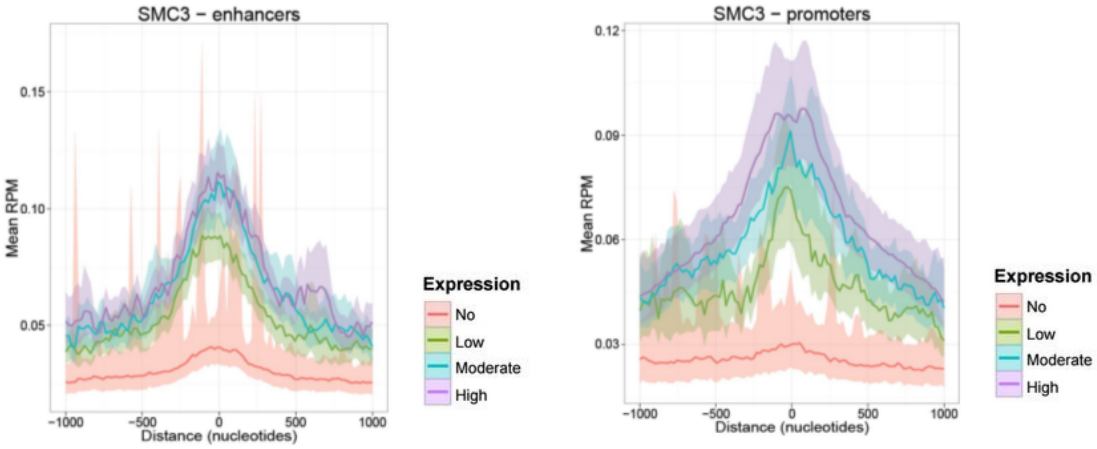
S5 Fig. Metagene plots H3K27ac in enhancers and H3K4me3 at promoters



S6 Fig. Boxplot of RATION_INTERSECT values for 105 experiments in GM12878



S7 Fig. Metagene plots of the cofactor SMC3 at promoters and enhancers



S8 Fig. Metagene plots of the cofactor EP300 at promoters and enhancers

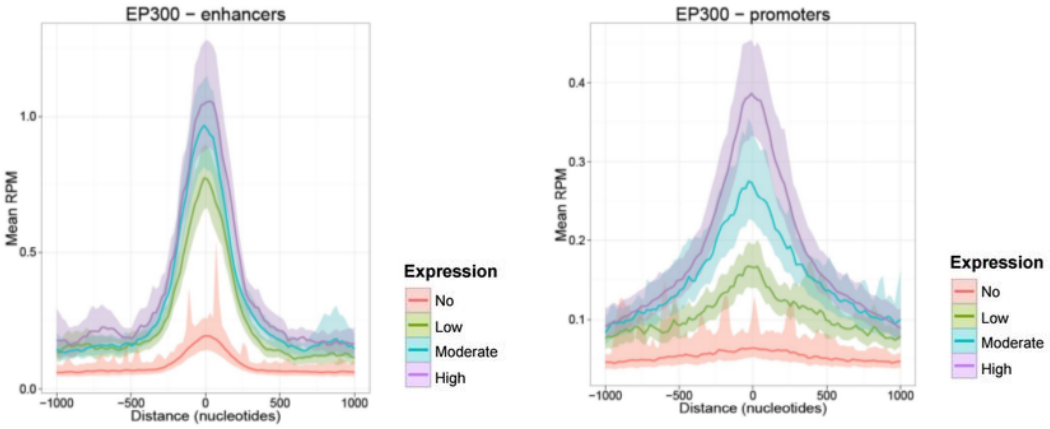


TABLEAU B.1 – Description of the 276 bam files used in this article.

experiment_accession	file_accession	target	controls	biosample_name	assembly
ENCSR000BQK	ENCFF000NRS	ATF2	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000BQK	ENCFF000NRT	ATF2	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000BJY	ENCFF000NSA	ATF3	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000BJY	ENCFF000NSD	ATF3	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000BGT	ENCFF000NSJ	BATF	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BGT	ENCFF000NSN	BATF	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BHA	ENCFF000NSQ	BCL11A	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BHA	ENCFF000NSU	BCL11A	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BNQ	ENCFF000NSY	BCL3	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000BNQ	ENCFF000NTA	BCL3	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000BJZ	ENCFF000NTF	BCLAF1	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000BJZ	ENCFF000NTG	BCLAF1	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000DZJ	ENCFF000VSB	BHLHE40	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZJ	ENCFF000VSI	BHLHE40	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZS	ENCFF000VSJ	BRCA1	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZS	ENCFF000VSK	BRCA1	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000BRX	ENCFF000NTO	CEBPB	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000BRX	ENCFF000NTR	CEBPB	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000DZE	ENCFF000VTE	CHD1	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZE	ENCFF000VTQ	CHD1	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZR	ENCFF000VTR	CHD2	/experiments/ENCSR000EAF/	GM12878	hg19

experiment_accession	file_accession	target	controls	biosample_name	assembly
ENCSR000DZR	ENCFF000VTS	CHD2	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000AKJ	ENCFF000AQZ	Control	NA	GM12878	hg19
ENCSR000AKJ	ENCFF000ARA	Control	NA	GM12878	hg19
ENCSR000BGH	ENCFF000OCL	Control	NA	GM12878	hg19
ENCSR000BGH	ENCFF000OCM	Control	NA	GM12878	hg19
ENCSR000BGH	ENCFF000OCN	Control	NA	GM12878	hg19
ENCSR000BGH	ENCFF000OCQ	Control	NA	GM12878	hg19
ENCSR000BGJ	ENCFF000ODC	Control	NA	GM12878	hg19
ENCSR000BGJ	ENCFF000ODD	Control	NA	GM12878	hg19
ENCSR000BMS	ENCFF000ODH	Control	NA	GM12878	hg19
ENCSR000BNF	ENCFF000ODL	Control	NA	GM12878	hg19
ENCSR000BVP	ENCFF000ODM	Control	NA	GM12878	hg19
ENCSR000BVP	ENCFF000ODN	Control	NA	GM12878	hg19
ENCSR000BVP	ENCFF000ODR	Control	NA	GM12878	hg19
ENCSR000DKW	ENCFF000ROW	Control	NA	GM12878	hg19
ENCSR000EYX	ENCFF000VWO	Control	NA	GM12878	hg19
ENCSR000DRV	ENCFF001HHW	Control	NA	GM12878	hg19
ENCSR000BUF	ENCFF000NTW	CREB1	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000BUF	ENCFF000NTZ	CREB1	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000AKB	ENCFF000ARG	CTCF	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AKB	ENCFF000ARI	CTCF	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000DKV	ENCFF000ROM	CTCF	/experiments/ENCSR000DKW/	GM12878	hg19
ENCSR000DKV	ENCFF000RON	CTCF	/experiments/ENCSR000DKW/	GM12878	hg19
ENCSR000DKV	ENCFF000ROQ	CTCF	/experiments/ENCSR000DKW/	GM12878	hg19
ENCSR000DZN	ENCFF000VUC	CTCF	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000DZN	ENCFF000VUE	CTCF	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000DRZ	ENCFF001HHN	CTCF	/experiments/ENCSR000DRV/	GM12878	hg19
ENCSR000DRZ	ENCFF001HHS	CTCF	/experiments/ENCSR000DRV/	GM12878	hg19
ENCSR000DYR	ENCFF000VSQ	CUX1	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DYR	ENCFF000VSU	CUX1	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DYY	ENCFF000VUG	E2F4	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DYY	ENCFF000VUH	E2F4	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000BGU	ENCFF000NUE	EBF1	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BGU	ENCFF000NUG	EBF1	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000DZQ	ENCFF000VUQ	EBF1	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000DZQ	ENCFF000VUS	EBF1	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000BRG	ENCFF000NUO	EGR1	/experiments/ENCSR000BGJ/	GM12878	hg19
ENCSR000BMQ	ENCFF000NUS	EGR1	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000BMQ	ENCFF000NUT	EGR1	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000BMB	ENCFF000NUX	ELF1	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000BMB	ENCFF000NUY	ELF1	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000DZB	ENCFF000VUV	ELK1	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZB	ENCFF000VVD	ELK1	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000BHB	ENCFF000NYS	EP300	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BHB	ENCFF000NYW	EP300	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000DZD	ENCFF000WAJ	EP300	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZD	ENCFF000WAR	EP300	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZG	ENCFF000WAS	EP300	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000DZG	ENCFF000WAT	EP300	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000DYQ	ENCFF000VVE	ESRRA	/experiments/ENCSR000EAB/	GM12878	hg19
ENCSR000DYQ	ENCFF000VVL	ESRRA	/experiments/ENCSR000EAB/	GM12878	hg19
ENCSR000BKA	ENCFF000NVG	ETS1	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000BKA	ENCFF000NVL	ETS1	/experiments/ENCSR000BMS/	GM12878	hg19

experiment_accession	file_accession	target	controls	biosample_name	assembly
ENCSR000ARD	ENCFF000ARL	EZH2	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000ARD	ENCFF000ARN	EZH2	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000EYZ	ENCFF000VSZ	FOS	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000EYZ	ENCFF000VTA	FOS	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000EYZ	ENCFF000VTB	FOS	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000BRU	ENCFF000NVR	FOXM1	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000BRU	ENCFF000NVU	FOXM1	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000BGC	ENCFF000NVY	GABPA	/experiments/ENCSR000BGJ/	GM12878	hg19
ENCSR000BGC	ENCFF000NWB	GABPA	/experiments/ENCSR000BGJ/	GM12878	hg19
ENCSR000AOV	ENCFF000ARR	H2AFZ	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AOV	ENCFF000ARS	H2AFZ	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AKC	ENCFF000ASG	H3K27ac	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AKC	ENCFF000ASI	H3K27ac	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AKD	ENCFF000ASK	H3K27me3	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AKD	ENCFF000ASL	H3K27me3	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AKD	ENCFF000ASQ	H3K27me3	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000DRX	ENCFF001EWW	H3K27me3	/experiments/ENCSR000DRV/	GM12878	hg19
ENCSR000DRX	ENCFF001EWZ	H3K27me3	/experiments/ENCSR000DRV/	GM12878	hg19
ENCSR000AKE	ENCFF000ASX	H3K36me3	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AKE	ENCFF000ATB	H3K36me3	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000DRW	ENCFF001EXF	H3K36me3	/experiments/ENCSR000DRV/	GM12878	hg19
ENCSR000DRW	ENCFF001EXG	H3K36me3	/experiments/ENCSR000DRV/	GM12878	hg19
ENCSR000AKF	ENCFF000ARY	H3K4me1	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AKF	ENCFF000ATE	H3K4me1	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AKG	ENCFF000ATG	H3K4me2	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AKG	ENCFF000ATH	H3K4me2	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AKA	ENCFF000ASE	H3K4me3	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AKA	ENCFF000ATS	H3K4me3	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000DRY	ENCFF001EXQ	H3K4me3	/experiments/ENCSR000DRV/	GM12878	hg19
ENCSR000DRY	ENCFF001EXR	H3K4me3	/experiments/ENCSR000DRV/	GM12878	hg19
ENCSR000AOW	ENCFF000ATT	H3K79me2	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AOW	ENCFF000ATV	H3K79me2	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AKH	ENCFF000ATY	H3K9ac	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AKH	ENCFF000AUA	H3K9ac	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AOX	ENCFF000AUG	H3K9me3	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AOX	ENCFF000AUH	H3K9me3	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AOX	ENCFF000AUJ	H3K9me3	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AKI	ENCFF000AUR	H4K20me1	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000AKI	ENCFF000AUT	H4K20me1	/experiments/ENCSR000AKJ/	GM12878	hg19
ENCSR000EUJ	ENCFF000VVS	IKZF1	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000EUJ	ENCFF000VVT	IKZF1	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000DZX	ENCFF000VWS	IRF3	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZX	ENCFF000VWU	IRF3	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000BGY	ENCFF000NWF	IRF4	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BGY	ENCFF000NWK	IRF4	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000DYS	ENCFF000VWX	JUND	/experiments/ENCSR000EAB/	GM12878	hg19
ENCSR000DYS	ENCFF000VWY	JUND	/experiments/ENCSR000EAB/	GM12878	hg19
ENCSR000EYV	ENCFF000VXL	JUND	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000EYV	ENCFF000VXM	JUND	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000DNO	ENCFF000VVM	KAT2A	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000DNO	ENCFF000VVO	KAT2A	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000DYV	ENCFF000VXN	MAFK	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DYV	ENCFF000VXO	MAFK	/experiments/ENCSR000EAF/	GM12878	hg19

experiment_accession	file_accession	target	controls	biosample_name	assembly
ENCSR000DZF	ENCFF000VXV	MAX	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZF	ENCFF000VXW	MAX	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZA	ENCFF000VYJ	MAZ	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZA	ENCFF000VYK	MAZ	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000BKB	ENCFF000NWO	MEF2A	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000BKB	ENCFF000NWR	MEF2A	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000BNG	ENCFF000NWW	MEF2C	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000BNG	ENCFF000NWZ	MEF2C	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000EAF	ENCFF000VWA	mouse-IgG-control	NA	GM12878	hg19
ENCSR000EAF	ENCFF000VWB	mouse-IgG-control	NA	GM12878	hg19
ENCSR000BRH	ENCFF000NXH	MTA3	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000BRH	ENCFF000NXI	MTA3	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000DZI	ENCFF000VYR	MXI1	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZI	ENCFF000VYY	MXI1	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DKU	ENCFF000ROE	MYC	/experiments/ENCSR000DKW/	GM12878	hg19
ENCSR000DKU	ENCFF000ROF	MYC	/experiments/ENCSR000DKW/	GM12878	hg19
ENCSR000BQL	ENCFF000NXM	NFATC1	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000BQL	ENCFF000NXP	NFATC1	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000DZY	ENCFF000VYZ	NFE2	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000DZY	ENCFF000VZB	NFE2	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000BRN	ENCFF000NXX	NFIC	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000BRN	ENCFF000NXY	NFIC	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000DNN	ENCFF000VZO	NFYA	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DNN	ENCFF000VZP	NFYA	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DNM	ENCFF000VZX	NFYB	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DNM	ENCFF000WAC	NFYB	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000EUL	ENCFF000WFL	NR2C2	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000EUL	ENCFF000WFM	NR2C2	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000DZO	ENCFF000WAF	NRF1	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZO	ENCFF000WAH	NRF1	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000BHD	ENCFF000NYZ	PAX5	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BHD	ENCFF000NZF	PAX5	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BHJ	ENCFF000NZI	PAX5	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BHJ	ENCFF000NZL	PAX5	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BGR	ENCFF000NZP	PBX3	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BGR	ENCFF000NZR	PBX3	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BQM	ENCFF000NZY	PML	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000BQM	ENCFF000OAB	PML	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000BGD	ENCFF000OAO	POLR2A	/experiments/ENCSR000BGJ/	GM12878	hg19
ENCSR000BGD	ENCFF000OAR	POLR2A	/experiments/ENCSR000BGJ/	GM12878	hg19
ENCSR000DKT	ENCFF000RPA	POLR2A	/experiments/ENCSR000DKW/	GM12878	hg19
ENCSR000DKT	ENCFF000RPC	POLR2A	/experiments/ENCSR000DKW/	GM12878	hg19
ENCSR000EAD	ENCFF000WBH	POLR2A	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000EAD	ENCFF000WBI	POLR2A	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000EAD	ENCFF000WBJ	POLR2A	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000EAD	ENCFF000WBK	POLR2A	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000EAD	ENCFF000WBL	POLR2A	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000EAD	ENCFF000WBN	POLR2A	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000EAD	ENCFF000WBO	POLR2A	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000EYW	ENCFF000WBY	POLR2A	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000EYW	ENCFF000WCB	POLR2A	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000DZK	ENCFF000WCD	POLR2AphosphoS2	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZK	ENCFF000WCF	POLR2AphosphoS2	/experiments/ENCSR000EAF/	GM12878	hg19

experiment_accession	file_accession	target	controls	biosample_name	assembly
ENCSR000BIF	ENCFF000OAG	POLR2AphosphoS5	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BIF	ENCFF000OAJ	POLR2AphosphoS5	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000EYU	ENCFF000WCJ	POLR3G	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000EYU	ENCFF000WCK	POLR3G	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000BGP	ENCFF000OAW	POU2F2	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BGP	ENCFF000OAX	POU2F2	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BGP	ENCFF000OBC	POU2F2	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000EAB	ENCFF000VWE	rabbit-IgG-control	NA	GM12878	hg19
ENCSR000DZH	ENCFF000VWP	rabbit-IgG-control	NA	GM12878	hg19
ENCSR000DZH	ENCFF000VWQ	rabbit-IgG-control	NA	GM12878	hg19
ENCSR000BMY	ENCFF000OBV	RAD21	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000BMY	ENCFF000OCB	RAD21	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000EAC	ENCFF000WCN	RAD21	/experiments/ENCSR000EAB/	GM12878	hg19
ENCSR000EAC	ENCFF000WCQ	RAD21	/experiments/ENCSR000EAB/	GM12878	hg19
ENCSR000DZC	ENCFF000VTU	RCOR1	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZC	ENCFF000VUB	RCOR1	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000EAG	ENCFF000VZF	RELA	/experiments/ENCSR000DZH/	GM12878	hg19
ENCSR000EAG	ENCFF000VZG	RELA	/experiments/ENCSR000DZH/	GM12878	hg19
ENCSR000EAG	ENCFF000VZI	RELA	/experiments/ENCSR000DZH/	GM12878	hg19
ENCSR000EAG	ENCFF000VZJ	RELA	/experiments/ENCSR000DZH/	GM12878	hg19
ENCSR000BQS	ENCFF000NYC	REST	/experiments/ENCSR000BNF/	GM12878	hg19
ENCSR000BQS	ENCFF000NYF	REST	/experiments/ENCSR000BNF/	GM12878	hg19
ENCSR000BGF	ENCFF000NYO	REST	/experiments/ENCSR000BGJ/	GM12878	hg19
ENCSR000BGF	ENCFF000NYP	REST	/experiments/ENCSR000BGJ/	GM12878	hg19
ENCSR000DZW	ENCFF000WCV	RFX5	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZW	ENCFF000WCW	RFX5	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000BRI	ENCFF000OCD	RUNX3	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000BRI	ENCFF000OCK	RUNX3	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000BJD	ENCFF000ODU	RXRA	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BJD	ENCFF000ODW	RXRA	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000DYX	ENCFF000WCZ	SIN3A	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DYX	ENCFF000WDA	SIN3A	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000BJE	ENCFF000OED	SIX5	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BJE	ENCFF000OEF	SIX5	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000DZP	ENCFF000WDH	SMC3	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZP	ENCFF000WDN	SMC3	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000BHK	ENCFF000OEL	SP1	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BHK	ENCFF000OEO	SP1	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BGQ	ENCFF000OBI	SPI1	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BGQ	ENCFF000OBJ	SPI1	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BGQ	ENCFF000OBO	SPI1	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000DYU	ENCFF000WDT	SREBF1	/experiments/ENCSR000EAB/	GM12878	hg19
ENCSR000DYU	ENCFF000WDU	SREBF1	/experiments/ENCSR000EAB/	GM12878	hg19
ENCSR000DYT	ENCFF000WED	SREBF2	/experiments/ENCSR000EAB/	GM12878	hg19
ENCSR000DYT	ENCFF000WEK	SREBF2	/experiments/ENCSR000EAB/	GM12878	hg19
ENCSR000BGE	ENCFF000OEY	SRF	/experiments/ENCSR000BGJ/	GM12878	hg19
ENCSR000BGE	ENCFF000OEZ	SRF	/experiments/ENCSR000BGJ/	GM12878	hg19
ENCSR000BMI	ENCFF000OFD	SRF	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000BMI	ENCFF000OFG	SRF	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000DZM	ENCFF000WEL	STAT1	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000DZM	ENCFF000WEN	STAT1	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000DZV	ENCFF000WER	STAT3	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZV	ENCFF000WES	STAT3	/experiments/ENCSR000EAF/	GM12878	hg19

experiment_accession	file_accession	target	controls	biosample_name	assembly
ENCSR000BQZ	ENCFF000OFP	STAT5A	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000BQZ	ENCFF000OFQ	STAT5A	/experiments/ENCSR000BVP/	GM12878	hg19
ENCSR000DNP	ENCFF000WDP	SUPT20H	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000DNP	ENCFF000WDR	SUPT20H	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000BGS	ENCFF000OFT	TAF1	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BGS	ENCFF000OFW	TAF1	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000DYZ	ENCFF000WEV	TBL1XR1	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DYZ	ENCFF000WEW	TBL1XR1	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZZ	ENCFF000WFF	TBP	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZZ	ENCFF000WFG	TBP	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000BGZ	ENCFF000OGB	TCF12	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BGZ	ENCFF000OGF	TCF12	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BQT	ENCFF000OGJ	TCF3	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BQT	ENCFF000OGM	TCF3	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BGI	ENCFF000OGR	USF1	/experiments/ENCSR000BGJ/	GM12878	hg19
ENCSR000BGI	ENCFF000OGU	USF1	/experiments/ENCSR000BGJ/	GM12878	hg19
ENCSR000DZU	ENCFF000WFT	USF2	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DZU	ENCFF000WFU	USF2	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000EAA	ENCFF000WFW	WRNIP1	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000EAA	ENCFF000WFY	WRNIP1	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000BNP	ENCFF000OGZ	YY1	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000BNP	ENCFF000OHC	YY1	/experiments/ENCSR000BMS/	GM12878	hg19
ENCSR000EUM	ENCFF000WGH	YY1	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000EUM	ENCFF000WGI	YY1	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000BHC	ENCFF000OHG	ZBTB33	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BHC	ENCFF000OHK	ZBTB33	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BND	ENCFF000OHP	ZEB1	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000BND	ENCFF000OHS	ZEB1	/experiments/ENCSR000BGH/	GM12878	hg19
ENCSR000DZL	ENCFF000WGJ	ZNF143	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000DZL	ENCFF000WGL	ZNF143	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000DYW	ENCFF000WGO	ZNF274	/experiments/ENCSR000EYX/	GM12878	NA
ENCSR000DYW	ENCFF000WGQ	ZNF274	/experiments/ENCSR000EYX/	GM12878	NA
ENCSR000DYP	ENCFF000WHE	ZNF384	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DYP	ENCFF000WHF	ZNF384	/experiments/ENCSR000EAF/	GM12878	hg19
ENCSR000DNQ	ENCFF000WHH	ZZZ3	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000DNQ	ENCFF000WHK	ZZZ3	/experiments/ENCSR000EYX/	GM12878	hg19
ENCSR000EJD	ENCFF000SKV	NA	NA	GM12878	hg19
ENCSR000EJD	ENCFF000SKW	NA	NA	GM12878	hg19
ENCSR000EJD	ENCFF000SKZ	NA	NA	GM12878	hg19
ENCSR000EJD	ENCFF000SLB	NA	NA	GM12878	hg19
ENCSR000EJD	ENCFF000SLD	NA	NA	GM12878	hg19
ENCSR000EMT	ENCFF001CTZ	NA	NA	GM12878	hg19
ENCSR000EMT	ENCFF001CUI	NA	NA	GM12878	hg19

TABLEAU B.2 – Description of similaRpeak's pseudometrics. (continued below)

Pseudometric	Definition	Threshold
RATIO_AREA	The ratio between the profile areas.	1- minimal area threshold
DIFF_POS_MAX	The difference between the maximal peaks positions. When a profile has more than one position with the maximum value, the median position is used. A maximal distance threshold ensure that the pseudometric is not calculated in presence of multiple peaks on the same profile. A tolerance threshold unable the inclusion of all positions within a certain range of the maximum value.	1- minimum peak value threshold 2- maximal distance threshold between two maximum positions 3- tolerance threshold
RATIO_MAX_MAX	The ratio between the peaks values in each profile.	1- minimal peak value threshold
RATIO_INTERSECT	The ratio between the intersection area and the total area of the two profiles. For the divisor, the intersection area is only counted once in the total area.	1 - minimal area threshold
RATIO_NORMALIZED_INTERSECT	The ratio between the intersection area and the total area of two normalized profiles. The profiles are normalized by dividing them by their average value.	1- minimal area threshold
PEARMAN_CORRELATION	The Spearman's rho statistic between profiles.	NA

TABLEAU B.4 – Classification of GM12878 factors.

target	type	RATIO_INTERSECT	class
ATF2	enhancers	0.7496309	gradient
ATF2	promoters	0.8361148	gradient
ATF3	enhancers	0.8737663	threshold
ATF3	promoters	0.8226083	gradient
BATF	enhancers	0.8953338	threshold
BATF	promoters	0.8467073	gradient
BCL11A	enhancers	0.8588480	threshold
BCL11A	promoters	0.8383785	gradient
BCL3	enhancers	0.8164351	gradient
BCL3	promoters	0.8408951	gradient
BCLAF1	enhancers	0.7812544	gradient
BCLAF1	promoters	0.8000127	gradient
BHLHE40	enhancers	0.8494647	gradient
BHLHE40	promoters	0.7713868	gradient
BRCA1	enhancers	0.7669265	gradient
BRCA1	promoters	0.7697468	gradient
CEBPB	enhancers	0.7812174	gradient
CEBPB	promoters	0.8811076	threshold
CEBPZ	enhancers	0.8010971	gradient
CEBPZ	promoters	0.7881752	gradient
CHD1	enhancers	0.8024852	gradient
CHD1	promoters	0.7917686	gradient
CHD2	enhancers	0.7149570	gradient
CHD2	promoters	0.6873916	gradient
CREB1	enhancers	0.8174688	gradient
CREB1	promoters	0.6057792	gradient
CREM	enhancers	0.8579512	threshold
CREM	promoters	0.7082010	gradient
CTCF	enhancers	0.6950783	gradient
CTCF	promoters	0.8780933	threshold
CUX1	enhancers	0.8080430	gradient
CUX1	promoters	0.8372762	gradient
DNase	enhancers	0.8590574	threshold
DNase	promoters	0.7729776	gradient
E2F4	enhancers	0.8630400	threshold
E2F4	promoters	0.6692210	gradient
EBF1	enhancers	0.7373044	gradient
EBF1	promoters	0.8201636	gradient
EED	enhancers	0.9493303	threshold
EED	promoters	0.8789210	threshold
EGR1	enhancers	0.8890021	threshold
EGR1	promoters	0.7729942	gradient
ELF1	enhancers	0.8824260	threshold
ELF1	promoters	0.6635204	gradient
ELK1	enhancers	0.8583322	threshold
ELK1	promoters	0.7656027	gradient
EP300	enhancers	0.8562707	threshold
EP300	promoters	0.8332010	gradient
ESRRA	enhancers	0.8704275	threshold
ESRRA	promoters	0.8690370	threshold
ETS1	enhancers	0.7735212	gradient

target	type	RATIO_INTERSECT	class
ETS1	promoters	0.7609079	gradient
ETV6	enhancers	0.9300437	threshold
ETV6	promoters	0.8097995	gradient
EZH2	enhancers	0.9219432	threshold
EZH2	promoters	0.9039275	threshold
FOS	enhancers	0.7135716	gradient
FOS	promoters	0.6510312	gradient
FOXM1	enhancers	0.7780945	gradient
FOXM1	promoters	0.8362155	gradient
GABPA	enhancers	0.8442373	gradient
GABPA	promoters	0.4964659	gradient
H3K27ac	enhancers	0.7728215	gradient
H3K27ac	promoters	0.6597663	gradient
H3K27me3	enhancers	0.8436130	gradient
H3K27me3	promoters	0.5230462	gradient
H3K36me3	enhancers	0.8795926	threshold
H3K36me3	promoters	0.8164637	gradient
H3K4me1	enhancers	0.9125790	threshold
H3K4me1	promoters	0.7984258	gradient
H3K4me2	enhancers	0.9171471	threshold
H3K4me2	promoters	0.8460399	gradient
H3K4me3	enhancers	0.7924974	gradient
H3K4me3	promoters	0.6421751	gradient
H3K79me2	enhancers	0.9218091	threshold
H3K79me2	promoters	0.5131972	gradient
H3K9ac	enhancers	0.7521190	gradient
H3K9ac	promoters	0.6442444	gradient
H3K9me3	enhancers	0.9065338	threshold
H3K9me3	promoters	0.8638066	threshold
H4K20me1	enhancers	0.7979631	gradient
H4K20me1	promoters	0.9023709	threshold
IKZF1	enhancers	0.9397585	threshold
IKZF1	promoters	0.8856953	threshold
IRF3	enhancers	0.8891108	threshold
IRF3	promoters	0.7998822	gradient
IRF4	enhancers	0.9312747	threshold
IRF4	promoters	0.8202070	gradient
JUND	enhancers	0.8999446	threshold
JUND	promoters	0.8396737	gradient
KAT2A	enhancers	0.8417074	gradient
KAT2A	promoters	0.8551702	threshold
MAFK	enhancers	0.9003880	threshold
MAFK	promoters	0.8725800	threshold
MAX	enhancers	0.8574277	threshold
MAX	promoters	0.7941603	gradient
MAZ	enhancers	0.8612428	threshold
MAZ	promoters	0.8201461	gradient
MEF2A	enhancers	0.8033372	gradient
MEF2A	promoters	0.7926042	gradient
MEF2C	enhancers	0.8207161	gradient
MEF2C	promoters	0.8314029	gradient
MTA3	enhancers	0.6775226	gradient
MTA3	promoters	0.8411555	gradient

target	type	RATIO_INTERSECT	class
MXI1	enhancers	0.8183073	gradient
MXI1	promoters	0.7168870	gradient
MYC	enhancers	0.8807865	threshold
MYC	promoters	0.8042135	gradient
NFATC1	enhancers	0.6603011	gradient
NFATC1	promoters	0.8769123	threshold
NFE2	enhancers	0.8399090	gradient
NFE2	promoters	0.7465599	gradient
NFIC	enhancers	0.8033133	gradient
NFIC	promoters	0.8310981	gradient
NFYA	enhancers	0.8267284	gradient
NFYA	promoters	0.6990368	gradient
NFYB	enhancers	0.6817663	gradient
NFYB	promoters	0.5915779	gradient
NR2C2	enhancers	0.9167941	threshold
NR2C2	promoters	0.7836656	gradient
NRF1	enhancers	0.8639965	threshold
NRF1	promoters	0.7017489	gradient
PAX5	enhancers	0.8184942	gradient
PAX5	promoters	0.8226313	gradient
PBX3	enhancers	0.7140275	gradient
PBX3	promoters	0.6960342	gradient
PML	enhancers	0.5465648	gradient
PML	promoters	0.6228788	gradient
POLR2AphosphoS2	enhancers	0.6586316	gradient
POLR2AphosphoS2	promoters	0.6818314	gradient
POLR2AphosphoS5	enhancers	0.5787311	gradient
POLR2AphosphoS5	promoters	0.5954451	gradient
POLR2A	enhancers	0.5686428	gradient
POLR2A	promoters	0.6039013	gradient
POLR3G	enhancers	0.8269562	gradient
POLR3G	promoters	0.9012033	threshold
POU2F2	enhancers	0.6915335	gradient
POU2F2	promoters	0.7209998	gradient
RAD21	enhancers	0.8929759	threshold
RAD21	promoters	0.8948873	threshold
RCOR1	enhancers	0.8815286	threshold
RCOR1	promoters	0.8444167	gradient
RELA	enhancers	0.7918326	gradient
RELA	promoters	0.7831202	gradient
REST	enhancers	0.8010338	gradient
REST	promoters	0.8558240	threshold
RFX5	enhancers	0.7844170	gradient
RFX5	promoters	0.7436267	gradient
RUNX3	enhancers	0.8862939	threshold
RUNX3	promoters	0.7728822	gradient
RXRA	enhancers	0.8270339	gradient
RXRA	promoters	0.8359216	gradient
SIN3A	enhancers	0.8813798	threshold
SIN3A	promoters	0.7335123	gradient
SIX5	enhancers	0.7425485	gradient
SIX5	promoters	0.6891345	gradient
SMAD5	enhancers	0.6305831	gradient

target	type	RATIO_INTERSECT	class
SMAD5	promoters	0.6729919	gradient
SMC3	enhancers	0.9166033	threshold
SMC3	promoters	0.8809240	threshold
SP1	enhancers	0.7982942	gradient
SP1	promoters	0.6772314	gradient
SPI1	enhancers	0.8242364	gradient
SPI1	promoters	0.7869797	gradient
SREBF1	enhancers	0.8509871	threshold
SREBF1	promoters	0.8359284	gradient
SREBF2	enhancers	0.7400525	gradient
SREBF2	promoters	0.8418863	gradient
SRF	enhancers	0.6992078	gradient
SRF	promoters	0.6785688	gradient
STAT1	enhancers	0.8948733	threshold
STAT1	promoters	0.8858612	threshold
STAT3	enhancers	0.8730606	threshold
STAT3	promoters	0.8027093	gradient
STAT5A	enhancers	0.7257072	gradient
STAT5A	promoters	0.8933277	threshold
SUPT20H	enhancers	0.8936713	threshold
SUPT20H	promoters	0.9484882	threshold
TAF1	enhancers	0.7125450	gradient
TAF1	promoters	0.6376661	gradient
TBL1XR1	enhancers	0.9039882	threshold
TBL1XR1	promoters	0.8197264	gradient
TBP	enhancers	0.6004067	gradient
TBP	promoters	0.6182798	gradient
TCF12	enhancers	0.7941260	gradient
TCF12	promoters	0.8151668	gradient
TCF3	enhancers	0.8079836	gradient
TCF3	promoters	0.8102212	gradient
TCF7	enhancers	0.8511006	threshold
TCF7	promoters	0.7975571	gradient
USF1	enhancers	0.7876330	gradient
USF1	promoters	0.6810510	gradient
USF2	enhancers	0.7983729	gradient
USF2	promoters	0.7544840	gradient
WRNIP1	enhancers	0.9091523	threshold
WRNIP1	promoters	0.7635192	gradient
YY1	enhancers	0.3440772	gradient
YY1	promoters	0.6711835	gradient
ZBED1	enhancers	0.7079517	gradient
ZBED1	promoters	0.7263530	gradient
ZBTB33	enhancers	0.6649304	gradient
ZBTB33	promoters	0.6948484	gradient
ZEB1	enhancers	0.8879040	threshold
ZEB1	promoters	0.8747564	threshold
ZNF143	enhancers	0.5934893	gradient
ZNF143	promoters	0.6777972	gradient
ZNF274	enhancers	0.8942597	threshold
ZNF274	promoters	0.8655489	threshold
ZNF384	enhancers	0.8630051	threshold
ZNF384	promoters	0.7625790	gradient

target	type	RATIO_INTERSECT	class
ZZZ3	enhancers	0.8544116	threshold
ZZZ3	promoters	0.8129802	gradient

S1 Text. *metagene* profiles analyses reveal regulatory elements specific recruitment pattern

Charles Joly Beauparlant[1,2,5], Fabien C. Lamaze[1,3,5], Astrid Deschênes[1], Rawane Samb[1], Audrey Lemaçon[1], Pascal Belleau[1], Steve Bilodeau[1,3,4] and Arnaud Droit[1,2,*]

[1]Centre de Recherche du CHU de Québec, Université Laval, Québec, Québec, G1V 0A6
[2]Département de Médecine Moléculaire, Faculté de médecine, Québec, Canada
[3]Centre de Recherche sur le Cancer de l'Université Laval, 9, rue McMahon, Québec, Québec, G1R 3S3
[4]Département de Biologie Moléculaire, Biochimie Médicale et Pathologie, Faculté de médecine, Québec, canada
[5]Co-first authors.

Data collection

First, we stratified four groups of enhancer and promoter regions based on their transcriptional levels using cap analysis of gene expression (CAGE) levels [28]. Transcription start sites (TSS) with a transcript per million (TPM) value greater than 0 were extracted from Fantom5 for the GM12878 and split into low expression (smaller or equal to the 33th percentile), moderate expression (between the 33th and 66th percentile) and high expression (greater than 66th percentile) groups. Regions with a TPM value of 0 were included into the "no expression" group. The Fantom consortium defines enhancers as regions with balanced bidirectional capped transcripts [28], as measured by CAGE. Fantom also defines a subset of TSS as 'robust' using a tag evidence thresholds approach [38]. Each enhancer region contains between 0 and 20 robust TSS. We studied enhancers from the Fantom 5 release that overlap at least one 'robust TSS' and we defined the expression level or the enhancer as the maximal TPM value of the overlapped TSS. Enhancers were resized to a final width of 2000 nucleotides to produce the robust enhancers regions. The promoters were defined using the Bioconductor's TxDb.Hsapiens.UCSC.hg19.knownGene package [16] with 1500 nucleotides upstream and 500 nucleotides downstream of each Entrez gene TSS position and were filtered to include only the regions overlapping at least one robust TSS. Their expression level was defined as for the enhancers. Each region was then centered on the Entrez gene TSS position. The files were downloaded with the ENCODExplorer package [39].

Bootstrap

The metagene package produces one curve for each combination of group of genomic regions and sample. To calculate the value of the points in the curve, metagene calculates the estimator (mean or median) value of each row in a matrix where columns correspond to a bin and each row to a genomic region. The bootstrapping step is performed to calculate the confidence interval (CI) on the estimator. The CI of each curve is calculated independently. For each bin position, metagene will sample with replacement the values. By default, metagene will produce 1000 samples (can be changed with the sample_count parameter of the produce_data_frame function) of the size of the smallest number of rows of all the matrices (can be changed with the sample_size parameter of the produce_data_frame function). The estimator value is computed for each sample and the result is used to calculate the 0.025 and 0.975 CI (alpha = 0.05).

Permutation

It is possible to compare two profiles using a metric of interest (see Metrics description section of this document for a list of recommended metrics implemented in the similaRpeak package). The metagene package implements a permutation strategy to determine if two profiles are

statistically different for a given metric. The permutation analysis will perform multiple iterations (the number of permutation can be defined with the sample_count parameter). For each iteration, two new profiles are produced by randomly sampling elements for the combination of the two matrices used to produce the original profiles. The number of element sampled for each profile is defined with the sample_size parameter. The metric value is then calculated using the new profiles. The ratio of metrics values with a score greater or equal to the original metric value can be used to compute a p-value and determine if the two original profiles are statistically different.

Supplementary Bibliography:

[38] Consortium TF, et al. A promoter-level mammalian expression atlas. Nature. 2014;507(7493):462–470.
[39] Beauparlant CJ, Lemacon A, Droit A. ENCODExplorer: A compilation of ENCODE metadata; 2015. R package version 1.2.1.

Bibliographie

, G. S. (2013). Integrative genomic analysis by interoperation of bioinformatics tools in GenomeSpace. `http://www.genomespace.org`. (18 June 2013, date last accessed).

Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J.-J., Mayer, P., and Kawashima, E. (2000). Solid phase dna amplification : characterisation of primer attachment and amplification mechanisms. *Nucleic acids research*, 28(20) :e87–e87.

Adli, M. and Bernstein, B. E. (2011). Whole-genome chromatin profiling from limited numbers of cells using nano-chip-seq. *Nature protocols*, 6(10) :1656–1668.

Albert, I., Wachi, S., Jiang, C., and Pugh, B. F. (2008). Genetrack—a genomic data processing and visualization framework. *Bioinformatics*, 24(10) :1305–1306.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast : a new generation of protein database search programs. *Nucleic acids research*, 25(17) :3389–3402.

Ambrosini, G., Dreos, R., Kumar, S., and Bucher, P. (2016). The chip-seq tools and web server : a resource for analyzing chip-seq and other types of genomic data. *BMC genomics*, 17(1) :938.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11(10) :1.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493) :455–461.

Andersson, R., Sandelin, A., and Danko, C. G. (2015). A unified architecture of transcriptional regulatory elements. *Trends in Genetics*, 31(8) :426–433.

Angelini, C., Heller, R., Volkinshtein, R., and Yekutieli, D. (2015). Is this the right normalization ? a diagnostic tool for chip-seq normalization. *BMC bioinformatics*, 16(1) :1.

Anonymous (2014). Credit for code.

Arner, E., Daub, C. O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drabløs, F., Lennartsson, A., Rönnerblad, M., Hrydziuszko, O., Vitezic, M., et al. (2015). Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, 347(6225) :1010–1014.

Arnosti, D. N. and Kulkarni, M. M. (2005). Transcriptional enhancers : Intelligent enhanceosomes or flexible billboards ? *Journal of cellular biochemistry*, 94(5) :890–898.

Ashoor, H., Hérault, A., Kamoun, A., Radvanyi, F., Bajic, V. B., Barillot, E., and Boeva, V. (2013). Hmcan : a method for detecting chromatin modifications in cancer samples using chip-seq data. *Bioinformatics*, 29(23) :2979–2986.

Aszódi, A. (2012). Multovl : fast multiple overlaps of genomic regions. *Bioinformatics*, 28(24) :3318–3319.

Auerbach, R. K., Euskirchen, G., Rozowsky, J., Lamarre-Vincent, N., Moqtaderi, Z., Lefrançois, P., Struhl, K., Gerstein, M., and Snyder, M. (2009). Mapping accessible chromatin regions using sono-seq. *Proceedings of the National Academy of Sciences*, 106(35) :14926–14931.

Bailey, T., Krajewski, P., Ladunga, I., and Lefebvre, C. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput . . .*.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). Meme suite : tools for motif discovery and searching. *Nucleic acids research*, page gkp335.

Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). Meme : discovering and analyzing dna and protein sequence motifs. *Nucleic acids research*, 34(suppl 2) :W369–W373.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote sv40 dna sequences. *Cell*, 27(2) :299–308.

Bardet, A. F., Steinmann, J., Bafna, S., Knoblich, J. A., Zeitlinger, J., and Stark, A. (2013). Identification of transcription factor binding sites from chip-seq data at high resolution. *Bioinformatics*, page btt470.

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P., and Marth, G. T. (2011). Bamtools : a c++ api and toolkit for analyzing and managing bam files. *Bioinformatics*, 27(12) :1691–1692.

Barolo, S. and Posakony, J. W. (2002). Three habits of highly effective signaling pathways : principles of transcriptional control by developmental cell signaling. *Genes & development*, 16(10) :1167–1181.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4) :823–837.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *nature*, 456(7218) :53–59.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., et al. (2010). The nih roadmap epigenomics mapping consortium. *Nature biotechnology*, 28(10) :1045–1048.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146) :799–816.

Blahnik, K. R., Dou, L., O'Geen, H., McPhillips, T., Xu, X., Cao, A. R., Iyengar, S., Nicolet, C. M., Ludäscher, B., Korf, I., et al. (2010). Sole-search : an integrated analysis program for peak detection and functional annotation using chip-seq data. *Nucleic acids research*, 38(3) :e13–e13.

Boeva, V., Surdez, D., Guillon, N., Tirode, F., Fejes, A. P., Delattre, O., and Barillot, E. (2010). De novo motif identification improves the accuracy of predicting transcription factor binding sites in chip-seq data analysis. *Nucleic acids research*, 38(11) :e126–e126.

Boyle, A. P., Guinney, J., Crawford, G. E., and Furey, T. S. (2008). F-seq : a feature density estimator for high-throughput sequence tags. *Bioinformatics*, 24(21) :2537–2538.

Brenner, C. (2012). HOMER : Software for motif discovery and next-gen sequencing analysis. `http://www.genomespace.org`. (21 December 2016, date last accessed).

Calo, E. and Wysocka, J. (2013). Modification of enhancer chromatin : what, how, and why ? *Molecular cell*, 49(5) :825–837.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., et al. (2005). The transcriptional landscape of the mammalian genome. *Science*, 309(5740) :1559–1563.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A., Taylor, M. S., Engström, P. G., Frith, M. C., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature genetics*, 38(6) :626–635.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2016). *shiny : Web Application Framework for R*. R package version 0.14.1.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6) :1106–1117.

Chen, Y., Meyer, C. A., Liu, T., Li, W., Liu, J. S., and Liu, X. S. (2011). Mm-chip enables integrative analysis of cross-platform and between-laboratory chip-chip or chip-seq data. *Genome biology*, 12(2) :1.

Chen, Y., Negre, N., Li, Q., Mieczkowska, J. O., and Slattery, M. (2012). Systematic evaluation of factors influencing ChIP-seq fidelity. *Nature*.

Chun, H.-J. E., Lim, E. L., Heravi-Moussavi, A., Saberi, S., Mungall, K. L., Bilenky, M., Carles, A., Tse, K., Shlafman, I., Zhu, K., et al. (2016). Genome-wide profiles of extra-cranial malignant rhabdoid tumors reveal heterogeneity and dysregulated developmental pathways. *Cancer Cell*, 29(3) :394–406.

Chung, D., Park, D., Myers, K., Grass, J., Kiley, P., Landick, R., and Keleş, S. (2013). dpeak : high resolution identification of transcription factor binding sites from pet and set chip-seq data. *PLoS Comput Biol*, 9(10) :e1003246.

Cohen, A., Najarian, D., Paulus, A., Guttman, A., Smith, J. A., and Karger, B. (1988). Rapid separation and purification of oligonucleotides by high-performance capillary gel electrophoresis. *Proceedings of the National Academy of Sciences*, 85(24) :9660–9663.

Consortium, E. P. et al. (2004). The encode (encyclopedia of dna elements) project. *Science*, 306(5696) :636–640.

Consortium, E. P. et al. (2011). A user's guide to the encyclopedia of dna elements (encode). *PLoS biol*, 9(4) :e1001046.

Consortium, E. P. et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414) :57–74.

Consortium, T. F. et al. (2014). A promoter-level mammalian expression atlas. *Nature*, 507(7493) :462–470.

Corradin, O. and Scacheri, P. C. (2014). Enhancer variants : evaluating functions in common disease. *Genome medicine*, 6(10) :85.

Craig, R. and Beavis, R. C. (2004). Tandem : matching proteins with tandem mass spectra. *Bioinformatics*, 20(9) :1466–1467.

Cuscó, P. and Filion, G. (2016). Zerone : a chip-seq discretizer for multiple replicates with built-in quality control. *Bioinformatics*, page btw336.

Dagum, L. and Menon, R. (1998). Openmp : an industry standard api for shared-memory programming. *IEEE computational science and engineering*, 5(1) :46–55.

Dao, L. T., Galindo-Albarrán, A. O., Castro-Mondragon, J. A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., Charbonnier, G., Griffon, A., Vanhille, L., Stephen, T., et al. (2017). Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nature Genetics*.

de Boer, B. A., van Duijvenboden, K., van den Boogaard, M., Christoffels, V. M., Barnett, P., and Ruijter, J. M. (2014). Occupeak : Chip-seq peak calling based on internal background modelling. *PloS one*, 9(6) :e99844.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B. K., Muller, H., Ragoussis, J., Wei, C.-L., and Natoli, G. (2010). A large fraction of extragenic rna pol ii transcription sites overlap enhancers. *PLoS Biol*, 8(5) :e1000384.

Dharmalingam, G. and Carroll, T. (2015). *soGGi : Visualise ChIP-seq, MNase-seq and motif occurrence as aggregate plots Summarised Over Grouped Genomic Intervals*. R package version 1.4.4.

Döring, A., Weese, D., Rausch, T., and Reinert, K. (2008). Seqan an efficient, generic c++ library for sequence analysis. *BMC bioinformatics*, 9(1) :11.

Dressman, D., Yan, H., Traverso, G., Kinzler, K. W., and Vogelstein, B. (2003). Transforming single dna molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences*, 100(15) :8817–8822.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910) :133–138.

Elo, L. L., Kallio, A., Laajala, T. D., Hawkins, R. D., Korpelainen, E., and Aittokallio, T. (2011). Optimized detection of transcription factor-binding sites in chip-seq experiments. *Nucleic acids research*, page gkr839.

Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology*, 28(8) :817–825.

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J., Shishkin, A. A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539) :337–343.

Farnham, P. J. (2009). Insights from genomic profiling of transcription factors. *Nature Publishing Group*, 10(9) :605–616.

Faulkner, G. J., Kimura, Y., Daub, C. O., Wani, S., Plessy, C., Irvine, K. M., Schroder, K., Cloonan, N., Steptoe, A. L., Lassmann, T., et al. (2009). The regulated retrotransposon transcriptome of mammalian cells. *Nature genetics*, 41(5) :563–571.

Fedurco, M., Romieu, A., Williams, S., Lawrence, I., and Turcatti, G. (2006). Bta, a novel reagent for dna attachment on glass and efficient generation of solid-phase amplified dna colonies. *Nucleic acids research*, 34(3) :e22–e22.

Fejes, A. P., Robertson, G., Bilenky, M., Varhol, R., Bainbridge, M., and Jones, S. J. (2008). Findpeaks 3.1 : a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24(15) :1729–1730.

Feng, X., Grossman, R., and Stein, L. (2011). Peakranger : a cloud-enabled peak caller for chip-seq data. *BMC bioinformatics*, 12(1) :1.

Fitzgerald, K. A., McWhirter, S. M., Faia, K. L., Rowe, D. C., Latz, E., Golenbock, D. T., Coyle, A. J., Liao, S.-M., and Maniatis, T. (2003). Ikk&epsi ; and tbk1 are essential components of the irf3 signaling pathway. *Nature immunology*, 4(5) :491–496.

Fournier, F., Beauparlant, C. J., Paradis, R., and Droit, A. (2014). rtandem, an r/bioconductor package for ms/ms protein identification. *Bioinformatics*, 30(15) :2233–2234.

Fournier, M., Bourriquen, G., Lamaze, F. C., Côté, M. C., Fournier, É., Joly-Beauparlant, C., Caron, V., Gobeil, S., Droit, A., and Bilodeau, S. (2016). Foxa and master transcription factors recruit mediator and cohesin to the core transcriptional regulatory circuitry of cancer cells. *Scientific reports*, 6.

Frank, C. L., Liu, F., Wijayatunge, R., Song, L., Biegler, M. T., Yang, M. G., Vockley, C. M., Safi, A., Gersbach, C. A., Crawford, G. E., et al. (2015). Regulation of chromatin accessibility and zic binding at enhancers in the developing cerebellum. *Nature neuroscience*, 18(5) :647–656.

Frietze, S., O'Geen, H., Blahnik, K. R., Jin, V. X., and Farnham, P. J. (2010). Znf274 recruits the histone methyltransferase setdb1 to the 3' ends of znf genes. *PloS one*, 5(12) :e15082.

Frith, M. C., Fu, Y., Yu, L., Chen, J.-F., Hansen, U., and Weng, Z. (2004). Detection of functional dna motifs via statistical over-representation. *Nucleic acids research*, 32(4) :1372–1381.

Furey, T. S. (2012). ChIP-seq and beyond : new and improved methodologies to detect and characterize protein-DNA interactions. *Nature Publishing Group*, 13(12) :840–852.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor : open software development for computational biology and bioinformatics. *Genome biology*, 5(10) :1.

Gerasimova, T. I. and Corces, V. G. (2001). Chromatin insulators and boundaries : effects on transcription and nuclear organization. *Annual review of genetics*, 35(1) :193–208.

Ghosh, D. and Qin, Z. S. (2010). Statistical issues in the analysis of ChIP-Seq and RNA-Seq data. *Genes*.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., et al. (2005). Galaxy : a platform for interactive large-scale genome analysis. *Genome research*, 15(10) :1451–1455.

Gjoneska, E., Pfenning, A. R., Mathys, H., Quon, G., Kundaje, A., Tsai, L.-H., and Kellis, M. (2015). Conserved epigenomic signals in mice and humans reveal immune basis of alzheimer/'s disease. *Nature*, 518(7539) :365–369.

Gogol-Döring, A. and Chen, W. (2010). Finding optimal sets of enriched regions in chip-seq data. In *GCB*, pages 113–121.

Guo, Y., Mahony, S., and Gifford, D. K. (2012). High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol*, 8(8) :e1002638.

Halbritter, F., Vaidya, H. J., and Tomlinson, S. R. (2012). Geneprof : analysis of high-throughput sequencing experiments. *Nature methods*, 9(1) :7–8.

Hansen, P., Hecht, J., Ibrahim, D. M., Krannich, A., Truss, M., and Robinson, P. N. (2015). Saturation analysis of chip-seq data for reproducible identification of binding peaks. *Genome research*, 25(9) :1391–1400.

Harmanci, A., Rozowsky, J., and Gerstein, M. (2014). Music : identification of enriched regions in chip-seq experiments using a mappability-corrected multiscale signal processing framework. *Genome biology*, 15(10) :1.

Hayashizaki, Y. (2003). Special fantom2 issue. *Genome research*, 13(6b) :1265–1265.

He, X., Cicek, A. E., Wang, Y., Schulz, M. H., Le, H.-S., and Bar-Joseph, Z. (2015). De novo chip-seq analysis. *Genome biology*, 16(1) :1.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell*, 38(4) :576–589.

Heinz, S., Romanoski, C. E., Benner, C., and Glass, C. K. (2015). The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology*, 16(3) :144–154.

Ho, J. W., Bishop, E., Karchenko, P. V., Nègre, N., White, K. P., and Park, P. J. (2011). Chip-chip versus chip-seq : lessons for experimental design and data analysis. *BMC genomics*, 12(1) :1.

Hower, V., Evans, S. N., and Pachter, L. (2011). Shape-based peak identification for chip-seq. *BMC bioinformatics*, 12(1) :1.

Huang, X. C., Quesada, M. A., and Mathies, R. A. (1992). Capillary array electrophoresis using laser-excited confocal fluorescence detection. *Analytical Chemistry*, 64(8) :967–972.

Humburg, P., Helliwell, C. A., Bulger, D., and Stone, G. (2011). Chipseqr : analysis of chip-seq experiments. *BMC bioinformatics*, 12(1) :1.

Ibrahim, M. M., Lacadie, S. A., and Ohler, U. (2015). Jamm : a peak finder for joint analysis of ngs replicates. *Bioinformatics*, 31(1) :48–55.

Jalili, V., Matteucci, M., Masseroli, M., and Morelli, M. J. (2015). Using combined evidence from replicates to evaluate chip-seq peaks. *Bioinformatics*, 31(17) :2761–2769.

Ji, H., Jiang, H., Ma, W., and Wong, W. H. (2011). Using cisgenome to analyze chip-chip and chip-seq data. *Current Protocols in Bioinformatics*, pages 2–13.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830) :1497–1502.

Jorgenson, J. W. and Lukacs, K. D. (1983). Capillary zone electrophoresis. *Science*, 222(4621) :266–272.

Jothi, R., Cuddapah, S., Barski, A., Cui, K., and Zhao, K. (2008). Genome-wide identification of in vivo protein–dna binding sites from chip-seq data. *Nucleic acids research*, 36(16) :5221–5231.

Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314) :430–435.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., et al. (2010). Variation in transcription factor binding among humans. *science*, 328(5975) :232–235.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C., Suzuki, M., Kawai, J., et al. (2005). Antisense transcription in the mammalian transcriptome. *Science*, 309(5740) :1564–1566.

Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., et al. (2001). Functional annotation of a full-length mouse cdna collection. *Nature*, 409(6821) :685–690.

Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature biotechnology*, 26(12) :1351–1359.

Kidder, B. L., Hu, G., and Zhao, K. (2011). ChIP-Seq : technical considerations for obtaining high-quality data. *Nature immunology*, 12(10) :918–922.

Kim, H., Kim, J., Selby, H., and Gao, D. (2011). A short survey of computational analysis methods in analysing ChIP-seq data. *Human ....*

Kim, N.-K., Jayatillake, R. V., and Spouge, J. L. (2013). Next-peak : a normal-exponential two-peak model for peak-calling in chip-seq data. *BMC genomics*, 14(1) :1.

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295) :182–187.

Kowalczyk, M. S., Hughes, J. R., Garrick, D., Lynch, M. D., Sharpe, J. A., Sloane-Stanley, J. A., McGowan, S. J., De Gobbi, M., Hosseini, M., Vernimmen, D., et al. (2012). Intragenic enhancers act as alternative promoters. *Molecular cell*, 45(4) :447–458.

Kuan, P. F., Chung, D., Pan, G., and Thomson, J. A. (2011a). A statistical framework for the analysis of ChIP-Seq data. *Journal of the ....*

Kuan, P. F., Chung, D., Pan, G., Thomson, J. A., Stewart, R., and Keleş, S. (2011b). A statistical framework for the analysis of chip-seq data. *Journal of the American Statistical Association*, 106(495) :891–903.

Kumar, V., Muratani, M., Rayan, N. A., Kraus, P., Lufkin, T., Ng, H. H., and Prabhakar, S. (2013). Uniform, optimal signal processing of mapped deep-sequencing data. *Nature biotechnology*, 31(7) :615–622.

Kundaje, A., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M., Smith, C. L., Raha, D., Winters, E. E., Johnson, S. M., Snyder, M., Batzoglou, S., and Sidow, A. (2012). Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome research*, 22(9) :1735–1747.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539) :317–330.

Lan, X., Bonneville, R., Apostolos, J., Wu, W., and Jin, V. X. (2011). W-chipeaks : a comprehensive web application tool for processing chip-chip and chip-seq data. *Bioinformatics*, 27(3) :428–430.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822) :860–921.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J., and Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research*, 22(9) :1813–1831.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4) :357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3) :1.

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Comput Biol*, 9(8) :e1003118.

Lefrançois, P., Zheng, W., and Snyder, M. (2010). ChIP-Seq : using high-throughput DNA sequencing for genome-wide identification of transcription factor binding sites. *Methods in enzymology*.

Leleu, M. and Lefebvre, G. (2010). Processing and analyzing ChIP-seq data : from short reads to regulatory interactions. *Briefings in functional . . .*.

Lemaçon, A., Joly Beauparlant, C., Soucy, P., Allen, J., Easton, D., Kraft, P., Simard, J., and Droit, A. (2017). Vexor : an integrative environment for prioritization of functional variants in fine-mapping analysis. *Bioinformatics*.

Lenhard, B., Sandelin, A., and Carninci, P. (2012). Metazoan promoters : emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics*, 13(4) :233–245.

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., et al. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1) :84–98.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv :1303.3997*.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14) :1754–1760.

Li, J. J. and Biggin, M. D. (2015). Statistics requantitates the central dogma. *Science*, 347(6226) :1066–1067.

Li, Q., Barkess, G., and Qian, H. (2006). Chromatin looping and the probability of transcription. *Trends in Genetics*, 22(4) :197–202.

Li, Q., Brown, J. B., Huang, H., and Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *The annals of applied statistics*, pages 1752–1779.

Li, Y., Umbach, D. M., and Li, L. (2014). T-kde : a method for genome-wide identification of constitutive protein binding sites from multiple chip-seq data sets. *BMC genomics*, 15(1) :1.

Liang, K. and Keleş, S. (2012). Normalization of chip-seq data with control. *BMC bioinformatics*, 13(1) :1.

Lihu, A. and Holban, Ş. (2015). A review of ensemble methods for de novo motif discovery in chip-seq data. *Briefings in bioinformatics*, page bbv022.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. *BioMed Research International*, 2012.

Liu, T., Ortiz, J. A., Taing, L., Meyer, C. A., Lee, B., Zhang, Y., Shin, H., Wong, S. S., Ma, J., Lei, Y., et al. (2011). Cistrome : an integrative platform for transcriptional regulation studies. *Genome biology*, 12(8) :1.

Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., et al. (2015). Gateways to the fantom5 promoter level mammalian expression atlas. *Genome biology*, 16(1) :1.

Lun, D. S., Sherrid, A., Weiner, B., Sherman, D. R., and Galagan, J. E. (2009). A blind deconvolution approach to high-resolution mapping of transcription factor binding sites from chip-seq data. *Genome biology*, 10(12) :1.

Lund, E., Oldenburg, A. R., and Collas, P. (2014). Enriched domain detector : a program for detection of wide genomic enrichment domains robust against local variations. *Nucleic acids research*, 42(11) :e92–e92.

Mandoli, A., Singh, A. A., Prange, K. H., Tijchon, E., Oerlemans, M., Dirks, R., Ter Huurne, M., Wierenga, A. T., Janssen-Megens, E. M., Berentsen, K., et al. (2016). The hematopoietic transcription factors runx1 and erg prevent aml1-eto oncogene overexpression and onset of the apoptosis program in t (8 ; 21) amls. *Cell reports*, 17(8) :2087–2100.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in genetics*, 24(3) :133–141.

Marinov, G. K., Kundaje, A., and Park, P. J. (2014). Large-scale quality analysis of published ChIP-seq data. *G3 : Genes| Genomes| . . ..*

Mendenhall, E. M. and Bernstein, B. E. (2008). Chromatin state maps : new technologies, new insights. *Current opinion in genetics & development*, 18(2) :109–115.

Mercier, E., Droit, A., Li, L., Robertson, G., Zhang, X., and Gottardo, R. (2011). An integrated pipeline for the genome-wide analysis of transcription factor binding sites from chip-seq. *PLoS One*, 6(2) :e16432.

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature reviews genetics*, 11(1) :31–46.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153) :553–560.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7) :621–628.

Muiño, J. M., Kaufmann, K., van Ham, R. C., Angenent, G. C., and Krajewski, P. (2011). Chip-seq analysis in r (csar) : An r package for the statistical detection of protein-bound genomic regions. *Plant Methods*, 7(1) :1.

Murakami, K., Elmlund, H., Kalisman, N., Bushnell, D. A., Adams, C. M., Azubel, M., Elmlund, D., Levi-Kalisman, Y., Liu, X., Gibbons, B. J., et al. (2013). Architecture of an rna polymerase ii transcription pre-initiation complex. *Science*, 342(6159) :1238724.

Nagy, G., Czipa, E., Steiner, L., Nagy, T., Pongor, S., Nagy, L., and Barta, E. (2016). Motif oriented high-resolution analysis of chip-seq data reveals the topological order of ctcf and cohesin proteins on dna. *BMC genomics*, 17(1) :637.

Nakato, R., Itoh, T., and Shirahige, K. (2013). Drompa : easy-to-handle peak calling and visualization software for the computational analysis and validation of chip-seq data. *Genes to Cells*, 18(7) :589–601.

Nechaev, S. and Adelman, K. (2011). Pol ii waiting in the starting gates : Regulating the transition from transcription initiation into productive elongation. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1809(1) :34–45.

Newkirk, D., Biesinger, J., Chon, A., Yokomori, K., and Xie, X. (2011). Arem : aligning short reads from chip-sequencing by expectation maximization. *Journal of Computational Biology*, 18(11) :1495–1505.

Nix, D. A., Courdy, S. J., and Boucher, K. M. (2008). Empirical methods for controlling false positives and estimating confidence in chip-seq peaks. *BMC bioinformatics*, 9(1) :1.

Ogbourne, S. and Antalis, T. M. (1998). Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochemical Journal*, 331(1) :1–14.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cdnas. *Nature*, 420(6915) :563–573.

Ong, C.-T. and Corces, V. G. (2011). Enhancer function : new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics*, 12(4) :283–293.

Park, P. J. (2009). ChIP-seq : advantages and challenges of a maturing technology. *Nature Publishing Group*, 10(10) :669–680.

Plank, J. L. and Dean, A. (2014). Enhancer function : mechanistic and genome-wide insights come together. *Molecular cell*, 55(1) :5–14.

Qin, Z. S., Yu, J., Shen, J., Maher, C. A., Hu, M., Kalyana-Sundaram, S., Yu, J., and Chinnaiyan, A. M. (2010). Hpeak : an hmm-based algorithm for defining read-enriched regions in chip-seq data. *BMC bioinformatics*, 11(1) :1.

Quinlan, A. R. and Hall, I. M. (2010). Bedtools : a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6) :841–842.

Rashid, N. U., Giresi, P. G., Ibrahim, J. G., Sun, W., and Lieb, J. D. (2011). Zinba integrates local covariates with dna-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome biology*, 12(7) :1.

Rautio, S. and Lähdesmäki, H. (2015). Mixchip : a probabilistic method for cell type specific protein-dna binding analysis. *BMC bioinformatics*, 16(1) :1.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., et al. (2000). Genome-wide location and function of dna binding proteins. *Science*, 290(5500) :2306–2309.

Rengarajan, J., Mowen, K. A., McBride, K. D., Smith, E. D., Singh, H., and Glimcher, L. H. (2002). Interferon regulatory factor 4 (irf4) interacts with nfatc2 to modulate interleukin 4 gene expression. *The Journal of experimental medicine*, 195(8) :1003–1012.

Rhee, H. S. and Pugh, B. F. (2012). Chip-exo method for identifying genomic location of dna-binding proteins with near-single-nucleotide accuracy. *Current Protocols in Molecular Biology*, pages 21–24.

Rigden, D. J., Fernández-Suárez, X. M., and Galperin, M. Y. (2015). The 2016 database issue of nucleic acids research and an updated molecular biology database collection. *Nucleic acids research*, 44(D1) :D1–D6.

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., et al. (2007). Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods*, 4(8) :651–657.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger : a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1) :139–140.

Roeder, R. (1998). Role of general and gene-specific cofactors in the regulation of eukaryotic transcription. In *Cold Spring Harbor symposia on quantitative biology*, volume 63, pages 201–218. Cold Spring Harbor Laboratory Press.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., and Nyrén, P. (1996). Real-time dna sequencing using detection of pyrophosphate release. *Analytical biochemistry*, 242(1) :84–89.

Ronaghi, M., Uhlén, M., and Nyren, P. (1998). A sequencing method based on real-time pyrophosphate. *Science*, 281(5375) :363.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., Leamon, J. H., Johnson, K., Milgrew, M. J., Edwards, M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356) :348–352.

Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M. B. (2009). Peakseq enables systematic scoring of chip-seq experiments relative to controls. *Nature biotechnology*, 27(1) :66–75.

Sainsbury, S., Bernecky, C., and Cramer, P. (2015). Structural basis of transcription initiation by rna polymerase ii. *Nature Reviews Molecular Cell Biology*, 16(3) :129–143.

Samb, R., Khadraoui, K., Belleau, P., Deschênes, A., Lakhal-Chaieb, L., and Droit, A. (2015). Using informative multinomial-dirichlet prior in a t-mixture with reversible jump estimation of nucleosome positions for genome-wide profiling. *Statistical applications in genetics and molecular biology*, 14(6) :517–532.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12) :5463–5467.

Sawado, T., Halow, J., Bender, M., and Groudine, M. (2003). The β-globin locus control region (lcr) functions primarily by enhancing the transition from transcription initiation to elongation. *Genes & development*, 17(8) :1009–1018.

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome research*, 22(9) :1748–1759.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235) :467.

Schmidt, D., Wilson, M. D., Spyrou, C., and Brown, G. D. (2009). ChIP-seq : using high-throughput sequencing to discover protein–DNA interactions. *Methods*.

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.-M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S. W., et al. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome research*, 25(4) :582–597.

Schones, D. E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5) :887–898.

Shen, L., Shao, N., Liu, X., and Nestler, E. (2014). ngs. plot : Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC genomics*, 15(1) :1.

Shin, H., Liu, T., Manrai, A. K., and Liu, X. S. (2009). Ceas : cis-regulatory element annotation system. *Bioinformatics*, 25(19) :2605–2606.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S., and Hood, L. E. (1985). Fluorescence detection in automated dna sequence analysis. *nature*, 321(6071) :674–679.

Song, Q. and Smith, A. D. (2011). Identifying dispersed epigenomic domains from chip-seq data. *Bioinformatics*, 27(6) :870–871.

Spitz, F. and Furlong, E. E. (2012). Transcription factors : from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9) :613–626.

Spyrou, C., Stark, R., Lynch, A. G., and Tavaré, S. (2009). Bayespeak : Bayesian analysis of chip-seq data. *BMC bioinformatics*, 10(1) :1.

Stanton, K. P., Jin, J., Weissman, S. M., and Kluger, Y. (2015). Ritornello : High fidelity control-free chip-seq peak calling. *bioRxiv*, page 034090.

Stark, R. and Brown, G. (2011). Diffbind : differential binding analysis of chip-seq peak data. *R package version*, 100.

Starmer, J. and Magnuson, T. (2016). Detecting broad domains and narrow peaks in chip-seq data with hiddendomains. *BMC bioinformatics*, 17(1) :1.

Steinhauser, S., Kurzawa, N., Eils, R., and Herrmann, C. (2016). A comprehensive comparison of tools for differential chip-seq analysis. *Briefings in bioinformatics*, page bbv110.

Stempor, P. (2014). *seqplots : An interactive tool for visualizing NGS signals and sequence motif densities along genomic features using average plots and heatmaps.* R package version 1.10.2.

Stormo, G. D. (2000). Dna binding sites : representation and discovery. *Bioinformatics*, 16(1) :16–23.

Stunnenberg, H. G., Hirst, M., Consortium, I. H. E., et al. (2016). The international human epigenome consortium : a blueprint for scientific collaboration and discovery. *Cell*, 167(5) :1145–1149.

Suzuki, H., Forrest, A. R., Van Nimwegen, E., Daub, C. O., Balwierz, P. J., Irvine, K. M., Lassmann, T., Ravasi, T., Hasegawa, Y., De Hoon, M. J., et al. (2009). The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature genetics*, 41(5) :553–562.

Szalkowski, A. M. and Schmid, C. D. (2011). Rapid innovation in chip-seq peak-calling algorithms is outdistancing benchmarking efforts. *Briefings in bioinformatics*, 12(6) :626–633.

Tuteja, G., White, P., Schug, J., and Kaestner, K. H. (2009). Extracting transcription factor targets from chip-seq data. *Nucleic acids research*, 37(17) :e113–e113.

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M., and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods*, 5(9) :829–834.

Vernimmen, D. and Bickmore, W. A. (2015). The hierarchy of transcriptional activation : From enhancer to promoter. *Trends in Genetics*, 31(12) :696–708.

Wang, C., Xu, J., Zhang, D., Wilson, Z. A., and Zhang, D. (2010). An effective approach for identification of in vivo protein-dna binding sites from paired-end chip-seq data. *BMC bioinformatics*, 11(1) :1.

Wang, J., Lunyak, V. V., and Jordan, I. K. (2013). Broadpeak : a novel algorithm for identifying broad peaks in diffuse chip-seq datasets. *Bioinformatics*, page bts722.

Wang, X. and Zhang, X. (2011). Pinpointing transcription factor binding sites from chip-seq data with seqsite. *BMC systems biology*, 5(2) :1.

Weinmann, A. S. and Farnham, P. J. (2002). Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. *Methods*, 26(1) :37–47.

Whiteford, N., Haslam, N., Weber, G., Prügel-Bennett, A., Essex, J. W., Roach, P. L., Bradley, M., and Neylon, C. (2005). An analysis of the feasibility of short read sequencing. *Nucleic acids research*, 33(19) :e171–e171.

Wilder, S. (2010). Swembl : a generic peak-calling program.

Williams, R., Peisajovich, S. G., Miller, O. J., Magdassi, S., Tawfik, D. S., and Griffiths, A. D. (2006). Amplification of complex gene libraries by emulsion pcr. *Nature methods*, 3(7) :545–550.

Wu, D.-Y., Bittencourt, D., Stallcup, M. R., and Siegmund, K. D. (2015). Identifying differential transcription factor binding in chip-seq. *Frontiers in genetics*, 6.

Wu, H. and Ji, H. (2014). Polyapeak : detecting transcription factor binding sites from chip-seq using peak shape information. *PloS one*, 9(3) :e89694.

Xu, H., Handoko, L., Wei, X., Ye, C., Sheng, J., Wei, C.-L., Lin, F., and Sung, W.-K. (2010). A signal–noise model for significance analysis of chip-seq with negative control. *Bioinformatics*, 26(9) :1199–1204.

Xu, S., Grullon, S., Ge, K., and Peng, W. (2014). Spatial clustering for identification of chip-enriched regions (sicer) to map regions of histone methylation patterns in embryonic stem cells. *Stem Cell Transcriptional Networks : Methods and Protocols*, pages 97–111.

Yang, Y., Fear, J., Hu, J., Haecker, I., Zhou, L., Renne, R., Bloom, D., and McIntyre, L. M. (2014). Leveraging biological replicates to improve analysis in chip-seq experiments. *Computational and structural biotechnology journal*, 9(13) :1–10.

Young, M. D., Willson, T. A., Wakefield, M. J., Trounson, E., Hilton, D. J., Blewitt, M. E., Oshlack, A., and Majewski, I. J. (2011). Chip-seq analysis reveals distinct h3k27me3 profiles that correlate with transcriptional activity. *Nucleic acids research*, 39(17) :7415–7427.

Yu, G., Wang, L.-G., and He, Q.-Y. (2015). Chipseeker : an r/bioconductor package for chip peak annotation, comparison and visualization. *Bioinformatics*, page btv145.

Zeng, X., Li, B., Welch, R., Rojo, C., Zheng, Y., Dewey, C. N., and Keleş, S. (2015). Perm-seq : Mapping protein-dna interactions in segmental duplication and highly repetitive regions of genomes with prior-enhanced read mapping. *PLoS Comput Biol*, 11(10) :e1004491.

Zeng, X., Sanalkumar, R., Bresnick, E. H., Li, H., Chang, Q., and Keleş, S. (2013). jmosaics : joint analysis of multiple chip-seq datasets. *Genome biology*, 14(4) :1.

Zhang, X., Robertson, G., Krzywinski, M., Ning, K., Droit, A., Jones, S., and Gottardo, R. (2011). Pics : Probabilistic inference for chip-seq. *Biometrics*, 67(1) :151–163.

Zhang, X., Robertson, G., Woo, S., Hoffman, B. G., and Gottardo, R. (2012). Probabilistic inference for nucleosome positioning with mnase-based or sonicated short-read data. *PloS one*, 7(2) :e32095.

Zhang, Y., Lin, Y.-H., Johnson, T. D., Rozek, L. S., and Sartor, M. A. (2014). Pepr : a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated chip-seq data. *Bioinformatics*, page btu372.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., et al. (2008). Model-based analysis of chip-seq (macs). *Genome biology*, 9(9) :1.

Zhu, L. J., Gazin, C., Lawson, N. D., Pagès, H., Lin, S. M., Lapointe, D. S., and Green, M. R. (2010). Chippeakanno : a bioconductor package to annotate chip-seq and chip-chip data. *BMC bioinformatics*, 11(1) :1.