Régulation de l'expression génique par le facteur de transcription SPI1/PU.1 dans l'érythroleucémie : mécanismes de répression des gènes par sa liaison à l'ADN, conséquences de sa liaison à l'ARN

Lélia Polit

université
PARIS-SACLAY

Régulation de l'expression génique par le facteur de transcription SPI1/PU.1 dans l'érythroleucémie : mécanismes de répression des gènes par sa liaison à l'ADN, conséquences de sa liaison à l'ARN

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 577, Structure et Dynamique des Systèmes Vivants (SDSV)

Spécialité de doctorat : Sciences de la Vie et de la Santé

Unités de recherche (en co-encadrement) :
Université de Paris, Inserm, CNRS, Institut Cochin, Département Développement, Reproduction et Cancer, 75014 Paris, France
Université Paris-Saclay, Inserm, Institut Gustave Roussy, Dynamique moléculaire de la transformation hématopoïétique, 94805 Villejuif, France

Référent : Faculté des sciences d'Orsay

Thèse présentée et soutenue à Paris, le 11 Décembre 2020, par

Lélia POLIT

Composition du jury:

Daniel GAUTHERET Professeur, Université Paris Saclay	Président
Morgane THOMAS-CHOLLIER Maître de conférence (HDR), ENS Ulm	Rapportrice & Examinatrice
Jacques VAN HELDEN Professeur, Université d'Aix-Marseille	Rapporteur & Examinateur
Olivier KOSMIDER PU-PH, Université de Paris	Examinateur
Valentina BOEVA Maître de conférence, ETH Zurich	Directrice de thèse
Christel GUILLOUF DR, Institut Gustave Roussy	Directrice de thèse

Thèse de doctorat

NNT: 2020UPASL064

.

Il faut beaucoup pardonner à cette vie incompréhensible. Il faut tout lui pardonner pour cette douceur inouïe qu'elle exerce par surprise.

La nuit du coeur, Christian Bobin

Remerciements

Je souhaite tout d'abord exprimer ma reconnaissance aux membres de mon jury, les professeurs Morgane Thomas-Chollier et Jacques Van Helden d'avoir accepté de rapporter le manuscrit de ma thèse. Je souhaite également remercier les professeurs Daniel Gautheret et Olivier Kosmider, d'avoir accepté de participer à ce jury.

Je remercie bien évidemment mes deux directrices de thèse, Christel Guillouf et Valentina Boeva, pour toutes les connaissances et les échanges fructueux que nous avons pu avoir depuis Octobre 2017. Ce travail de thèse se trouve à la croisée des disciplines que sont l'informatique, les mathématiques et la biologie, leurs compétences dans ces disciplines m'ont permis de progresser dans ces différents domaines.

Merci à ma famille, sans qui rien de tout cela n'aurait été possible.
Mon papa et ma maman, des exemples de vie pour moi. Papa, merci de m'avoir donné ce goût pour la science, de m'avoir toujours poussée à aller plus loin et de m'avoir guidée, merci de m'avoir toujours soutenue et écoutée et d'avoir toujours su me faire relativiser les problèmes et obstacles que je voyais comme immenses. Maman, merci d'avoir toujours été une oreille très attentive, de m'avoir écoutée parler de tout et de rien, de m'avoir toujours soutenue dans mes choix et de t'être si bien occupée de moi comme à mon plus jeune âge. Titouan, mon petit frère, lorsque j'ai commencé ma thèse tu étais encore au lycée à la maison avec papa et maman, te voilà aujourd'hui étudiant à l'Université à Paris, marchant dans mes "traces" sans s'être trompé de chemin avant ! Merci d'être la personne que tu es aujourd'hui, contribuant à ma culture cinématographique et informatique.

Un grand merci à toutes les personnes qui sont passées par l'équipe "(Epi-)Génétique Computationnelle du cancer" à l'Institut Cochin, particulièrement Floriane, Mélissa, Malo, Samira. Je souhaiterai remercier plus particulièrement Gwenneg, sans qui la thèse n'aurait pas été pareille. Merci de m'avoir écoutée parler de SPI1 et de toutes les autres choses pendant presque 3 ans, merci pour tes très bons conseils et ton analyse scientifique toujours très exacte.

Merci à Sebastian pour ta pleine disponibilité, merci d'avoir pris le temps de m'expliquer des concepts biologiques parfois abstraits pour moi. Nous formons une bonne équipe. J'espère

que tu trouveras chaussure à ton pied après ta soutenance de l'autre côté de l'Atlantique.

J'aimerai également et surtout remercier Camille et Emmanuel qui ont été là du début à la fin de ma thèse, Selin, Boris et Bolaji qui sont partis en cours de route pour de nouveaux horizons : maternels, chinois ou privé, Pauline, qui, sans jamais avoir travaillé avec nous occupe une place tout aussi importante dans ma vie, merci pour votre soutien sans faille dans les moments où tout allait bien mais aussi et surtout dans les moments où la thèse et/ou la vie a été un peu moins douce...

Laura, cette oreille attentive pour parler de science ou bien de tout autre chose. Tout simplement, merci d'être toi. Manon, merci pour ces moments partagés depuis notre rencontre et pour ceux qui viendront. Tu es toujours de bon conseil et une oreille attentive.

Je souhaiterai adresser un remerciement particulier à Renaud Dentin qui m'a gentiment accueillie dans ses locaux pour ne pas que j'écrive mon manuscrit et termine ma thèse seule dans un immense bureau, autrefois plein.

Merci finalement à tous les membres de l'Institut Cochin avec qui j'ai pu discuter plusieurs fois, une seule fois, partager un verre (dans la vie d'avant) ou bien simplement échanger un bonjour dans un couloir.

Marie, merci d'être cette personne à qui je peux toujours tout dire, merci de m'avoir écoutée me plaindre de la thèse et du reste pendant 3 ans, d'avoir toujours été là à n'importe quelle heure et à n'importe quel endroit, même depuis l'autre bout du monde.

Elodie, sans qui ma vie serait beaucoup moins drôle, merci d'être venue me voir presque tous les week-end à Paris au début de la thèse pour faire la fête et profiter de la vie. Merci d'être cette personne avec qui j'ai l'impression que tout va toujours bien.

Lucas, Louis, les copaaaains merci de m'avoir toujours écoutée et soutenue pendant ces trois années qui furent longues et compliquées surtout sur la fin. Merci pour ces moments partagés et pour tous ceux à venir.

Et tous les autres, Sophie, Mathilde, Manon, Charlotte, Laura, Julien, Lucas, les Baptiste, Sarah, Guillaume, Timothé, Luc, Agathe, Paul, Lise, Alice, Maxence, Ben, Abi, Marion. Merci pour tous les moments partagés qui sont souvent joyeux et festifs.

Un petit clin d'oeil aussi à Valérie, Jérôme, Jérémy, Charlotte, Yéluna, Jess sans qui les après-travail n'auraient pas été pareils. Merci d'avoir toujours apporté de la bonne humeur à notre table dans ce lieu unique. Dommage qu'en ce moment on se voit moins mais cela aura peut être au moins contribué à la qualité de ce manuscrit.

Résumé

Français

Dans le lignage érythroïde, une expression anormale et non-contrôlée du facteur de transcription SPI1 entraîne une leucémie aiguë, en partie en inhibant la différenciation érythroïde et l'apoptose des progéniteurs engagés dans la différenciation érythroïde. Mon travail de thèse concerne la caractérisation des mécanismes par lesquels SPI1 réprime l'expression génique dans un modèle d'érythroleucémie murine. En utilisant des cellules pré-leucémiques issues de souris transgéniques pour spi1, dans lesquelles l'expression de *spi1* peut être contrôlée, j'ai comparé des données de séquençage à haut débit pour caractériser l'accessibilité de la chromatine, les modifications épigénétiques des protéines histones et l'expression des gènes en fonction de la présence ou de l'absence de SPI1. Pour comparer les signaux de ChIP-seq entre différentes conditions, nous avons développé un package R : ChIP-seq Intersample Normalization (CHIPIN). Nous avons ainsi démontré que la répression des gènes par SPI1, dont certains sont liés à la différenciation érythroïde et à l'apoptose, est basée sur la coordination de deux mécanismes qui impliquent et sont contrôlés par l'histone dé-acétylase 1 (HDAC1) et le complexe répressif Polycomb (PRC2). La caractérisation de la fixation de SPI1 à l'ARN en utilisant des données de CLIP-seq nous a permis de montrer que cette fixation n'était pas liée à la régulation de l'expression génique. Nous proposons un nouveau mécanisme pour la répression de l'expression génique par SPI1 dans l'érythroleucémie en coopération avec deux facteurs épigénétiques : PRC2 et HDAC1, et un nouveau package R pour la normalisation des données de ChIP-seq.

English

In the erythroid lineage, abnormal and uncontrolled expression of the transcription factor SPI1 leads to acute leukaemia, in part by inhibiting erythroid differentiation and apoptosis of progenitors involved in erythroid differentiation. My PhD thesis is dedicated to the characterization of the mechanisms by which SPI1 represses gene expression in a mouse erythroleukemia model. Using pre-leukemic cells from *spi1* transgenic mice, in which spi1 expression can be controlled, I compared high-throughput sequencing data to characterise chromatin accessibility, epigenetic modifications of histone proteins and gene expression as a function of the presence or absence of SPI1. To compare ChIP-seq signals between different conditions, we developed an R package : ChIP-seq Intersample Normalization (CHIPIN). We demonstrated that gene repression by SPI1, including genes coding for apoptosis and erythroid differentiation, is based on the coordination of two mechanisms which involve and are controlled by histone deacetylase 1 (HDAC1) and the Polycomb repressive complex (PRC2). Characterisation of the binding of SPI1 to RNA using CLIP-seq data allowed us to show that this binding was not linked to regulation of gene expression. We propose a new mechanism for the repression of gene expression by SPI1 in erythroleukemia in cooperation with two epigenetic factors : PRC2 and HDAC1, and a new R package for the normalization of ChIP-seq data.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

AP-1	Activator Protein 1
ATAC-seq	Assay for Transposase- Accessible Chromatin with high-throughput sequencing
BRD4	BromoDomain containing 4
C/EBP	CAAT Enhancer Binding Protein
ChIP-seq	Chromatin ImmunoPrecipitation followed by high-throughput sequencing
CLIP-seq	CrossLink ImmunoPrecipiatation with high-throughput sequencing
CTCF	11-zinc finger protein
FACT	Facilitate Chromatin Transcription
G-CSF	Granulocyte Colony Stimulating Factor
HDAC1	Histone Desacetylase 1
IRF4	Interferon Regulatory Factor 4
MAPK	Mitogene Activated Protein Kinase
M-CSF	Macrophage Colony Stimulating Factor
MeCP2	Methyl CpG Binding Protein 2
MG-CSF	Macrophage Granulocyte Colony Stimulating Factor
NDR	Nucleosome Depleted Region
NF-kappaB	Nuclear Factor Kappa B
OCT-1	POU Domain Class 2, Transcription Factor 1
PcG	PolyComb Gene complex
PCR	Poly Chain Reaction
PI3K	PhosphoInnositide 3 Kinase
RNAPolII	RNA Polymerase II
SP-1	Specificity Protein I
SPI1	Spleen Provirus Integration 1
SWI/SNF	SWItch/Sucrose Non-Fermentable
TEFb	Transcription Elongation Factor b
TFII	Transcription Factor II

TSS	Transcription Start Site
TES	Transcription End Site

Chapitre 1

Introduction

1.1 La régulation de l'expression génique

1.1.1 La transcription et ses acteurs

Un organisme est constitué d'une multitude de cellules, très différentes les unes des autres selon le tissu auquel elles appartiennent. Elles possèdent toutes exactement le même patrimoine génétique codé par plusieurs molécules d'ADN [1]. Toutefois, les différences entre un neurone, un érythrocyte et un hépatocyte sont si extrêmes qu'il est difficile d'imaginer que ces cellules contiennent le même ADN. Ainsi, chaque type cellulaire possède un programme d'expression génique qui lui est propre et qui est finement régulé, de la synthèse d'ARN [2] à l'assemblage de la protéine qui permettent la création et la survie des différents types cellulaires, avec leurs caractéristiques propres. L'expression génique est donc essentielle pour tous les systèmes vivants, tant pour leur création que pour le maintien de leur identité et leur fonctionnement.

Les éléments régulateurs de la transcription

La première étape de la transcription, appelée **initiation de la transcription** (Fig. 1.1), est régulée par le recrutement et l'association de protéines régulatrices nommées facteurs de transcription au niveau de séquences d'ADN plus ou moins proches du site d'initiation de la transcription (TSS) [1]. Les facteurs de transcription généraux TFIIA, TFIIB, TFIID, TFIIE, TFIIH sont nécessaires et indispensables pour que la transcription ait lieu. Ils font partie du complexe de pré-initiation, assemblé sur le promoteur (Fig. 1.1). Le promoteur est la région de l'ADN, propre à chaque gène, sur laquelle se fixe le complexe de pré-initiation et où se déroule l'initiation de la transcription, il est indispensable à la transcription de l'ADN. Il comprend le ou les TSSs du gène, son étendue n'est pas définie avec précision [2], mais on définit généra-

1. **ADN** : **A**cide **D**esoxyribo**N**ucléique
2. **ARN** : **A**cide **R**ibo**N**ucléique

lement le promoteur basal comme englobant le TSS [3] ainsi que les régions immédiatement en amont et en aval, tandis que la promoteur proximal est défini comme comprenant les régions en amont plus éloignées du TSS. Le promoteur basal contient des séquences particulières qui sont reconnues par les facteurs de transcription généraux, la plus connue est le motif TATA-box (Fig. 1.1 A). Il est reconnu par un composant du TFIID ce qui induit le recrutement de la RNAPolII et l'assemblage du complexe de pré-initiation de la transcription. Le motif TATA-box a une position fixe par rapport au TSS le plus utilisé (autour de 30 paires de base (bp)) [3]. Dans les promoteurs proximaux on trouve des motifs CAAT-box et des des ilots CG.

La régulation de cette étape d'initiation de la transcription est également assurée par des facteurs de transcription spécifiques [1]. Ils reconnaissent de petites séquences d'ADN localisées en amont du TSS, le facteur SP-1 en est un exemple. Il existe également des facteurs de transcription spécifiques à certains gènes. La particularité de ces facteurs de transcription est d'avoir un domaine de liaison à l'ADN, cela leur permet de reconnaître et de se fixer sur une séquence d'ADN spécifique qui peut être une région cis-régulatrice distale (nommée enhancer dans la suite) ou un promoteur. Une région enhancer est une région de l'ADN capable de moduler l'expression d'un gène positionnée en 5' ou en 3' du TSS et pouvant être éloignée du TSS jusqu'à 1.7M bases [4]. Elles peuvent être activatrices ou répressives. Les enhancers et les promoteurs proximaux ont la fonction de moduler les niveaux de transcription. Les promoteurs proximaux sont des régions contenant des sites de reconnaissance permettant le recrutement de protéines stimulatrices ou inhibitrices de la transcription. L'agencement tridimensionnel de la molécule d'ADN permet le contact entre ces régions enhancers et les promoteurs de leurs gènes cibles. La formation d'une boucle promoteur-enhancer est médiée par la cohésine, la protéine CTCF et le complexe médiateur permettant le rapprochement de l'enhancer vers son promoteur cible afin de favoriser l'activation de l'expression génique. CTCF et la cohésine peuvent également créer une structure tridimensionnelle qui interdit le glissement de l'ADN sur les nucléosomes pour empêcher certaines interactions promoteur-enhancer. Les Topological Associated Domains (TAD) sont des régions du génome au sein desquelles les séquences d'ADN interagissent plus entre elles qu'avec les autres régions du génome. Un TAD mesure environ 880 kilobases et à ses bornes on trouve un enrichissement important en sites de liaison pour la protéine CTCF et pour la cohésine.

En plus de leur domaine de liaison à l'ADN, les facteurs de transcription spécifiques possèdent différents domaines fonctionnels ou structuraux pouvant inclure : (i) un domaine de transactivation, (ii) un site d'accrochage de ligand, (iii) un domaine de dimérisation [1]. Les facteurs de transcription peuvent agir seuls ou en complexe avec des co-facteurs, et sont capables de favoriser ou de réprimer l'activité transcriptionelle de leurs gènes cibles. Quatre groupes de facteurs de transcription sont principalement décrits sur la base de la structure de leur domaine de liaison à l'ADN [5] : (i) la famille hélice-coude-hélice ; (ii) la famille à doigt de zinc,

3. **TSS : T**ranscription **S**tart **S**ite

tous les facteurs de transcription de cette famille ont un ion Zn2+ qui stabilise les motifs sous forme de doigt, ces protéines sont organisées sous forme de répétition (le facteur SP-1 en est un exemple et possède trois doigts) ; (iii) la famille à glissière de leucine, les facteurs de transcription de cette famille fixent l'ADN sous forme de dimères via les deux branches du Y qui lient l'ADN (C/ECBP en est un exemple) ; (iv) la famille hélice-boucle-hélice, la liaison de la protéine à l'ADN entraîne sa dimérisation. Récemment, une classification plus fine des facteurs de transcription a été publiée (TFClass [6]).

Pré-initiation de la transcription

La mise en place du complexe de pré-initiation de la transcription est principalement réalisée par le facteur TFIID dont la sous unité de liaison (TBP pour "TATA box-binding protein" Fig. 1.1A) reconnait la TATA-box. Certains promoteurs ne possèdent pas de TATA-box, d'autres sous unités du facteur TFIID peuvent alors reconnaître des séquences particulières présentes dans le promoteur basal. Une fois que le facteur TFIID est en place sur le promoteur (Fig. 1.1B) les autres éléments du complexe d'initiation de la transcription vont être recruté soit selon un assemblage séquentiel des différents facteurs selon un ordre défini, soit par recrutement d'un complexe préformé qui contient tous les facteurs de transcription généraux, la RNAPolII ainsi que ses co-régulateurs [7] (Fig. 1.1C). Les niveaux de transcription peuvent être modulés par l'action de co-régulateurs présents au niveau des séquences régulatrices, le plus connu est le complexe Médiateur. Il est composé de 26 sous unités et est décrit comme permettant d'établir un pont entre les facteurs de régulation et la machinerie de transcription basale. Il interagit avec un grand nombre de facteurs de transcription (dont les généraux) et avec la RNAPolII [8]. Il peut augmenter le niveau de transcription d'un gène mais également stabiliser la formation du complexe de pré-initiation. En outre il régule le phénomène d'entrée et/ou sortie de pause de la RNAPolII.

Elongation de la transcription

Le passage d'initiation de la transcription à élongation est assuré par la phosphorylation de la sérine 5 de la RNAPolII par TFIIH [9]. La phosphorylation de la sérine 5 entraîne la déstabilisation de l'interaction entre la RNAPolII et les autres composants du complexe de pré-initiation pour favoriser le processus d'échappement de la RNAPolII depuis le promoteur. La phase de pause évoquée à la fin du paragraphe précédent et régulée par le complexe Médiator se caractérise par un ralentissement de la RNAPolII de 20 à 50bp en aval du TSS. Le rôle de cette étape de pause serait de déclencher rapidement la transcription en réponse à un stimulus et d'assurer le maintien de la molécule d'ADN dans une forme sujette à la transcription. A la suite de la phase de pause deux choses sont possibles : terminaison de la transcription par relargage de petits ARN ou élongation productive. La sortie de la phase de pause est assurée par le recrutement

d'un complexe dont le rôle est de lever l'inhibition de l'élongation et de favoriser le recrutement de divers complexes associés à la polymérase durant l'élongation [10]. Il est associé à la phosphorylation de la sérine 2, impliquant p-TEFb lui même régulé par les protéines à bromodomaine reconnaissant des histones modifiées de la chromatine, tel que BRD4 [11] D'autres protéines telle que c-myc [12] sont impliquées dans la régulation de l'étape d'élongation.

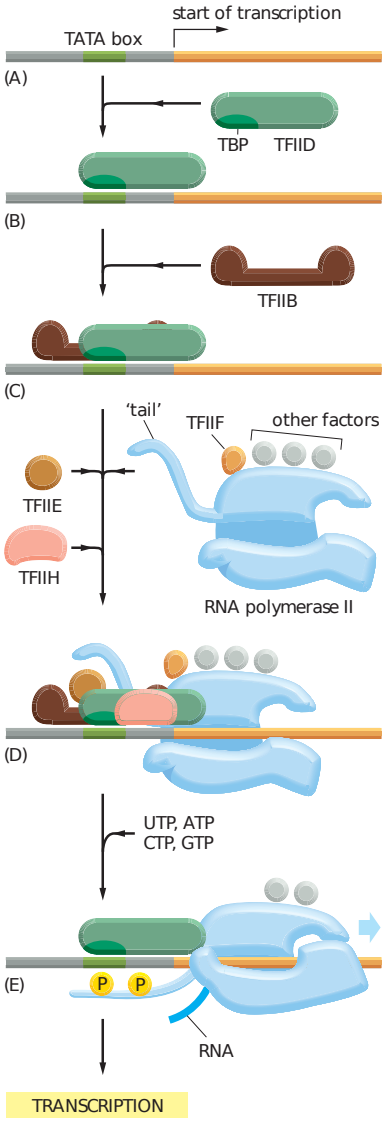


FIGURE 1.1 – **Étapes principales de l'initiation de la transcription.** *Issue de [5].* **A.** Les promoteurs contiennent des séquences particulières, dont la TATA box. **B.** Ces séquences sont reconnues par le facteur de transcription TFIID qui induit la liaison adjacente de TFIIB. **C, D.** Les autres facteurs de transcriptions généraux, TFIIE, TFIIH, TFIIF sont assemblés sur le promoteur, ainsi que la RNAPolII. **E.** La double hélice d'ADN s'ouvre au niveau du point de départ de la transcription, la RNAPolII est phosphorylée par TFIIH afin de commencer la phase d'élongation de la transcription.

1.1.2 La chromatine : organisation fonctionnelle du génome

Dans les cellules eucaryotes, le matériel génétique est organisé en une structure complexe composée d'ADN et de protéines. Chaque cellule possède environ deux mètres d'ADN qui doivent être contenus dans un noyau d'environ $10\mu m$, le matériel génétique est donc sujet à un haut niveau de compaction (Fig. 1.2). L'unité fondamentale de la chromatine est le nucléosome. Un nucléosome est composé d'une particule cœur dont la structure est très conservée parmi les espèces. Cette particule coeur est composée de 146 bp d'ADN enroulées selon environ 1.7 tours autour d'un octamère protéique comprenant deux exemplaires de chacune des histones H2A, H2B, H3, H4 (Fig. 1.3). Cette structure est répétée régulièrement tout au long de la molécule d'ADN pour former le nucléofilament. Celui-ci peut adopter des niveaux d'organisation plus compacts jusqu'au niveau de condensation le plus élevé : le chromosome métaphasique (Fig. 1.2). La longueur de la région entre les nucléosomes, appelée région inter-nucléosomale, varie selon les espèces et le type cellulaire, au niveau de cette région des histones inter-nucléosomales (H1) peuvent être incorporées. Ainsi la longueur d'ADN caractéristique d'un nucléosome varie selon les espèces entre 160bp et 241bp. Les protéines histones qui constituent la particule cœur sont de petites protéines riches en acides aminés positivement chargés (lysine et arginine). Ces charges positives leur permettent de se fixer étroitement à la molécule d'ADN qui elle, est chargée négativement.

1.1.3 La chromatine : régulation de l'activité transcriptionnelle

Le maintien de l'intégrité de l'information génétique comprend le maintien de la séquence nucléotidique mais aussi son organisation en chromatine. Outre les mutations qui sont des modifications de la séquence d'ADN, la chromatine est sujette à des modifications épigénétiques. Ces modifications épigénétiques consistent en des changements de la chromatine qui ne modifient pas la nature de la séquence ADN et qui sont réversibles et héritables. Ainsi, lors de la réplication du génome, en plus de la duplication de la séquence d'ADN, les caractères épigénétiques doivent également être transmis.

L'assemblage de l'ADN en chromatine se fait en plusieurs étapes : (i) formation de son unité fondamentale, le nucléosome via la mise en place des deux tétramères d'histones ; (ii) maturation au cours de laquelle les nucléosomes sont espacés de manière régulière pour former le nucléofilament ; (iii) incorporation des histones inter-nucléosomales et repliement du nucléofilament en fibre de 30 nm ; (iii) repliements successifs qui conduisent à des niveaux d'organisation supérieurs en domaines spécifiques dans le génome. Dès les premières étapes de l'assemblage de la chromatine, le nucléosome peut être soumis à des variations. L'ADN peut être méthylé [5] ou bien les histones peuvent présenter des modifications post-traductionnelles. Toutes ces variations peuvent induire des différences dans la structure de la chromatine. Les facteurs responsables de ces variations peuvent être de deux types : (i) des facteurs de remode-

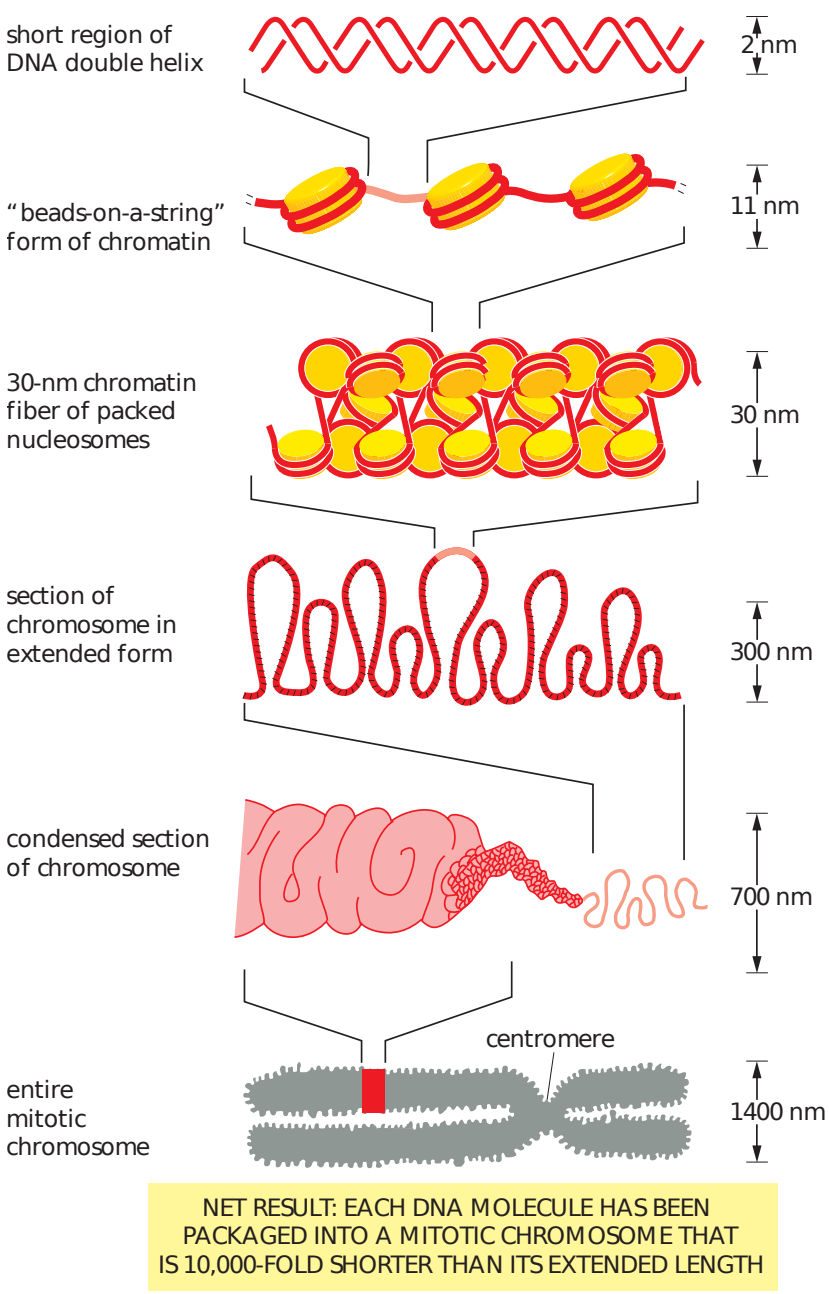


FIGURE 1.2 – **Différent niveaux de compaction de l'ADN : de la double hélice d'ADN au chromosome métaphasique**. *Issue de [5].*

lage de la chromatine qui ont une activité ATPase ; (ii) des enzymes qui modifient chimiquement les protéines histones post-traductionnellement. Les facteurs de remodelage sont des complexes multi-protéiques (familles SWI/SNF [13], ISWI [14], Mi2/NuRD [15]), dont l'activité ATPase permet le remaniement de l'organisation nucléosomique. Même ces trois familles semblent capables d'exercer l'activation et la répression de la transcription, par leurs actions sur la structure chromatinienne, la famille SWI/SNF est majoritairement impliquée dans l'activation de la trans-

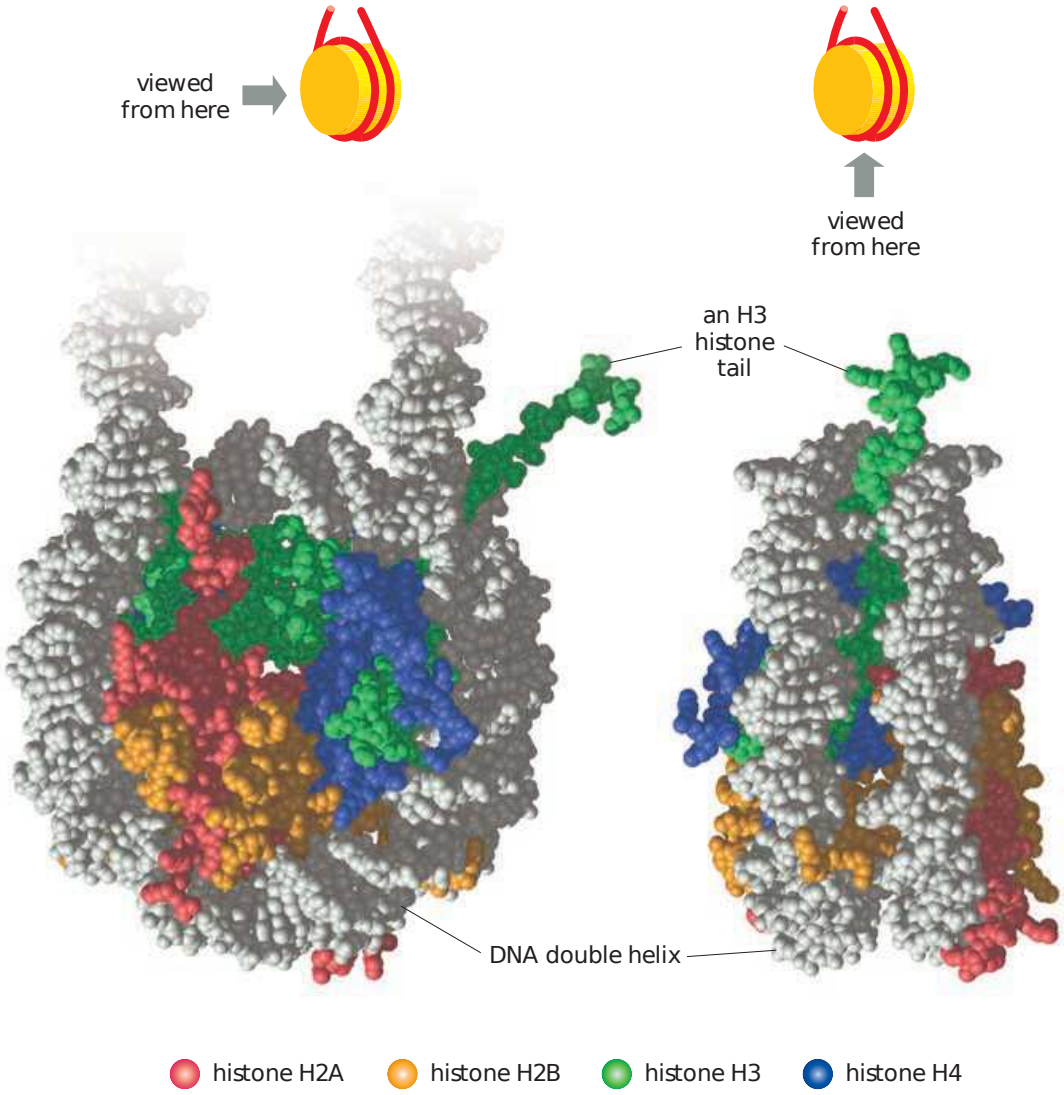


FIGURE 1.3 – **Structure de l'unité fondamentale de la chromatine : le nucléosome.** *Issue de [5].* Cette structure a été déterminée par diffraction des rayons X. Deux vues de la structure du nucléosomes sont montrées ici, la double hélice d'ADN est representée en gris tandis que les différentes protéines histones, H2A, H2B, H3 et H4 sont representées par des couleurs. L'extrémité de l'histone H3 sujette à des modifications chimiques est mise en évidence.

cription tandis que les complexes ISWI et Mi2/NuRD sont, eux, principalement responsables de la répression de la transcription.

Les extrémités N-terminales des histones (extrémité N-terminale de l'histone H3 en vert la

Fig. 1.3) pointent vers l'extérieur du nucléosome et sont la cible d'enzymes appelées facteurs épigénétiques. Les résidus tels que les lysines, sérines, thréonines, arginines peuvent être modifiés (Fig. 1.4) par acétylation, phosphorylation, méthylation, citrullination, O-N-acétylglucosaminylation, PARylation, ubiquitylation ou encore sumoylation. Ces modifications chimiques entrainent des modifications dans les interactions entre les charges des nucléotides et les charges des résidus des histones, cela affecte la stabilité des interactions chimiques intra- ou inter-nucléosomales entrainant ainsi la condensation ou la décondensation de la chromatine. Ces modifications sont réversibles grâce à l'action d'enzymes "d'effacement" par opposition aux enzymes "d'apposition", cela permet l'acquisition d'un équilibre entre les propriétés activatrices et répressives de ces modifications.

L'activité transcriptionnelle d'un gène est dépendante de l'état de sa chromatine. En effet, plus la chromatine d'un gène sera relâchée, plus les facteurs responsables de la régulation de l'étape de pré-initiation de la transcription, cités précédemment (section 1.1.1), auront la possibilité de se fixer sur l'ADN, afin de positionner la RNAPolII pour initier la transcription. De plus, pour mener à bien la phase d'élongation de la transcription la RNAPolII doit pouvoir parcourir l'ADN aisément. Pour cela, certaines protéines et certains complexes, comme le complexe protéique FACT ("facilitates chromatin transcription"), restructurent la chromatine et les nucléosomes. Le facteur général de transcription TFIIH a une activité hélicase et va faciliter la transcription également. Les protéines Spt [17] et FACT sont responsables de la reconstruction de la chromatine. En revanche, si la chromatine du gène est condensée, c'est à dire si les nucléosomes sont plus proches entre eux, alors le promoteur du gène ne sera pas accessible à ces facteurs. L'acétylation des protéines histones est associée à une chromatine accessible, en effet, cette modification chimique induit une modification de charge de la molécule d'ADN qui est plus relachée autour des protéines histones (Fig. 1.4). Cela induit également le recrutement des protéines à bromo-domaines telles que SWI/SNF et BRD4 (jouant un rôle dans l'élongation) par modification de la charge de la molécule d'ADN. En revanche, la méthylation (de certains résidus), par exemple de la lysine 27 de l'histone H3 (H3K27me3), est associée à une chromatine condensée et est liée à la répression de l'expression génique (Fig. 1.4) [18, 19]. Ainsi, les facteurs épigénétiques peuvent avoir des effets positifs ou négatifs sur l'expression génique. La majorité de la molécule ADN de la cellule eucaryote se trouve enroulée autour des nucléosomes et est donc occultée par les protéines histones. Il a été montré que les facteurs de transcription se fixaient préférentiellement sur des séquences qui se trouvent dans des régions à forte densité de nucléosomes [20]. Or les régions cis-régulatrices, cibles des facteurs de transcription, qui sont normalement peu présentes sur les nucléosomes, ont tendance à être incorporées dans ceux-ci [21]. Pour expliquer cela, il a été montré que ces régions cis-régulatrices sont déplacées par des facteurs de transcription qui sont capables de se lier à des régions de chromatine très condensées, à forte densité de nucléosomes. On appelle ces facteurs de transcription des facteurs pionniers [22]. Ceux-ci sont capables de rendre localement plus accessible la chromatine

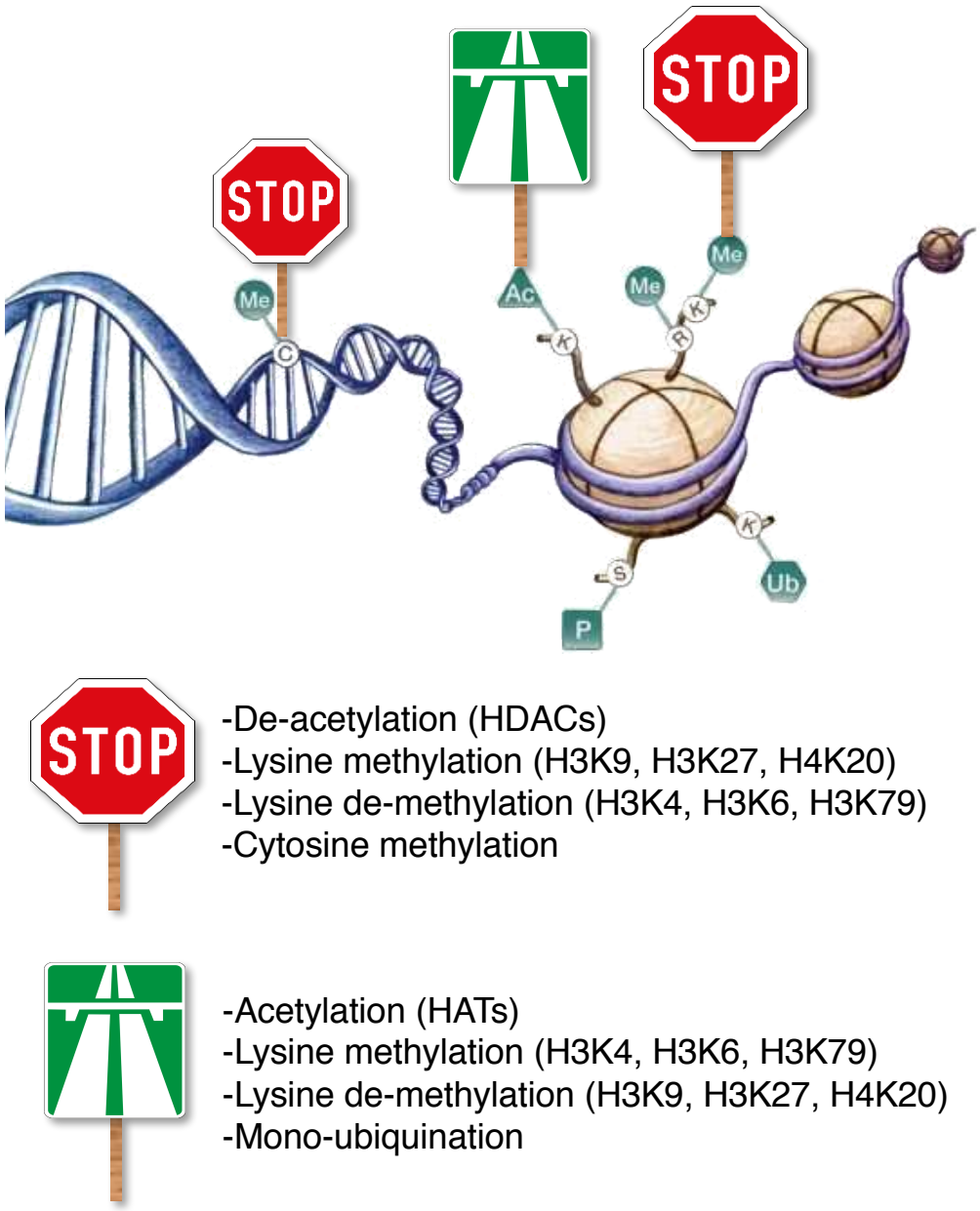


FIGURE 1.4 – **Modification des extrémités N-terminales des histones et conséquences sur l'expression génique.** *Adapté de [16].*

et d'augmenter l'accessibilité à celle-ci pour les facteurs de transcription ou les facteurs épigénétiques. Le facteur de transcription SPI1/PU.1 qui est l'objet de mon travail de thèse, est un de ces facteurs pionniers.

1.2 Le facteur de transcription SPI1, un acteur majeur de l'hématopoïèse

1.2.1 L'hématopoïèse

L'hématopoïèse assure la production de toutes les cellules du sang depuis une très petite population de cellules souches hématopoïétiques (CSH) pluripotentes. Chez l'homme et la souris, avant la naissance, l'hématopoïèse se produit d'abord dans les îlots de sang du sac vitellin, puis dans le foie, la rate et les ganglions lymphatiques. Après la naissance l'hématopoïèse se produit dans la moelle osseuse. Les CSH ont la capacité de générer tous les types de cellules sanguines différenciées, c'est un aspect fondamental de l'hématopoïèse. Toutefois, ce processus permettant la création des différentes cellules fonctionnelles et différenciées est encore largement discuté. Les premiers modèles ont proposé qu'une CSH génère deux types de progéniteurs pluripotents : le Progéniteur Lymphoïde Commun (CLP) et le Progéniteur Myéloïde Commun (CMP) (Fig. 1.5A). D'une part, la différenciation du CLP génère les progéniteurs monopotents des lymphocytes T et des lymphocytes B (Pro-T et Pro-B). D'autre part, le CMP génère deux types de progéniteurs bipotents : le progéniteur des granulocytes-macrophages (GMP) et le progéniteur érythrocytaire et mégacaryocytaire (MEP). Le progéniteur des granulocytes-macrophages génère ensuite des macrophages et des granulocytes tandis que le progéniteur érythrocytaire et mégacaryocytaire génère des lignées d'érythrocytes ou de plaquettes.

L'introduction d'autres marques de surface a, par la suite, suggéré plusieurs modifications de cet arbre classique, dont les destins lymphoïdes et myéloïdes qui restent associés plus longtemps ainsi que la subdivision du compartiment progéniteur multipotent en sous-population distinctes (Fig. 1.5B).

Les nouvelles technologies (single-cell RNA-seq, analyse de la structure de la chromatine par ATAC-seq [4] et ChIP-seq [5] des marques d'histones liées à la transcription) remettent en question la vision classique de la hiérarchie hématopoïétique comme une structure très compartimentée et stable. L'image qui se dessine est celle d'un ensemble de populations hétérogènes organisées de manière hiérarchique (Fig. 1.5C), avec une progression de l'une à l'autre, qui reste très flexible pour répondre aux besoins changeants de la demande de sang. Par exemple, de récentes observations démontrent une hétérogénéité et un engagement de lignée inexpliqué dès les progéniteurs hématopoïétiques et donc l'existence de progéniteurs multipotents capables de donner naissance à des CMP comme à des CLP (Fig. 1.5B) [24]. Des expériences de single-cell RNA-seq ont été réalisées pour mettre en évidence les caractéristiques spécifiques des progéniteurs hématopoïétiques ainsi que leur devenir [25, 26]. Des études de profilage de la chromatine sur l'ensemble du génome ont démontré de larges différences dans les modifications

4. **ATAC-seq** : **A**ssay for **T**ransposase-**A**ccessible **C**hromatin with highthroughput **seq**uencing
5. **ChIP-seq** : **C**hromatin **I**mmuno**P**recipitation followed by highthroughput **seq**uencing - expérience détaillée dans la section 1.3.4

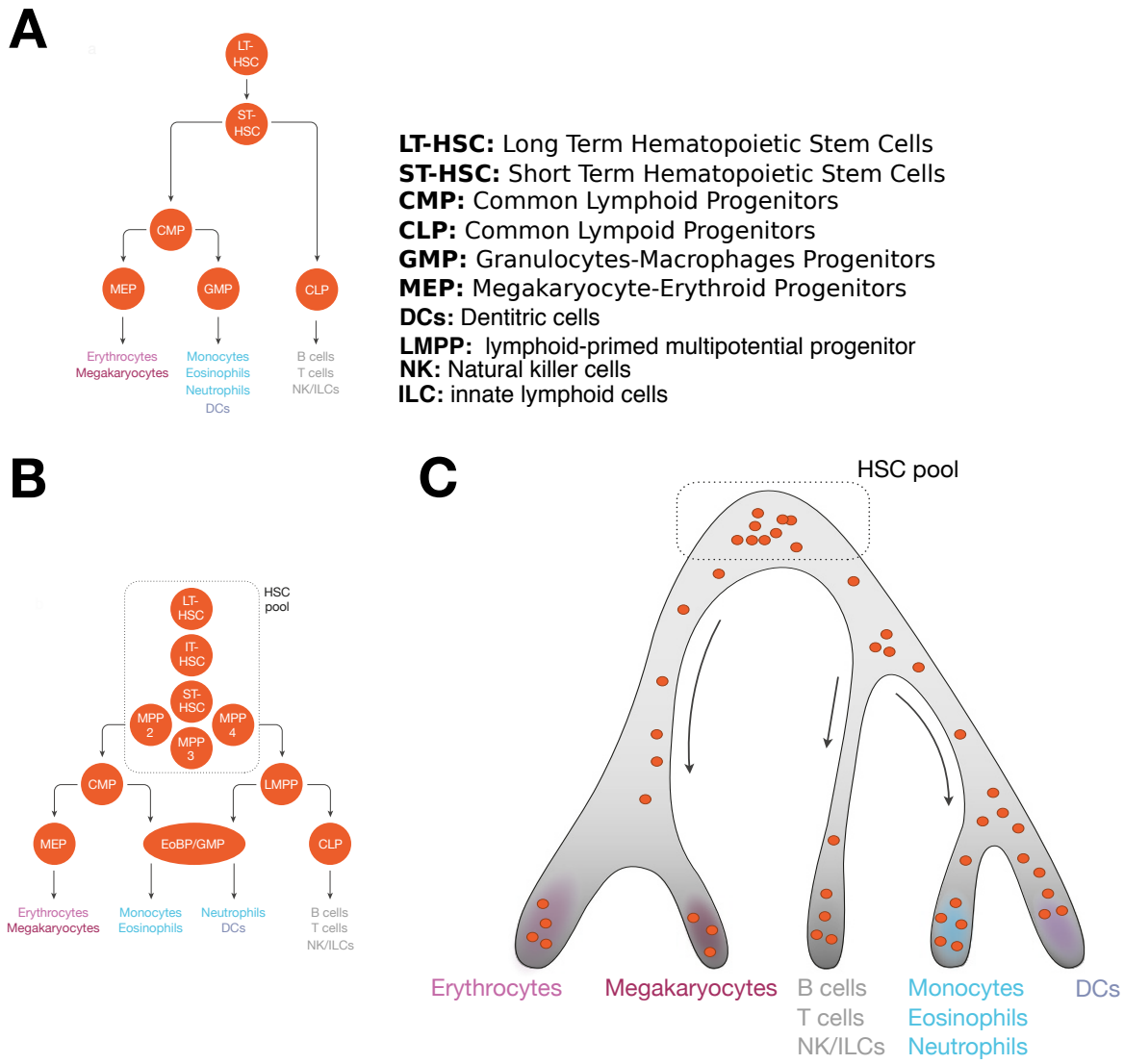


FIGURE 1.5 – **Les différents arbres d'hématopoïèse au cours du temps - figure et légende extraits de [23]. A.** Années 2000 : Les cellules souches hématopoïètiques (CSH) sont représentées comme une population homogène, en aval de laquelle la première bifurcation de lignée sépare les branches myéloïdes et lymphoïdes via les populations de progéniteurs myéloïdes communs (CLP) et de progéniteurs lymphoïdes communs (CLP). **B.** Années 2005-2015 : Intégration de nouveaux résultats, le groupe CSH est maintenant plus hétérogène à la fois en terme de propriétés d'auto-renouvellement (axe vertical) et de différenciation (axe horizontal), les branches myeloïdes et lymphoïdes restent associées plus bas via la population de LMPP qui est assez hétérogène. **C.** En 2016, les techniques de cellules uniques (single-cell) indiquent un continuum de différenciation. Chaque point rouge représente une cellule unique et sa localisation le long d'une trajectoire de différenciation.

des histones et au niveau des sites de liaison des facteurs de transcription dans les différentes cellules du sang matures. Par exemple, Lara-Astiaso et al [27] ont étudié quatre marques d'histone différentes (H3K4me1, H3K4me3, H3K4me2, H3K27ac) et l'expression génique associée

dans seize différents stades de développement hématopoïétique. Ils ont démontré que certains enhancers sont prédéterminés dans les CSH, tandis que d'autres sont établis de novo pendant l'hématopoïèse. Le même type de travail a permis de définir le répertoire des enhancers lors des modifications transcriptomiques allant des cellules souches/progénitrices hématopoïétiques humaines aux progéniteurs éryhtroïdes engagés impliquant les protéines GATA1 et 2 [28]. En particulier, il a été montré que le facteur de transcription SPI1, membre de la famille **ETS** (**E**-twentysix **T**ransformation **S**pecific), qui sera l'objet d'étude de ce travail, contrôlait la dynamique du répertoire des enhancers dans les macrophages.

1.2.2 La famille ETS

La famille ETS est l'une des plus grandes familles de facteurs de transcription [29]. Tous les membres de cette famille ont un domaine de liaison à l'ADN très conservé, ce domaine est appelé domaine ETS (Fig. 1.6). Il contient 85 acides aminés ayant une structure en hélice-coude-hélice. Ce domaine reconnait des séquences d'ADN qui ont en commun le coeur des séquences 5'-GGAA/T-3'. La famille ETS est séparée en 4 sous-classes (I, II, III, IV) (Fig. 1.6) selon la proximité phylogénique, structurale et la spécificité de la reconnaissance des bases entourant le coeur 5'-GGAA/T-3'. Autour du noyau commun 5'-GGAA/T-3', on trouve les régions flanquantes dans lesquelles une grande variabilité de séquence apparait [29]. Il y a 27 et 28 parologues différents dans la famille ETS chez l'humain et la souris, respectivement. Ils peuvent agir comme des activateurs ou des répresseurs de l'expression des gènes impliqués dans diverses fonctions telles que : la différenciation cellulaire, le contrôle du cycle cellulaire, la migration cellulaire, l'apoptose et l'angiogenèse. Chez les vertébrés, de nombreuses protéines à domaine ETS régulent l'hématopoïèse embryonnaire et adulte. Les protéines SPI1, SPIB et SPIC constituent la classe III (Fig. 1.6), la plus éloignée de l'ensemble des protéines ETS selon tous ces critères. Cette classe présente les séquences consensus les plus divergentes en amont du 5'-GGAA-3' comparée aux trois autres classes, et ne comporte que deux autres facteurs : SPI-B et SPI-C. SPI1 présente 43% d'identité de séquence (chez l'humain) avec ces derniers.

1.2.3 SPI1/PU.1 : structure, expression, régulation dans l'hématopoïèse

Structure de la protéine

Le gène *Spi1* est localisé sur le chromosome 2, comporte 5 exons pour une protéine de 272 acides aminés et d'un poids moléculaire de 35 kDa. Elle a trois domaines fonctionnels (Fig. 1.7) : un domaine de liaison à l'ADN en C-terminal (domaine ETS), un domaine de transactivation en N-terminal et un domaine PEST. Les domaines de transactivation et PEST sont respectivement impliqués dans l'activation transcriptionnelle et dans les interactions protéine-protéine. Via son domaine PEST, SPI1 peut interagir (non exhaustif) avec des protéines telles que la protéine du rétinoblastome [31], le facteur de transcription TFIID [31] et le facteur NF-EM5 [32]

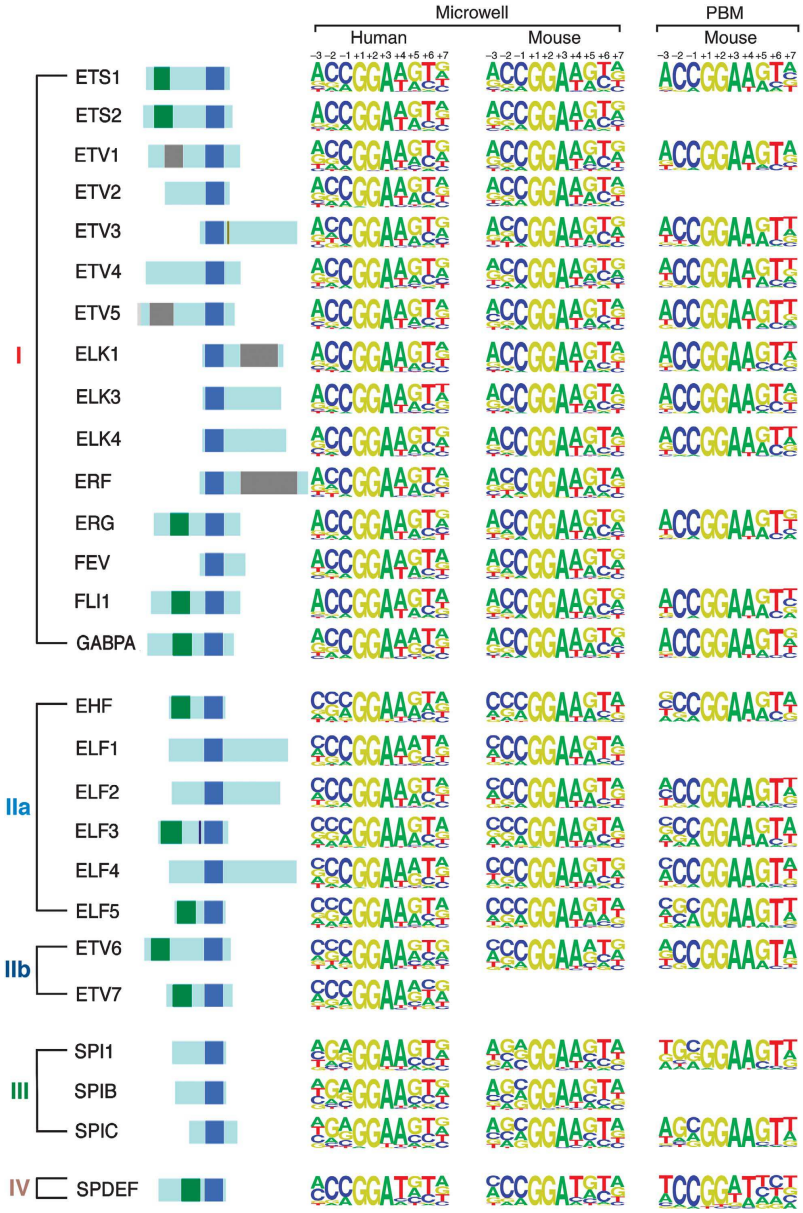


FIGURE 1.6 – **Représentation schématique des quatre sous-classes de facteurs de transcription de la famille ETS.** *Issue de [29]* Pour chaque protéine sont représentés à gauche l'agencement des domaines protéiques avec en bleu le domaine ETS, en vert le domaine pointé (PNT), en gris le domaine riche en proline et en noir et jaune doré les domaines riches en A/T et Nuc_crp_HMR_rcpt. Les motifs de liaison des différentes protéines ont été déterminés pour l'humain et la souris (au centre) à l'aide de la technique micro-well qui consiste en des tests de liaison du facteur de transcription à l'ADN basés sur des micro-puits. Les motifs de liaison des facteurs de transcription ont été également déterminés pour la souris uniquement (à droite) à l'aide de la technique Protein Binding Microarrays (PBM). La hauteur de chaque lettre à chaque position est proportionnelle à l'affinité de liaison de la protéine à la région. Les logos ont été dessinés en utilisant enoLOGOS [30].

(aussi appelé IRF4 ou Pip). Son domaine ETS se situe entre les acides aminés 165 à 272 (Fig. 1.7) et comporte trois hélices alpha et 4 feuillets beta antiparallèles, ce domaine lui permet de reconnaître des séquences comportant le motif minimal GGAA. La délétion de la région comportant les acides aminés 200 à 272 empêche la protéine SPI1 de se fixer à l'ADN. Les résidus sérine 41, 45, 132, 133 et 148 peuvent être phosphorylés. Plusieurs kinases sont impliquées dans la phosphorylation de SPI1 : la Caséine Kinase II [33], AKT [34], p38^{MAPK} [35] et la NF-kB-inducing kinase [36]. La phosphorylation peut augmenter l'activation de la transcription par SPI1, et moduler son activité d'interaction avec ses co-facteurs.

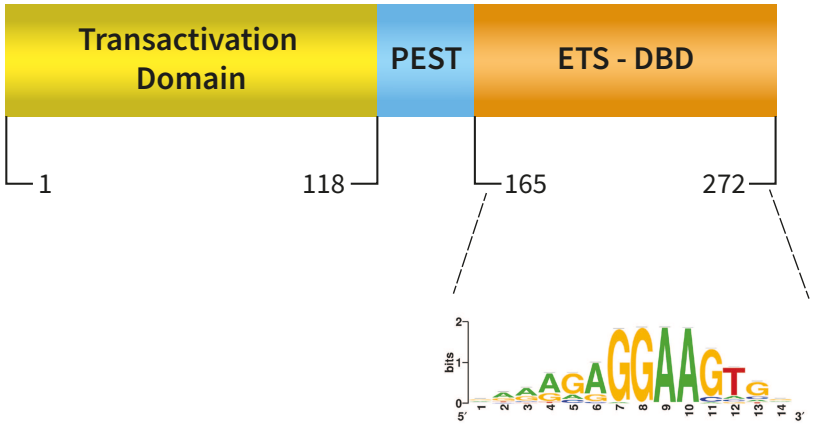


FIGURE 1.7 – **Structure de la protéine SPI1.** SPI1 est constitué de trois domaines fonctionnels : le domaine ETS de liaison à l'ADN en C-terminal ainsi que le motif de liaison de SPI1 à l'ADN [37], le domaine PEST dans la région centrale et le domaine transactivateur dans la partie N-terminale de la protéine. Le motif de bases reconnu par SPI1 sur l'ADN est indiqué.

Expression, rôle et régulation du gène *Spi1*

Régulation de l'expression de *Spi1* L'expression du gène *Spi1* est régulé par : (i) son promoteur proximal qui contient des sites de liaison pour OCT-1, SP1, GATA1 et SPI1 lui même ; (ii) un élément régulateur en amont du gène avec un site d'hypersensibilité à la DNAse I situé à 14kB en amont du TSS chez les souris où SPI1 lui-même, ELF1, FLI1, RUNX1, C/EBP et SATB1 se lient [38, 39]. La délétion de cet élément induit une diminution de 80% de l'expression de *Spi1* dans les cellules CSH, myéloïdes et lymphoïdes B et est associée au développement de leucémies myéloïdes [40,41]. En revanche, cet élément régulateur joue un rôle répresseur sur l'expression de *Spi1* dans la lignée des lymphocytes T. En effet, la suppression de cette région entraîne une augmentation de l'expression de *Spi1*, bloque la différenciation des lymphocytes T et induit l'apparition de lymphomes T [41]. Une région régulatrice à -12kB en amont du TSS a également été mise en évidence. Celle-ci fonctionnerait en synergie avec la région à -14kB afin de maintenir un niveau élevé de SPI1 dans les cellules myéloïdes [42]. Certains ARNs

non codants anti-sens transcrits à partir d'un promoteur intronique sous le contrôle de la région régulatrice à -14kB du TSS peuvent moduler la traduction de l'ARNm de *Spi1* [43].

Expression de *Spi1* dans les lignées hématopoïétiques L'expression de *Spi1* est finement régulée à travers les différents lignages hématopoïètiques. Il agit comme un déterminant transcriptionnel primaire du devenir des cellules hématopoïétiques, en particulier dans le contrôle du devenir de la population de CSHs et des progéniteurs associés. La délétion de *Spi1* chez la souris induit une létalité embryonnaire. Les embryons présentent des défauts de développement des lignages granulocytaire, lymphoïde et monocytaire [44]. En revanche, les lignages érythroïde et mégacaryocytaire ne sont pas affectés. Cela suggère un rôle crucial de SPI1 dans la lymphopoïèse et la myélopoïèse. A partir de modèles d'expression de *spi1* (souris SPI1/GFP) dans les différents lignages, il a été montré que l'expression de *Spi1* est dynamiquement régulée dans les différents lignages (Fig. 1.5B). *Spi1* est fortement exprimé dans les CSH, dans les progéniteurs CLP et GMP, et son expression augmente au cours de la différenciation myéloïde. Une augmentation de l'expression de *spi1* favorise la différenciation myéloïde tandis qu'un niveau plus faible de *Spi1* est nécessaire à la différenciation des lymphoïdes B [45, 46]. Ces modèles ont également montré la réduction d'expression de *Spi1* dans la phase terminale de la différenciation lymphocytaire T et érythroïde.

Rôle de *Spi-1* dans les lignées hématopoïétiques L'extinction inductible de *Spi1* chez les souris adultes (système inductible Cre-Lox) a permis de préciser le rôle de SPI1 dans les différents lignages. En particulier, SPI1 est nécéssaire à l'auto-renouvellement des CSHs [47], à la différenciation des CSHs en GMP et CLP, à l'engagement des cellules dans la différenciation lymphocytaire, à la différenciation en macrophages et granulocytes et à la rectriction de la granulopoïèse [48, 49, 49]. De plus, des souris ayant un ou deux allèles hypomorphes [6] de *Spi1* et exprimant respectivement 20% et 2% de la protéine SPI1 sauvage ont été générées. Cette étude a montré que SPI1 régule de manière dose-dépendante la transcription des gènes des différents lignages hématopoïétiques [50]. D'autre part, Back et al. ont montré qu'un faible niveau d'expression de *Spi1* est nécessaire au maintien du pool de progéniteurs érythroïdes immatures au cours de l'érythropoïèse foétale et au cours de l'érythropoïèse de stress chez l'adulte [46]. SPI1 est donc nécessaire :

- au renouvellement des cellules souches hématopoïètiques et à leur engagement vers les progéniteurs.
- à l'engagement des progéniteurs lymphoïdes B mais pas à leur maturation.
- à la différenciation myélomonocytaire.
- à la restriction de la granulopoïèse.

6. allèle hypomorphe : allèle dont la modification ne change pas la fonction du produit par rapport au gène sauvage. Le gène peut toutefois être moins exprimé ou son produit peut être moins actif.

— au renouvellement des progéniteurs érythroïdes immatures du foie fetal et de l'érythropoïèse de stress.

1.2.4 SPI1 : oncogène et oncosuppresseur dans les lignées hématopoïétiques

Les leucémies sont des cancers du tissu hématopoïétique caractérisées par l'accumulation et la prolifération incontrôlée de cellules souches ou progénitrices immatures. Le séquençage du gène *Spi1* chez des patients atteints de leucémie myéloïde aïgue (LAM) a révélé une fréquence de mutation faible (7% des patients) [51]. Des mutations inactivatrices du gène *Spi1* ont également été décrites dans la LAM humaine par fusion [52]. Néanmoins, la mutation du gène *Spi1* dans les LAM est un évènement rare. En revanche, d'autres études ont mis en évidence une diminution de son expression ou de son activité dans les LAM qui expriment des protéines de fusion telles que AML1-ETO [53], FLT3-ITD [54], PML-RARα [55] et dans les leucémies avec la mutation de NPM1c [56]. Par exemple, la protéine de fusion AML1-ETO se fixe à SPI1 et entraîne une inhibition de son activité transactivatrice [53]. Dans un type particulier de LAM, un SNP (Single-Nucleotide Polymorphism) localisé dans l'enhancer de SPI1 (à -14kbp du TSS) a été mis en évidence, ce SNP diminuerait la transcription du gène *Spi-1* [39]. Ces études suggèrent donc que la diminution de l'expression ou la diminution de l'activité de SPI1 est associée au développement des LAM, suggèrant que SPI1 joue un rôle suppresseur de tumeur dans le lignage myéloïde. Dans les modèle murins, la diminution de l'expression ou de l'activité transactivatrice de SPI1 entraîne également le développement de LAM. La délétion de l'enhancer de *Spi1* entraîne une diminution de 80% de son expression et ainsi le développement de LAM [40]. L'expression de *Spi1* ne peut pas être nulle, en effet, SPI1 joue un rôle anti-apoptotique permettant aux cellules de survivre. En particulier, Antony-Debré et al [57] suggèrent qu'il est maintenant clair que dans ce type de leucémie myéloïde, une activité résiduelle du SPI1 est néanmoins nécessaire pour que les cellules leucémiques puissent survivre et proliférer, ainsi l'inhibition de *Spi1* pourrait être une stratégie thérapeutique potentielle pour le traitement des LAM.

En plus de son rôle onco-suppresseur dans la lignée myéloïde, l'augmentation de l'expression ou de l'activité de SPI1 participe au développement leucémique. Son rôle d'oncogène s'établit à la suite d'un gain d'expression, dans la macroglobulinémie de Waldenström, une mutation somatique modifiant la spécificité des bases fixées par SPI1 sur l'ADN. En effet, outre les séquences qu'il reconnaît habituellement 5'AGA-GGAA-GTA-3', il reconnaît également celles spécifiques d'autres protéines de la famille ETS, telle que ETS1 5'ACC-GGAA/T-xxx-3' (Fig. 1.6). La mutation entraîne également une augmentation de sa capacité de transactivation. Cette liaison accrue de SPI1 à l'ADN, induit une augmentation de la prolifération cellulaire et une diminution de la différentiation lymphoïde B terminale [58]. Un autre rôle oncogénique de SPI1

est décrit dans la leucémie lymphoblastique aiguë pédiatrique des cellules T (T-ALL). SPI1 est impliqué dans plusieurs fusions de gènes (SMTN1-SPI1, TCF7-SPI1). Les protéines résultantes de ces fusions sont exprimées à un stade de différenciation dans la lignée lymphoïde T où l'expression de SPI1 devrait être éteinte, et induisent le blocage de leur maturation [59]. Enfin le troisième exemple de gain d'expression de SPI1 entraînant un rôle oncogénique a été décrit dans la lignée érythroïde qui est associé à une érythroleucémie murine (voir section 1.2.5). C'est sur ce type d'érythroleucémie que j'ai développé mes travaux de thèse.

1.2.5 SPI1 et érythroleucémie

Le virus de Friend est une souche du virus de la leucémie murine qui a été isolé par Charlotte Friend en 1957. Celui-ci est capable de s'intégrer de façon aléatoire dans les cellules qu'il infecte, et d'induire une leucémie en deux étapes. Le virus de Friend est un complexe composé de deux particules virales : le Spleen Focus Forming Virus (SFFV) qui est l'élément pathogène du virus de Friend, déficient pour la réplication et le Friend-Murine Leukemia Virus (F-MuLV) permettant la réplication [60]. L'érythroleucémie développée par les souris infectées évolue en deux étapes successives. L'étape précoce du développement de l'érythroleucémie est caractérisée par une polycythémie et une hépatosplénomégalie dûes à la prolifération incontrôlée de progéniteurs érythrocytaires dans le sang, la rate puis le foie des souris infectées. Ces cellules sont toujours capables de se différencier en globules rouges et prolifèrent indépendamment de leur facteur de croissance, l'érythropoïétine (Epo). L'indépendance à l'Epo est due à l'activation constitutive du récepteur à l'Epo par la glycoprotéine gp55 codée par le génome viral du SFFV. L'étape tardive est caractérisée par l'apparition d'une population clonale de cellules proérythroblastiques tumorigènes, dites cellules tumorales de Friend. La différentiation de ces cellules est bloquée [61] mais elles ont une capacité illimitée de renouvellement in-vitro (cellules appelées MEL pour **M**urine **E**rythro**L**eukemia). Ces cellules leucémiques présentent deux altérations génétiques [62] :

— inactivation du gène suppresseur de tumeur *p53*.
— mutagénèse insertionnelle par le virus SFFV du gène *Spi1* : c'est le clonage du site d'intégration du virus qui a permis d'identifier ce gène *Spi1* pour **S**pleen **P**rovirus **I**ntegration site 1.

Les séquences activatrices de la transcription présentes dans les Long Terminal Repeat (LTR) du SFFV induisent l'activation transcriptionnelle de *Spi1* et conduisent à la surexpression d'un ARNm de 1,4kb qui code pour la protéine normale de SPI1 [63]. Le traitement de ces cellules MEL avec des agents chimiques tels que le DiMethylSulFOxyde (DMSO) induit la différenciation de ces cellules en globules rouges anormaux [64].

Le modèle des souris transgéniques

Pour comprendre le rôle de SPI1 dans la transformation maligne, des souris transgéniques qui surexpriment le gène *Spi1* dans toutes les cellules de la souris, ont été générées par l'équipe de F. Moreau-Gachelin [65]. Ces souris développent à l'âge de 3 mois une érythroleucémie avec une incidence d'environ 50%. La maladie évolue en deux étapes : une première étape dite "préleucémique" et une seconde étape dite "leucémique" (Fig. 1.8). La première étape se caractérise par une anémie et une hépatosplénomégalie due à la prolifération aiguë des cellules blastiques qui envahissent le sang, la rate et le foie des souris malades (étape dite HS1 pour HépatoSplénomégalie 1 - Fig. 1.8). Ces cellules sont des progéniteurs érythroïdes appelés CFU-E ou des proérythroblastes [66]. Leur différenciation est bloquée [67], lorsque des lignées cellulaires issues de la rate des souris malades sont établies *in vitro* elles sont dépendantes de l'érythropoïétine (Epo) pour leur prolifération et leur survie. La greffe de ces cellules dans la souris nude n'entraîne pas l'apparition de tumeur, elles sont alors définies comme **cellules pré-leucémiques**. Ces cellules qui prolifèrent dans la rate sont donc bloquées dans leur différenciation terminale [65]. La perte de l'expression de SPI1 dans ces cellules entraîne la réactivation de la différenciation terminale en globule rouge, démontrant que l'expression anormale de SPI1 dans la lignée érythroïde suffit à bloquer la différenciation [67]. Ces résultats, associés à ceux montrant que l'inhibition de SPI1 entraine également la mort d'une partie des cellules, montrent que SPI1 joue un rôle majeur dans le développement de l'érythroleucémie, en bloquant la différenciation et l'apoptose [67]. L'hépatosplénomégalie et l'anémie entraine la mort des animaux. Des transfusions répétées de globules rouges (Fig. 1.8) peuvent provoquer une rémission transitoire de la maladie. En effet ils apportent un signal d'arrêt de la production d'Epo, induisant ainsi un arrêt de la prolifération des cellules. L'hépatosplénomégalie régresse, cela démontre la dépendance des cellules HS1 à l'Epo *in vivo*. Toutefois, malgré les transfusions, après 3 semaines, une seconde hépatosplénomégalie (HS2) apparaît. Les cellules érythroïdes sont toujours bloquées dans leur différenciation (cellules HS2 - Fig. 1.8). Au contraire des cellules HS1, les cellules HS2 peuvent être établies en lignée *in vitro* en absence d'Epo ; elles sont indépendantes de l'Epo. Dans la souris nude, la greffe de ces cellules entraîne des tumeurs, elles sont donc définies comme **cellules leucémiques**. La surexpression du facteur de transcription SPI1 est responsable du blocage de la différenciation érythroïde, toutefois il ne suffit pas au développement de la leucémie, puisque les cellules HS1 sont dans un état pré-leucémique. Il a été montré que ces cellules ont acquis des mutations activatrices du gène Kit, récepteur au Stem Cell Factor (SCF), qui induisent une activation constitutive de la voie de signalisation en aval du récepteur, sans SCF et sans EPO [68, 69]. Des mutations inactivatrices de *Trp53* ont également été identifiées dans 50% des souris.

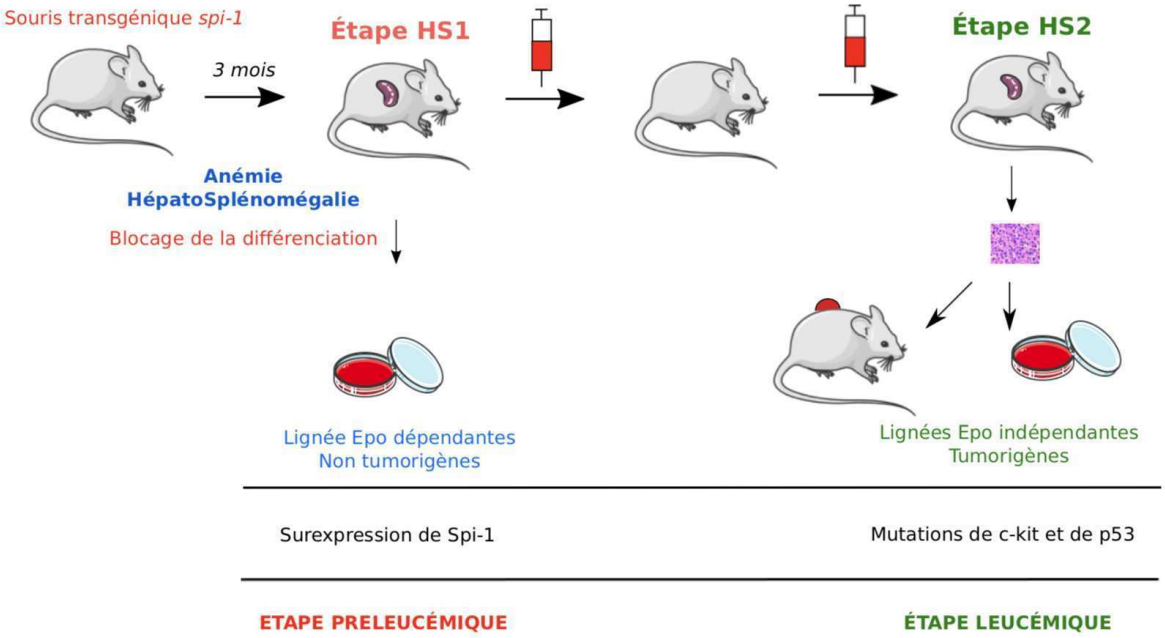


FIGURE 1.8 – **Processus érythroleucémique en deux étapes développé par les souris transgéniques *spi-1*.** L'étape 1 est caractérisée par la prolifération dépendante de l'Epo des progéniteurs érythroïdes qui ne se différencient pas en globules rouges. Au cours de l'étape 2, les cellules sont indépendantes de l'Epo et tumorales. Elles sont mutées dans le gène *kit* à presque 100% et dans le gène *p53* pour environ 50%.

1.2.6 SPI1 : un régulateur transcriptionnel

SPI1 est un facteur de transcription qui n'agit pas seul, en effet, il interagit avec des facteurs épigénétiques, des facteurs d'épissage et des co-facteurs. De plus il peut avoir une activité pionnière qui pourrait créer un relâchement de la chromatine afin de recruter d'autres facteurs de nature épigénétique ou d'autres facteurs de transcription.

SPI1 : un facteur pionnier

Les nucléosomes modulent les interactions des facteurs de transcription avec l'ADN, ainsi leur disposition est cruciale [70]. En effet, comme nous l'avons exposé dans la section 1.1.3, l'arrangement des nucléosomes c'est à dire l'état de la chromatine d'un gène et de ses régions régulatrices va influencer le niveau de transcription de celui-ci.

Il a été montré que dans les macrophages, l'expression de *spi1* est nécessaire pour produire et maintenir la marque H3K4me1 d'identité des enhancers [71, 72].

Dans une étude réalisée en 2014 [73], les chercheurs ont fait l'hypothèse que l'information reconnue par les facteurs de transcription pour se fixer devrait présenter des propriétés particulières, contrôlant l'incorporation de ces séquences dans les nucléosomes. Pour tester cette hypothèse, ils ont utilisé des macrophages de souris dans lesquels SPI1 se fixait virtuellement

sur toutes les régions portant la marque H3K4me1 et sur une grande fraction des régions TSSs. Ils ont montré que des NDR sont observées autour des sites fixés par SPI1. De plus, ces sites occupés par SPI1 sont protégés par des nucléosomes dans les cellules qui n'expriment pas le gène *spi1*. Ainsi, les régions sans nucléosomes autour des sites fixés par SPI1 semblent être dûes à la présence de SPI1. Les nucléosomes qui couvrent ces sites en absence de SPI1 présentent un large spectre d'occupation et de positionnement et suggèrent ainsi une complexité élevée dans l'interaction entre le facteur de transcription et l'occupation nucléosomiale.

Fernandez-Garcia et al [74] ont montré que les facteurs pionniers, dont SPI1-1 fait partie, reconnaissent l'ADN via leur domaine de liaison à l'ADN (scissor-like ou hélice-coude-hélice) qui leur permet de se fixer aux nucléosomes, tandis que les facteurs de transcription non-pionniers interagissent avec l'ADN via un enchaînement de structures secondaires en hélices ou boucles.

Néanmoins, cette activité a été remise en question récemment. En effet, Minderjahn et al [75] ont analysé la capacité de SPI1 à accéder à ses sites de liaison *in vivo* et *in vitro*. L'étude a montré que SPI1 ne fixe pas les CpG méthylés ni l'ADN enroulé autour des nucléosomes *in vitro* mais simplement des régions à +10bp des nucléosomes contrairement à ce qui a été montré par Fernandez et al [74]. Cela suggèrerait que SPI1 n'agit pas comme un facteur pionnier classique. De plus, l'analyse fonctionnelle de mutants pour les différents domaines de SPI1 indique que la capacité de SPI1 à accéder aux sites de la chromatine *de novo* remodelés repose sur son domaine d'activation N-terminal. L'introduction de SPI1 dans des lignées cellulaires pour lesquelles l'expression de SPI1 est manquante induit un remodelage de la chromatine aux sites de liaison de SPI1 et des conséquences immédiates sur le programme de différenciation cellulaire [75]. Une publication récente [76] suggère que SPI1 interagit avec les complexes de la famille SWI/SNF via son domaine N-terminal. Cette interaction permettrait la redistribution des co-facteurs de transcription recrutés par SPI1 et le remodelage de la chromatine.

SPI1 et ses partenaires épigénétiques

SPI1 est capable d'initier des changements dans la chromatine en recrutant des facteurs épigénétiques. Comme nous l'avons vu dans la section précédente, la position des nucléosomes autour de SPI1, et donc l'organisation de la chromatine dans les régions fixées par SPI1, est dûe à la présence de SPI1. Les protéines histone acétyltransférases, deacétylases, méthyltransférases, et la méthylcytosine dioxygénase sont des exemples de facteurs épigénétiques qui peuvent être recrutés pour modifier l'organisation de la chromatine. Dans cette partie, nous allons détailler les différents modificateurs épigénétiques qui interagissent avec SPI1.

Méthylation de l'ADN Largement étudiée depuis les années 1980, la méthylation de l'ADN chez les eucaryotes cible les cytosines, bases pyrimidiques constitutives de l'ADN, qui sont

alors converties en 5-méthylcytosines (5-mC) [77, 78]. Dans la plupart des cas, les cytosines méthylées sont adjacentes à une guanine et forment ensemble un couple de dinucléotides que l'on nomme CpG. Les régions du génome enrichies en CpG sont appelées îlots CpG. Elles sont principalement retrouvées au niveau des promoteurs ou du premier exon dans la majorité des gènes [79]. Les CpG ne sont donc pas distribués de façon homogène dans le génome. La présence d'îlots CpG au niveau des promoteurs est généralement associée à une répression transcriptionnelle car elle inhibe directement la liaison de certains facteurs de transcription et peut induire le recrutement de certains répresseurs [80].

La méthylation des cytosines est catalysée par les enzymes DNMT3A/B et DNMT1, leur déméthylation est quant à elle catalysée par l'enzyme TET2. Au cours de l'ostéoclastogénèse [7], des changements d'un état hypo-méthylé à un état hyper-méthylé sont observés pour de nombreux gènes, l'analyse d'enrichissement des motifs de facteurs de transcription autour des CpG hypo-/hyper-méthylées a révélé une surreprésentation des motifs de liaison de SPI1, NF-κB et AP-1 (Jun/Fos). Des expériences de ChIP-seq ont confirmé la présence de SPI1 sur ces sites [81]. Suzuki et al [82] ont montré que SPI1 interagissait avec ces deux DNA-methyltransferases dans les cellules B. L'interaction de SPI1 avec DNMT3B pourrait être un moyen de réprimer ses gènes cibles puisque la répression de ces derniers est stoppée par ajout de 5-aza-deoxycitidine, un inhibiteur des ADN-méthyltransférases. De plus, SPI1 interagit également avec TET2, une méthylcytosine dioxygénase. Le knock-down de SPI1 par un siRNA [8] dans les monocytes modifie la méthylation de l'ADN et réduit l'association TET2 et DNMT3B aux régions fixées par SPI1 durant la différenciation des ostéoclastes [81, 83].

Modifications de l'histone H3 L'histone H3 est l'une des cinq principales protéines histones impliquées dans la structure de la chromatine chez les eucaryotes. Les modifications (méthylation, acétylation) qui peuvent être apposées sur certains de ses résidus sont impliquées dans la dynamique et la régulation de l'expression des gènes. Les protéines du groupe PolyComb (PcG) sont une famille de complexes protéiques capables de remodeler la chromatine pour induire l'extinction des gènes [84]. Ce remodelage passe par la triméthylation de la lysine 27 de l'histone 3 (H3K27me3) catalysée par PolyComb Repressive Complex 2 (PRC2), l'une des deux protéines du groupe PcG. Le complexe PRC2 chez la souris possède quatre sous-unités : SUZ12, EED, EZH2, EZH1, et a un rôle dans l'inactivation du chromosome X. Une expression aberrante de EZH2 est un marqueur de maladie avancée et métastatique dans de nombreuses tumeurs solides, par exemple le cancer de la prostate et le cancer du sein [85]. Le groupe de C. Guillouf [66] a montré que SPI1 interagit avec SUZ12 et EZH2 afin de moduler l'activité de PRC2 et est capable d'augmenter le niveau de la marque H3K27me3 au promoteur d'un gène

7. Ostéoclastogénèse : Différenciation des ostéoclastes (un type de cellules osseuses) à partir des macrophages.

8. siRNA : ARN interférent pouvant se lier spécifiquement à un ARNmessager et le cliver pour empêcher l'expression de la protéine associée.

qu'il réprime transcriptonnellement. Cette interaction entre SPI1 et PRC2 joue un rôle majeur dans la répression de la transcription du gène *Bcl2l11*, un gène de l'apoptose.

HDAC1 est une enzyme qui catalyse le retrait d'un groupement acétyle sur un acide aminé lysine d'une histone, elle est en général en complexe avec mSin3A. Il a été montré que SPI1 peut former un complexe avec mSin3A et HDAC1 *in vivo* [86]. La formation de ce complexe se fait via le domaine C-terminal de SPI1 (acides aminés 101 à 272). De plus, SPI1 interagit avec la protéine MeCP2, une protéine capable de se fixer sur des régions d'ADN méthylé et de réprimer la transcription du ou des gènes associés à cette méthylation. L'interaction entre les deux protéines se fait via le domaine ETS de SPI1. La protéine MeCP2 est intégrée au complexe SPI1-mSin3A-HDAC1 [87].

La famille de co-activateurs CBP/p300 est composée de deux protéines co-activatrices : p300 et CBP. P300 et CBP interagissent avec un grand nombre de facteurs de transcription dans le but d'activer l'expression de leurs gènes cibles [88]. Pour cela, elles établissent le lien physique entre les activateurs transcriptionnels spécifiques à une séquence et les facteurs de transcription généraux qui se fixent près du TSS. Parmi les multiples fonctions de ce complexe, les deux protéines possèdent un domaine acétyltransférase qui leur confère la possibilité d'acétyler les protéines histones afin d'augmenter le niveau de transcription de leurs gènes cibles. Yamamoto et al. [89] ont démontré que CBP/p300 agit comme un co-activateur de SPI1 à travers une interaction par le domaine de transactivation de SPI1 (acides aminées 74 à 122). De plus, l'activité acétyltransférase de CBP augmente au cours de la différentiation des érythrocytes à mesure que le niveau de SPI1 diminue. Dans le cas particulier de la surexpression de SPI1 dans les pro-érythroblastes, SPI1 va réduire l'activité de CBP [90]. En effet, au cours de la différenciation érythrocytaire normale, la liaison de CBP à GATA1 entraîne son acétylation et son activation transcriptionnelle, il a été montré que la mutation des sites d'acétylation de GATA1 altère sa capacité à induire la différenciation érythroïde [91]. Dans les MEL, SPI1 en se fixant sur GATA1 lui même fixé à la chromatine, empêcherait l'acétylation de GATA1. SPI1 inhiberait ainsi l'activité transcriptionnelle de GATA1 et la transcription dépendante de l'acétylation, telle que la transcription du gène de la *globine*, nécessaire à la différenciation érythroïde [90].

SPI1 interagit avec SUV39H [92] qui est responsable de la triméthylation de la lysine 9 de l'histone H3 qui est une marque que l'on trouve majoritairement dans l'hétérochromatine. Cette interaction est médiée par la protéine du rétinoblastome (protéine RB) qui interagit avec SPI1 en particulier pour réprimer l'activité du facteur de transcription GATA1 [93]. En effet, SPI1 se fixe sur GATA1 par son domaine ETS, GATA1 est lui fixé sur l'ADN et SPI1 en coopération avec la protéine du RB par son domaine N-terminal réprimé l'activité de GATA1. La protéine RB co-localise avec SPI1 et GATA1 au niveau des gènes cibles de GATA1 qui sont alors réprimés. La liaison de SPI1 à GATA1 sur les gènes cibles de GATA1, où GATA1 est déjà fixé, entraîne

le recrutement de la protéine RB, de l'histone méthyltransférase SUV39H et de la protéine HP1α [92]. SUV39H est responsable de la méthylation de la lysine-9 sur les histones H3, cette marque épigénétique est une marque d'hétérochromatine. La formation du complexe répresseur SPI1-GATA1-SUV39H-HP1α-RB est suivie par une triméthylation locale de la lysine 9 de l'histone H3. La réduction de l'expression de SPI1 entraîne la dissociation du complexe de répression. La triméthylation de la lysine 9 de l'histone H3 est alors remplacée par l'acétylation de cette même lysine par le recrutement de p300/CBP.

Activité transcriptionnelle de SPI1 spécifique du lignage hématopoïétique

SPI1 en tant qu'activateur transcriptionnel La fonction de SPI1 en tant que régulateur transcriptionnel a essentiellement été décrite dans un contexte d'activation transcriptionnelle. L'activation de la transcription par SPI1 dépend de son association avec d'autres facteurs de transcription au sein de complexes multi-protéiques. Ces associations mettent en jeu divers mécanismes :

- les facteurs interagissent pour renforcer mutuellement leur lien à l'ADN.
- les facteurs n'interagissent pas directement mais leur liaison sur le promoteur permet une synergie de l'activation transcriptionnelle.
- Les facteurs se fixent à SPI1, lui même lié à l'ADN et activent ainsi son activité transcriptionnelle.

SPI1 interagit avec une grande variété de facteurs de transcription. Ces facteurs ont un profil d'expression ubiquitaire comme SP1 ou tissu spécifique comme OCT2 dans les cellules lymphoïdes B ou C/EBP dans les cellules myéloïdes. SP1 et SPI1 coopèrent pour activer la transcription de gènes myéloïdes tels que *Fes* [94] ou *Cd11b* [95]. L'interaction physique entre ces deux protéines n'a pas été démontrée. SPI1 et la protéine C/EBP interagissent physiquement et permettent la transcription de gènes dont ceux des récepteurs au G-CSF [96], M-CSF et GM-CSF [97]. Les protéines SPI1 et Pip participent à l'activation transcriptionnelle dans les cellules lymphoïdes B selon un mécanisme en deux étapes. L'interaction entre SPI1 et IRF4/8, active la transcription des chaines légères des immunoglobulines en se fixant sur un enhancer localisé en 3' du gène [98] et la transcription du marqueur lymphocytaire CD20 en se fixant sur son promoteur [99]. Le facteur de transcription c-JUN appartenant à la famille AP-1 est un co-activateur crucial de SPI1 pour de nombreux gènes myéloïdes notamment le récepteur au M-CSF. c-JUN ne se fixe pas directement au promoteur de M-CSF mais interagit via son domaine basique avec le domaine ETS de la protéine SPI1 [100]. Dans ce cas, l'activité transcriptionnelle de SPI1 est modulée par un co-facteur qui se fixe à SPI1 lui même lié à l'ADN.

Pham et al [101] ont montré que la liaison de SPI1 à l'ADN est sélective dans les monocytes et les macrophages. Ils ont, en effet, établi trois types de potentiels sites de liaison de SPI1 dans les macrophages : (i) des sites non fixés situés dans la chromatine inactive avec une affinité faible de SPI1 pour la région ; (ii) des sites de liaison inaccessibles à la DNase I où SPI1

a une forte affinité pour la région et se fixe de manière autonome correspondant à son activité dite "pionnier" ; (iii) des sites accessibles à la DNase, pour lesquels SPI1 a une moyenne/faible affinité, sa liaison à l'ADN est renforcée par la coopération avec des sites de liaisons voisins pour d'autres facteurs de transcription. L'augmentation de la concentration de SPI1 et la disponibilité de co-facteurs d'interactions spécifiques de lignages au cours de la différenciation induit une meilleure liaison de SPI1 à des sites spécifiques d'un type cellulaire.

Dans les macrophages, 80% des sites de liaison de SPI1 se trouvent dans des régions en dehors des promoteurs [71]. Ces régions sont préférentiellement des enhancers, elles sont caractérisées par une forte densité de la marque H3K4me1 et une faible densité de la marque promotrice H3K4me3. De plus, ces régions sont dépourvues de nucléosomes. Ces régions ne sont fixées par SPI1 ni dans les lymphocytes B ni dans les progéniteurs hématopoïétiques où SPI1 est exprimé [71]. Dans les macrophages, Heinz et al [72] ont montré que des facteurs de transcription inductibles par des stimulis externes (famille IRF, NF-kappaB) se fixent préférentiellement dans des régions génomiques pré-marquées et activées par SPI1, la liaison de ces facteurs inducibles ne se produit pas en absence de SPI1. En effet, la liaison de SPI1 générerait des régions de chromatine ouverte spécifiques au type de cellules. Ces régions serviraient de repères pour le recrutement de co-activateurs transcriptionnels en réponse aux stimuli. Cela suggère que la liaison de SPI1 est essentielle pour la fonctionnalité des enhancers dans les macrophages.

Les régions promotrices fixées par SPI1 dans les macrophages ne contiennent pas de marque canonique forte telle que la TATA-box, on y trouve des clusters de sites pour SPI1 [102]. De plus, ces régions présentent des niveaux d'occupation similaires aux régions promotrices fixées par SPI1 dans les lymphocytes B, ce qui n'est pas le cas pour les régions distales [72] (voir paragraphe précédent). En effet, SPI1 se fixe sur les promoteurs actifs des gènes non-spécifiques à un lignage, appelé housekeeping genes [103]. Yaneva et al [104] ont montré que SPI1 est essentiel pour recruter et positionner la machinerie transcriptionnelle, en particulier le facteur TFIID, en l'absence d'une TATA-box dans le promoteur.

Dans des cellules pré-leucémiques provenant de souris transgéniques qui sur-expriment *Spi1* ayant développé une hépatosplénomégalie, l'équipe dans laquelle j'ai effectué ma thèse a caractérisé comment SPI1 active les gènes en absence de co-facteurs érythroïdes connus [37]. Pour cela, l'expression de *Spi1* a été inhibée en utilisant un sh-RNA [9] inducible par la doxycycline. Ainsi les chercheurs ont pu caractériser les gènes activés et réprimés par SPI1 [37]. L'analyse des mécanismes d'activation des gènes montre que l'activation des gènes se fait principalement par une liaison de SPI1 sur les promoteurs à proximité des gènes qu'il active (Fig. 1.9). Elle ne fait pas intervenir de co-facteurs connus.

9. **shRNA** : **s**mall **h**airpin **RNA**

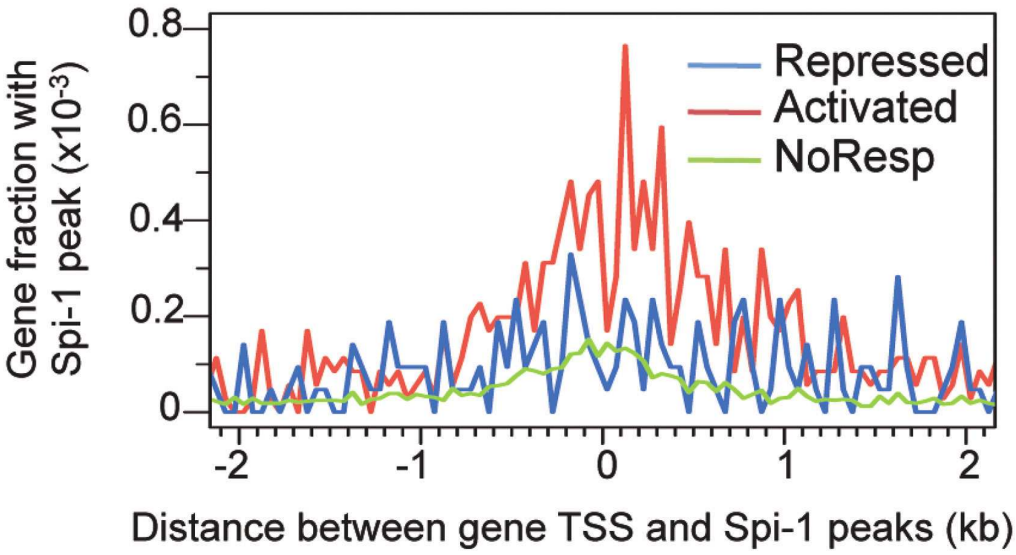


FIGURE 1.9 – **Distribution des pics de SPI1 autour du TSS du gène le plus proche.** *Issue de [37].* Les pics sont séparés en fonction de la régulation des gènes sur lesquels ils se trouvent. Les gènes activés par SPI1 présentent un enrichissement en pics de SPI1 autour du TSS. En revanche, les gènes réprimés par SPI1 ne présentent pas d'enrichissement particulier autour du TSS.

L'activation des gènes par SPI1 est sujette à différents paramètres. Elle est facilitée si SPI1 se fixe autour de TSS pauvre en îlots CpG. De plus, la multiplicité et l'orientation dans le même sens sur l'ADN des motifs de liaison de SPI1, au niveau des régions occupées, active d'autant plus la transcription du gène régulé. Ainsi, cette étude démontre que dans la lignée érythroïde, SPI1 se fixe majoritairement dans les promoteurs sans co-facteurs spécifiques de la lignée, et en particulier dans la région juste après le TSS, pour activer l'expression génique au contraire des cellules myéloïdes et lymphoïdes B où SPI1 se fixe dans des régions enhancers, accompagné de co-facteurs qui stabilisent sa liaison à l'ADN. De plus dans les cellules érythroïdes immatures, SPI1 régule de nombreux gènes des voies PI3K/Akt et ERK/MAPK qui constitutent les principales voies de contrôle de la croissance et de la survie des cellules érythroïdes immatures [105].

SPI1 en tant que répresseur transcriptionnel Dans les précurseurs des lymphocytes T (pro-T) SPI1 contrôle l'expression de nombreux gènes avant que les cellules n'entrent dans le thymus. SPI1 restreint l'expression de gènes et ralentit la progression vers l'engagement terminal permettant un timing correct de l'engagement dans la différenciation terminale des lymphocytes

T. Néanmoins, les sites de liaison de SPI1 se situent dans des régions de chromatine décondensée dans des gènes dont l'expression ne varie pas lorsque l'expression de *spi1* diminue. Au stade de différenciation suivant (pre-T), l'expression de *spi-1* diminue et ainsi de nombreuses régions de chromatine perdent l'accessibilité établie par SPI1 [106]. Hosokawa et al [107] ont montré que SPI1 réprime l'expression génique dans les précurseurs des lymphocytes T en modifiant les sites de liaison de deux de ses partenaires. En effet, il recrute ses partenaires (RUNX1, SATB1 qui sont deux facteurs de transcription indispensables à la différenciation terminale) sur ses propres sites de liaison, ainsi les sites initiaux de RUNX1 et SATB1 ne sont plus occupés, et les gènes régulés par ces régions sont réprimés. Ces gènes sont des gènes que SPI1 réprime malgré le fait qu'il ne se fixe pas directement dans leurs régions de régulation. Ainsi, dans ce modèle, SPI1 réprime des gènes en séquestrant des facteurs de transcription activateurs et en empêchant leur liaison sur leur gènes cibles.

Dans les neutrophiles, Fisher et al [108] ont démontré que la fonction prédominante de SPI1 est de restreindre la réponse immunitaire des neutrophiles en réponse à une infection en inhibant l'accessibilité aux enhancers par le recrutement de HDAC1. Cette fermeture de la chromatine empêche le facteur de transcription immuno-stimulé AP-1 de se fixer à la chromatine et d'activer ses cibles. Ainsi, SPI1 met en place un programme d'inhibition pour protéger l'épigénome des neutrophiles contre une activation incontrôlée ce qui protège l'hôte d'une réponse immunitaire inée exorbitante. En effet, les neutrophiles sont la première ligne de défense cellulaire contre les agents pathogènes envahissants et lorsqu'ils sont activés de façon inappropriée cela peut causer des dommages tissulaires collatéraux et contribuer aux maladies immunologiques.

SPI1 et sa liaison à l'ARN

Une recherche des interacteurs de SPI1 par GSC pull-down associé à un séquençage par masse spectrométrie a identifié que SPI1 interagit avec des protéines qui se fixent à l'ARN, essentiellement des protéines d'épissage dont P54 [109] une protéine similaire au facteur d'épissage PSF (polypyrimidine tract-binding protein-associated splicing factor), hnRNPA1 et TLS (Translocated in LipoSarcoma) [110], une protéine impliquée dans diverses translocations chromosomiques spécifiques de tumeurs humaines. Des expériences d'immunoprécipitation ont permis de montrer que ces complexes existaient *in cellulo* [109, 110]. La capacité de SPI1 à se lier à des protéines qui fixent l'ARN semble lui être spécifique car d'autres protéines de la famille ETS telles que ETS-2 et FLI-1 n'interagissent pas avec ces protéines [109, 110].

Les auteurs montrent en utilisant un minigène contenant les exons 1 et 2 de la β-globine, que la protéine SPI1 inhibe les fonctions de P54 dans l'épissage du transcrit [109]. En revanche, l'interaction SPI1/P54 ne modifie pas les fonctions de SPI1 dans la transcription. Il a également été démontré que la protéine TLS est capable d'empêcher la liaison de SPI1 à l'ADN sur ses cibles transcriptionnelles et que SPI1 et TLS modifient leur fonction respective dans l'épissage [110].

En ce qui concerne la fonction de SPI1 dans l'épissage, elle s'exerce de deux façons : une fonction dépendante de sa capacité à transactiver et à lier l'ADN et une fonction indépendante qui pourrait se faire par séquestration des facteurs d'épissage. Ces deux fonctions modifient l'épissage de façon opposée ; c'est à dire inclusion ou exclusion d'une région intronique [111].

Il a été montré que SPI1 est capable d'interagir avec l'ARN via son domaine de liaison à l'ADN [109]. Toutefois, la nature des bases des sites de liaison de SPI1 à l'ARN n'était pas connu au début de ce travail de thèse, dû à l'absence de méthode efficace pour étudier cet aspect. Etant donné la capacité de Spi1 à interagir avec des protéines impliqués dans le contrôle du processus d'épissage des ARN et son rôle direct dans l'épissage, la question se posait de savoir si SPI1 agit dans l'épissage uniquement par interaction fonctionnelle avec les protéines et/ou nécessite sa liaison à l'ARN. La liaison de Spi1 à l'ARN pourrait impliquer d'autres processus. En effet, depuis l'avènement des méthodes à haut débit, le rôle des ARNs dans la régulation des gènes s'est élargi, notamment avec l'identification des longs ARNs non codants ou des petits ARNs. Par exemple, une fonction de la liaison de SPI1 à l'ARN a été décrite récemment [112]. Il est montré que SPI1 recrute la hnRNPK (une protéine de liaison à l'ARN) dans des cellules leucémiques, protéine qui fixe l'ARN transcrit naissant. hnRNPK recrute ensuite NSUN3, une ARN-méthyl-cytosine, qui méthyle l'ARN et active ainsi la structure chromatinienne dans les leucémies. En conclusion, ces résultats ouvrent de nouvelles perspectives des conséquences biologiques de SPI1 fixée à l'ARN et ce champs d'étude reste à parcourir.

1.3 Le séquençage nouvelle génération au service de la génomique

1.3.1 Quelques notions d'histoire

La bioinformatique est une thématique qui a émergé à la fin du vingtième siècle à la suite de progrès en biologie et en physique. En effet à partir des années 70, le séquençage a révolutionné la biologie moléculaire moins de vingt ans après l'établissement de la structure de l'ADN par Watson et Crick [113]. Développée par Sanger [114, 115], la technique de séquençage dite de première génération comporte une étape d'amplification qui génère des fragments d'ADN de différentes longueurs, cela est dû à l'incorporation à l'extrémité des fragments d'ADN d'une base liée à un fluorochrome spécifique de chaque base. La terminaison se fait de façon statistique sur toutes les positions possibles. Une électrophorèse permet ensuite de regrouper et d'ordonner ce mélange de fragments d'ADN de tailles croissantes en fonction de leur longueur. La séquence des couleurs ainsi lue est utilisée pour décoder les séquences du fragment dans son ensemble.

En parallèle de cela, les techniques de construction d'ordinateurs ont évolué et de nouveaux langages de programmation (Fortran 77 (1977), C (1978)) ont vu le jour. Ainsi, l'arrivée à maturité simultanée de ces deux disciplines, biologie moléculaire et informatique, a permis, dès

1977, l'écriture des premiers programmes d'analyses de séquences. En 1981, le premier programme d'alignement local de séquences a vu le jour [116, 117]. Cet algorithme, l'algorithme de Smith et Waterman a pour but d'optimiser l'alignement de deux séquences d'ADN en maximisant le score qui correspond à la qualité de l'alignement. Ce score diminue à mesure que des insertions ou des délétions (indels) sont insérées dans l'alignement, ces indels permettent d'insérer des trous dans l'alignement afin d'avoir la meilleure correspondance entre les séquences. Dans les années suivantes, de nombreux outils d'alignement de séquences sont développés, en particulier l'algorithme BLAST [118] dont le but est de retrouver rapidement des séquences répertoriées présentant des similitudes avec la séquence d'entrée et de trouver ainsi des relations fonctionnelles ou évolutives entre les séquences. Les années 2000 commencèrent avec la publication de la séquence de l'ensemble du génome humain [119] et de nouvelles technologies de séquençage à très haut débit, dites de seconde génération virent le jour.

Les technologies de séquençage de seconde génération, dites à haut-débit ou "high throughput sequencing" ou "next-generation sequencing" (NGS), sont apparues en 2005 en réponse au prix élevé et au faible débit du séquençage de première génération. Des dizaines de milliers de séquences sont alors traitées ensembles, en parallèle. Les principales techniques sont soutenues par des entreprises du nom de Roche, Illumina et Life technologies. Aujourd'hui se sont ajoutées les entreprises PacBio et Oxford Nanopore. Au début, la bioinformatique correspondait à l'utilisation de l'informatique pour stocker/analyser les données de la biologie moléculaire. Aujourd'hui, cette définition a été étendue et la bioinformatique correspond à l'apport de l'informatique et de l'algorithmique pour résoudre les problèmes scientifiques posés par la biologie. C'est un champ de recherche pluri-disciplinaire associant informaticiens, mathématiciens, physiciens, biologistes. En particulier, les données biologiques brutes obtenues à l'issue d'une expérience de séquençage à haut débit sont traitées de manière informatique afin d'extraire des résultats exploitables en clinique ou en recherche. Ces nouvelles techniques de séquençage permettent de séquencer majoritairement des régions d'intérêts. On ne veut pas lire l'ensemble du génome, c'est une méthode ciblée.

1.3.2 Le séquençage à haut débit

La première étape du séquençage à haut débit consiste à préparer le matériel biologique qui sera séquencé. Ce matériel biologique initial peut être de l'ADN ou de l'ARN, une étape de sélection et traitement du matériel biologique permettent d'obtenir les régions d'intérêt. Après cela vient l'étape de préparation de la librairie et de séquençage. La quasi-intégralité des données de ce projet ont été séquencées en utilisant la technologie Illumina, celle-ci est la plus utilisée dorénavant. Un des jeux de données est issu d'une publication de 2012 [37] et le séquençage avait été réalisé en utilisant la technologie SOLiD (Sequencing by Oligonucleotide Ligation and Detection). Aujourd'hui cette méthode n'existe plus.

Technologie SOLiD

Les brins d'ADN sont obtenus par fractionnement aléatoire de l'ADN de l'échantillon à analyser en morceaux de 150 à 200bp. Chaque brin d'ADN est associé à une microbille magnétique d'environ 1 micron d'épaisseur par un petit brin d'ADN. Les billes sont émulsionnées avec les produits d'amplification dans un mélange eau-huile, la PCR en émulsion permet ensuite d'amplifier chaque séquence et d'obtenir plusieurs millions de copies pour chaque bille. Les billes sont ensuite déposées sur une lame où elles sont fixées de manière covalente [10].

Le séquençage est basé sur la ligation d'amorces qui s'hybrident sur les adaptateurs présents sur la matrice. Un jeu de quatre sondes de deux bases marquées en fluorescence sont associées aux amorces. La ligase permet la liaison de la sonde spécifique à la séquence à amplifier. Les produits d'extension sont retirés et une nouvelle amorce complémentaire de la position $n + 1$ est utilisée. Plusieurs cycles de ligation, détection et clivages sont effectués [10]. Chaque base de la séquence à amplifier est interrogée dans deux réactions de ligation indépendantes par deux amorces différentes. Le codage des résultats est effectué sur deux bases dans un espace de quatre couleurs. Ce système permet une très grande fidélité de la lecture des résultats et permet de faire la différence entre les erreurs de séquençages et les variants réels (SNP, insertions et délétions).

Technologie Illumina

L'étape de préparation de la librairie (Fig. 1.10) est faite en phase solide [10] et est détaillée ci dessous :

- (Optionnelle) Fractionnement aléatoire de l'échantillon à analyser en morceaux de 200 bp pour générer une banque d'ADN double brin
- Ajout d'adaptateurs spécifiques aux extrémités
- Dénaturation de l'ADN double brin pour obtenir de l'ADN simple brin
- Fixation des extrémités simple brin aléatoirement à la surface de la cellule à flux
- Formation d'un pont par chaque brin qui va s'hybrider avec l'amorce complémentaire de l'autre extrémité
- Amplification : synthèse du brin complémentaire (un nucléotide marqué par un marqueur fluorescent est incorporé au cours de la synthèse)
- Répétition de cette étape pour former des clusters du même fragment d'ADN

Le premier cycle de séquençage consiste en l'ajout des quatre terminateurs réversibles marqués afin de stopper l'élongation durant la lecture de la fluorescence, des amorces et de l'ADN polymérase. Un laser permet l'émission de la fluorescence de chaque cluster ainsi la première base est lue. Le second cycle commence par l'ajout des quatre terminateurs réversibles marqués, l'image est acquise après excitation par le laser et la deuxième base est lue (Fig. 1.10). Les cycles sont répétés pour lire chaque base les unes après les autres [10].

10. `http://genetique.snv.jussieu.fr/doc2012/120607_SequencageHautDebit.pdf`

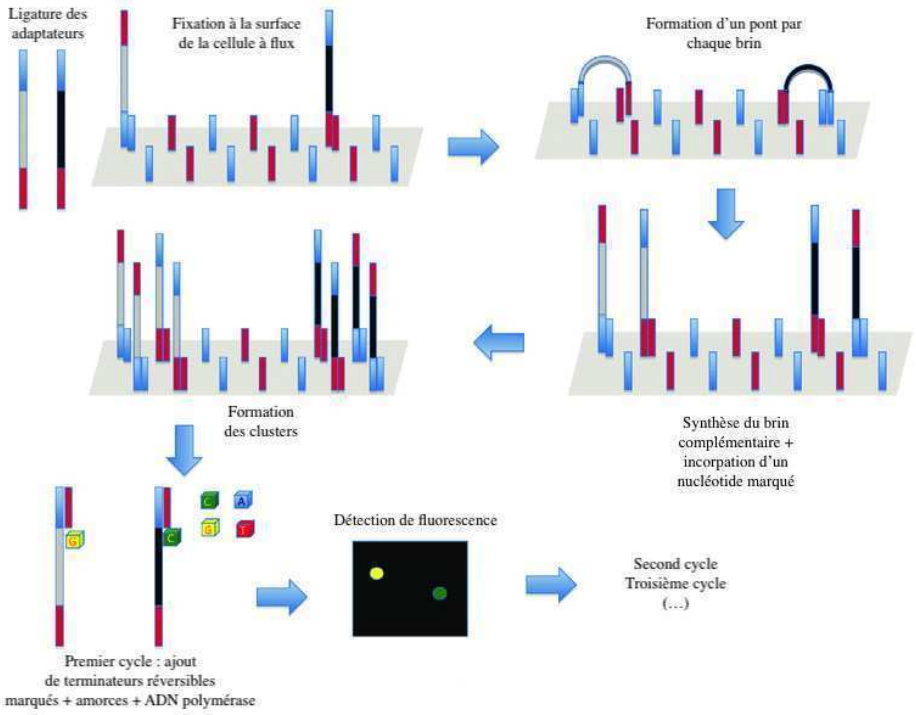


FIGURE 1.10 – **Préparation de la librairie et séquençage Illumina.** *Issue de [120]*

1.3.3 L'analyse du séquençage à haut débit

À l'issue du séquençage à haut débit, on obtient des lectures (ou reads) correspondant à des séquences d'ADN de taille de 50, 75 ou 100bp qu'il faut aligner sur le génome. Le nombre de lectures obtenues dépend de la quantité de matériel génétique initialement déposé sur la cellule de flux. En général il est préférable d'avoir un nombre de lectures assez élevé, c'est à dire une profondeur de séquençage élevée qui est variable suivant la cible de l'expérience, cela permet, dans la mesure du possible, après l'alignement des lectures sur le génome lors de l'analyse de différencier les régions d'intérêt des régions de bruit de fond. La taille des lectures obtenues dépend du kit utilisé pour la librairie. Après le séquençage et plusieurs processus dont celui de démultiplexage, on récupère des fichiers .fastq (Fig. 1.11) qui contiennent pour chaque lecture l'information sur l'identifiant de la lecture, la séquence encodée par celle-ci et sa qualité. L'encodage de la qualité permet, avant de commencer toute analyse, de vérifier que les lectures sont de qualité correcte. Ensuite, elles sont alignées sur le génome (Fig. 1.11), cette étape peut être compliquée à réaliser en raison de la petite taille des fragments séquencés et de la présence de régions de faible complexité sur le génome (séquences répétées). Le contrôle qualité et l'alignement des lectures sur le génome font partie des premières étapes de l'analyse bioinformatique, ainsi si elles sont est mal faites, les analyses qui en découleront pourront donner lieu à de faux résultats, il est donc très important d'utiliser des méthodes d'alignement de qualité.

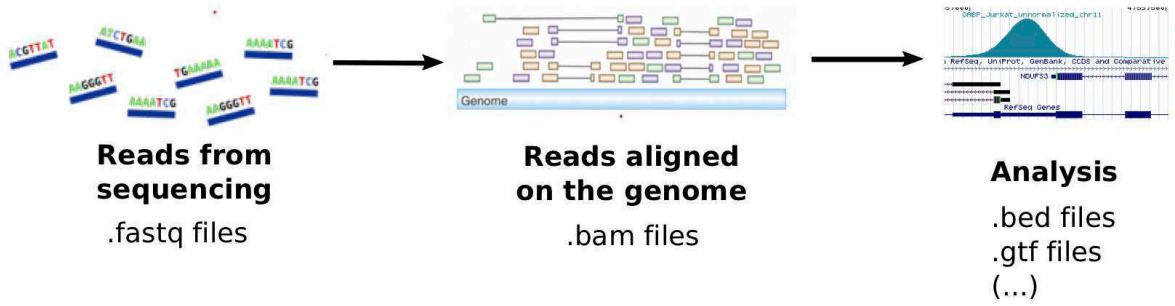


FIGURE 1.11 – **Schéma d'une analyse classique de données NGS**

Dans les sections suivantes, je détaillerai certaines des expériences de séquençage à haut débit qui ont été utilisées pour obtenir les données exploitées dans ce manuscrit et je m'intéresserai aux points importants de l'analyse bioinformatique de ces expériences. La seule analyse bioinformatique que je n'ai pas effectuée est celle concernant les données de RNA-seq et ne sera donc pas détaillée.

1.3.4 Cartographie des sites des interactions ADN-protéines : ChIP-seq

La section 1.2.6 détaille l'ensemble des facteurs épigénétiques avec lesquels SPI1 interagit. En particulier, SPI1 interagit avec de nombreuses enzymes capables de modifier chimiquement les histones. Au cours de ce travail, nous avons caractérisé le rôle de SPI1 dans la modification des protéines histones afin de cartographier leurs modifications et de les associer aux sites de liaison de SPI1 sur le génome. Le **ChIP-seq** (**Ch**romatin **I**mmuno**P**récipitation with high-throughput **seq**uencing) est une méthode de séquençage à haut débit permettant de localiser les protéines sur l'ADN.

Les principales étapes de l'expérience de ChIP-seq sont décrites sur la Fig. 1.12 et la méthode utilisée est décrite dans les annexes A.4. Brièvement, les protéines sont fixées à la molécule d'ADN par pontage au formaldéhyde, les cellules sont ensuite lysées et l'ADN est fragmenté (étapes 1 et 2 sur la Fig. 1.12). L'immunoprécipitation est réalisée avec un anticorps spécifique de la protéine d'intêret. Avant de procéder à cette étape, une fraction de la chromatine fragmentée (généralement appelée input) est isolée pour être utilisée comme contrôle de l'expérience. L'immunoprécipitation est réalisée à l'aide de billes sur lesquelles se fixent les anticorps ayant reconnu les fragments d'ADN avec la protéine d'intêret (étape 3 sur la Fig. 1.12). Plusieurs étapes de lavages permettent d'éliminer les complexes ADN-protéines qui n'ont pas été fixés par les billes. Les complexes ADN-protéine du ChIP et de l'input sont purifiées et le reverse crosslink est opéré (étape 4 sur la Fig. 1.12). Après cette étape, l'ADN est purifié, les adaptateurs sont accrochés aux fragments d'ADN dépourvus de la protéine d'intêret (étape 5 sur la Fig. 1.12), la librairie est préparée et l'ADN est séquencé (étape 6 sur la Fig. 1.12).

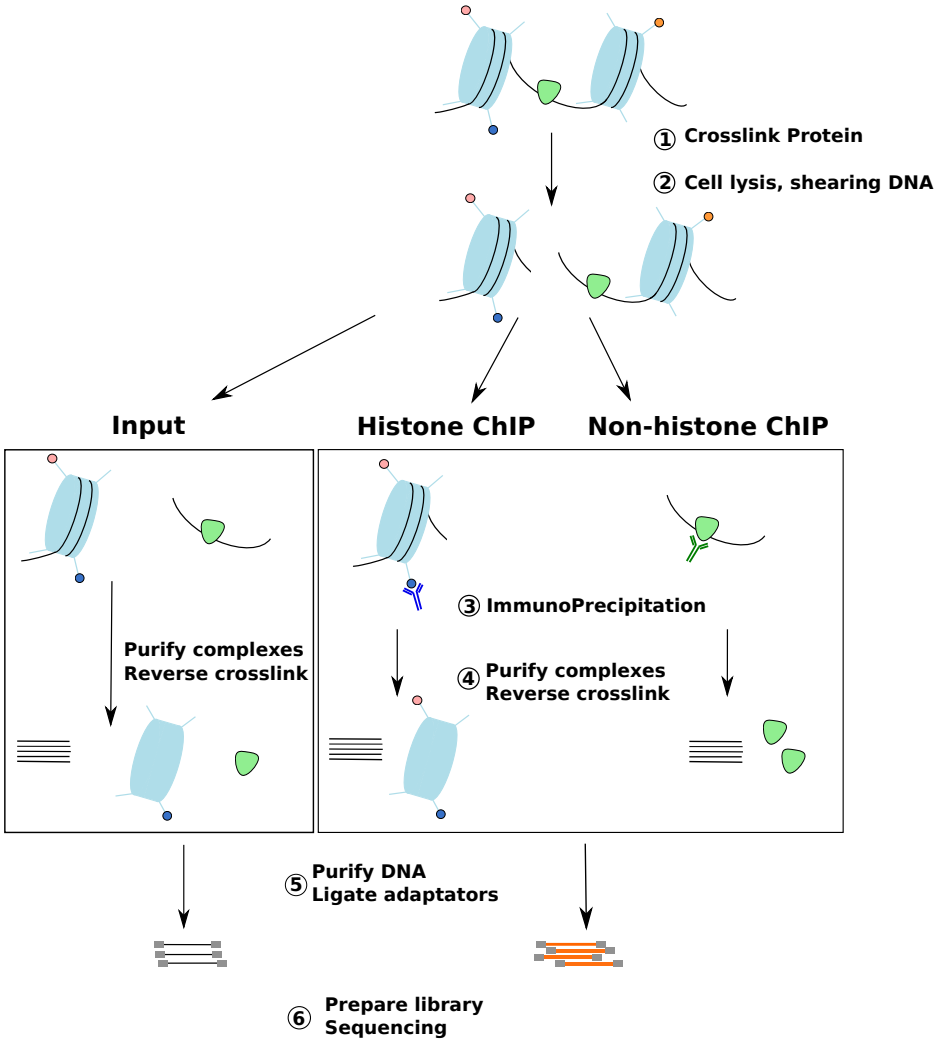


FIGURE 1.12 – **Protocole de Chromatin ImmunoPrecipitation followed by sequencing (ChIP-seq)**

Analyse du ChIP-seq

Les expériences de ChIP-seq pratiquées dans le but d'étudier la liaison d'un facteur de transcription ou la modification d'une marque d'histone dans un type de cellules particulier suivent un pipeline d'analyse assez classique. Tout d'abord les lectures sont alignées sur le génome en utilisant des logiciels tels que Bowtie [121] ou Burros Wheeler Aligner (BWA) [122], ensuite les potentiels sites de liaison du facteur de transcription ou d'enrichissement en la marque d'histone d'intêret sont repérés grâce au peak-calling avec des logiciels tels que MACS2 [123] ou HMCan [124]. MACS2 [123] est le logiciel le plus utilisé pour le peak calling, toutefois HMCan [124] présente des avantages pour les échantillons de tumeurs tels que la correction du biais en nombre de copie qui est inhérent aux échantillons cancéreux.

L'étude des motifs présents dans les séquences d'ADN des régions identifiées à l'étape précédente peut se faire de deux façons : (i) recherche de motifs déjà connus enrichis ; (ii) recherche

de motif *de novo*. Par exemple, dans des régions qui portent une marque d'histone la recherche de motifs connus peut permettre d'identifier des facteurs de transcription particuliers pouvant être liés aux région d'intêret. De plus, dans des régions de liaison du facteur de transcription analysé, la recherche de motifs connus peut permettre de trouver des facteurs de transcription qui coopèrent avec le facteur de transcription d'intêret. La recherche de motif *de novo* sur un jeu de données de ChIP-seq pour un facteur de transcription inconnu permet d'identifier le motif de celui-ci. Plusieurs outils ont été développés pour réaliser ces deux types d'analyses, on peut citer icisTarget [125], HOMER [72], ChIPMunk [126], **R**egulatory **S**equence **A**nalysis **T**ools (**RSAT**) [127–131], AME ou CentriMO de la suite MEME [132].

Certaines analyses peuvent être plus compliquées, l'analyse différentielle en fait partie. En effet, la comparaison des signaux de ChIP-seq entre différentes conditions ou entre différents échantillons a toujours été une question compliquée. Plusieurs paramètres sont à prendre à compte lors de la comparaison de signaux de ChIP-seq entre différents échantillons ou conditions. Tout d'abord, l'anticorps utilisé peut avoir une efficacité variable suivant les échantillons ou conditions. Ensuite, les expériences de ChIP-seq sont sujettes à une variabilité dans la longueur des fragments d'ADN et peuvent présenter des différences dans le comptage des lectures à cause d'un biais dans l'amplification de l'ADN. L'ensemble de ces biais peut causer des variations artificielles dans les signaux de ChIP-seq et amener à des conclusions biologiques erronées. Ainsi, une technique de normalisation permettant de prendre en compte tous ces biais est nécessaire, surtout lorsque l'information de spike-in n'est pas disponible [133]. Cette technique repose sur l'ajout de chromatine et d'anticorps (spike-in) provenant d'un organisme différent de celui étudié. Ainsi, toute variation introduite dans la réaction de ChIP-seq sera également retrouvé dans la chromatine spike-in. Malheureusement, la plupart des expériences ne comprennent pas de spike-in encore [133].

Dans la plupart des études, la normalisation de ChIP-seq est réalisée en utilisant le nombre total de fragments séquencés par échantillon. Les données sont corrigées en utilisant un facteur constant, par exemple en utilisant la fonction *"bamCompare"* fournie par le package Deeptools [134]. Malheureusement, cette façon de normaliser ne permet pas de prendre en compte les différences introduites par les biais liés à l'efficacité différente de l'anticorps et à l'amplification de l'ADN entre les différents échantillons. D'autres méthodes telles que **N**ormalization of **ChIP**-**S**eq (**NCIS**) [135], qui étend la méthode CisGenome [136], utilise l'échantillon input pour normaliser. Dans la plupart des processus de normalisation, le génome est découpé en bins qui ne s'overlappent pas. Si l'on savait quels bins appartiennent au bruit de fond, alors un facteur de normalisation serait :

$$r = \frac{\sum_{i \in B} x_i^{ChIP}}{\sum_{i \in B} x_i^{input}} \tag{1.1}$$

où B correspond à l'ensemble des bins i appartenant au bruit de fond. Dans CisGenome [136], ils déterminent les régions de bruit de fond comme l'ensemble des régions du génome qui

ont un nombre de lectures inférieur $t \leq 1$ et une largeur $w = 100$, le facteur de normalisation est calculé en utilisant ces paramètres. Dans NCIS [135], le rapport marginal ChIP/Input par rapport aux comptes totaux est utilisé pour déterminer la valeur optimale de t et w de façon itérative sur l'ensemble des données pour déterminer l'ensemble B.

Des méthodes plus sophistiquées [137–139] permettent à l'utilisateur de déterminer les régions différentiellement enrichies pour la protéine d'intêret, tout en traitant le problème de normalisation mais ne fournissent pas les profils de ChIP-seq normalisés. Par exemple, une méthode qui repose sur la détermination des régions différentiellement enrichies dans le génome étant donné deux bibliothèques de modifications d'histones de différents types cellulaires a été suggérée [138]. Tout d'abord, le bruit de fond est estimé, puis le biais génomique local est éliminé et enfin le problème de normalisation est traité en utilisant la normalisation quantile. En outre, la plupart des analyses de ChIP-seq sont réalisées en utilisant l'outil MACS2 [123], cet outil comprend une fonction appelée *bdgdiff*. Celle-ci permet d'identifier les pics différentiels, mais les profils normalisés résultant ne sont pas fournis. Une méthode de normalisation non linéaire en deux étapes basée sur l'approche de régression pondérée localement (LOESS) a également été développée [137]. Le but est de comparer les données de ChIP-seq entre plusieurs échantillons et de modéliser la différence en utilisant un modèle statistique de mélange de loi exponentielles. Le modèle ajusté est utilisé pour identifier les gènes associés aux sites de liaison différentiels en utilisant un FDR [11]. ChIPSeqSpikeInFree [139] est une méthode de normalisation basée sur le calcul de la pente de la courbe du nombre cumulé de lectures pour chaque échantillon. L'équipe dans laquelle j'ai effectué ma thèse a développé une méthode de normalisation simple appelée LILY qui est basée sur l'appariement du signal à l'intérieur des pics communs les plus forts [140]. Le facteur de normalisation est calculé comme le rapport des valeurs de densité dans ces régions communes.

Pour résumer, les méthodes existantes évaluent les paramètres de normalisation en se basant sur les pics de ChIP-seq communs, ou en utilisant l'input mais n'accepte pas d'informations supplémentaires pouvant permettre une détermination plus fine de ces paramètres de normalisation. Au cours de ma thèse, j'ai eu à analyser beaucoup de données de ChIP-seq dans des conditions différentes pour des marques d'histones différentes. Ainsi, une partie de ma thèse a été consacrée au développement d'un package R pour la normalisation de données de ChIP-seq entre différents échantillons ou différentes conditions. Cette méthode est basée sur l'hypothèse biologique selon laquelle les gènes dont l'expression est constante entre les différents échantillons ont, en moyenne des signaux de ChIP-seq similaires. La sortie du package correspond aux fichiers .bigwig normalisés, directement visualisables sur le logiciel IGV. Cette méthode est détaillée dans la section 3.

11. **FDR** : **F**alse **D**iscovery **R**ate

1.3.5 Cartographie des sites des interactions ARN-protéines : CLIP-seq

SPI1 est un facteur de transcription capable de se fixer à l'ARN comme nous l'avons décrit dans la partie 1.2.6, toutefois la méthode utilisée SELEX [141] ne permettait pas de caractériser la nature des bases et motifs reconnus par SPI1 à l'ARN (non publié). Afin de caractériser la liaison de SPI1 à l'ARN une méthode de séquençage à haut débit nommée **CLIP-seq** (**C**ross-**L**ink **I**mmuno**P**recipitation with highthroughput **seq**uencing) a été utilisée [142]. Le principe de l'expérience est décrit sur la Fig. 1.13, la liaison de la protéine d'intêret à l'ARN est renforcée en traitant les cellules aux rayons ultra-violets (UV) à 254nm (étape 1 sur la Fig. 1.13). Les ARN non fixés à la protéine sont traités à la RNase afin d'être éliminés et ne garder que ceux dont la liaison à la protéine empêche la dégradation (étape 2 sur la Fig. 1.13). Les ARN ainsi récupérés sont d'environ 50 bases. Une immunoprécipitation avec l'anticorps spécifique reconnaissant la protéine d'intêret permet de récupérer les ARN sur lesquels est fixée la protéine (étape 3 sur la Fig. 1.13), une électrophorèse en milieu dénaturant permet de séparer ces complexes ARN-protéines selon la taille de la protéine +/-20kD (étape 5 sur la Fig. 1.13). Le complexe ARN-protéine est alors récupéré directement dans le gel puis extrait grâce à la protéinase K et purifié (étape 6 sur la Fig. 1.13). Les ARN rétro-transcrits en ADNc* [12] après RT-PCR sont ensuite séquencés (étapes 7, 8 et 9 sur la Fig. 1.13).

12. **ADNc** : **ADN c**omplémentaire

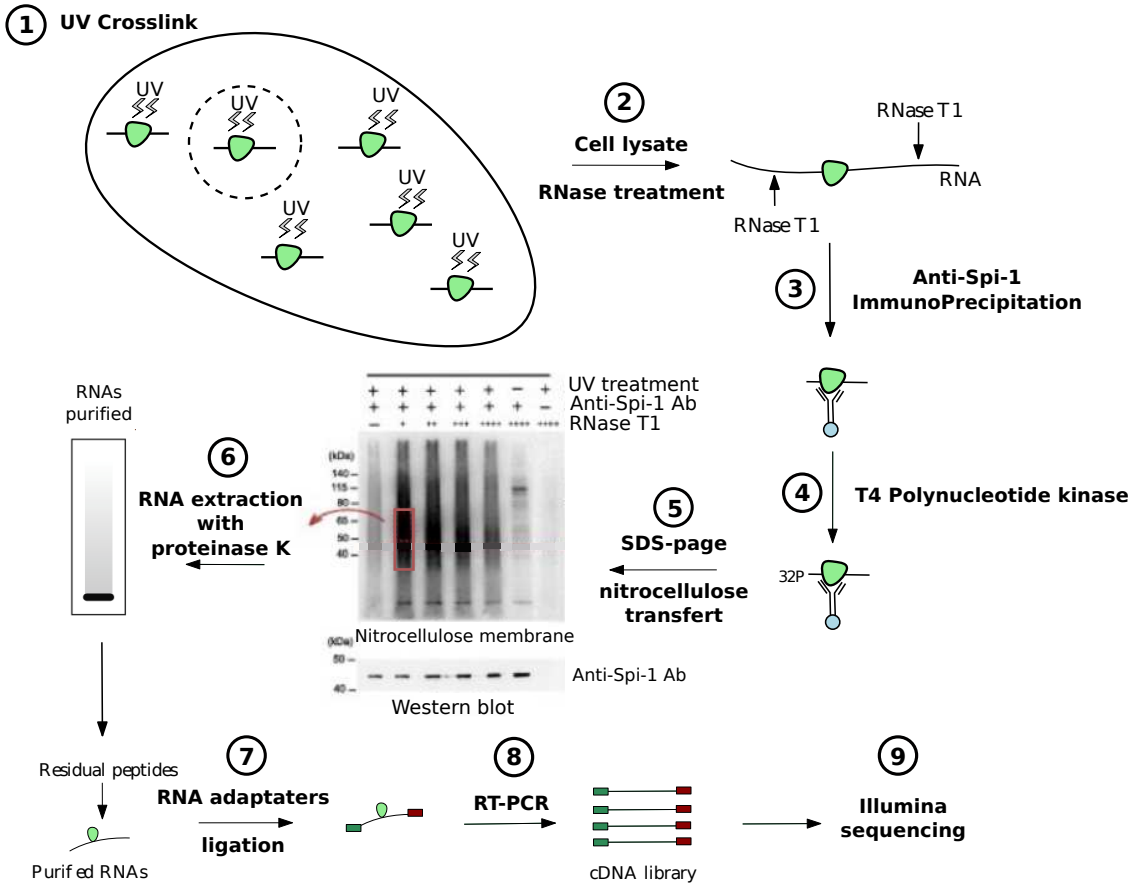


FIGURE 1.13 – **Protocole de Cross-Linking ImmunoPrecipitation followed by sequencing (CLIP-seq)**

La technique de CLIP-seq également nommée HITS-CLIP (**HI**gh **T**hroughput **S**equencing-**CLIP**) est apparue en 2008, c'est-à-dire il y a plus de dix ans, depuis de nouvelles techniques pour caractériser la liaison des protéines sur l'ARN ont fait leur apparition. En 2010, le PAR-CLIP pour **P**hoto**A**ctivatable-**R**ibonucleoside-Enhanced **C**ross**L**inking and **I**mmuno**P**recipitation apparaît et repose sur l'incorporation de ribonucléosides photo-réactifs dans les cellules en culture qui seront accrochés de façon plus efficace à l'ARN grâce aux UV à une longueur d'onde égale à 254 nm. Cette méthode a été effectuée sur les cellules préleucémiques d'intêret surexprimant SPI1, mais n'a pas fonctionné. En 2014, le iCLIP (**i**ndividual-nucleotide **CLIP**) [143] reprend le principe du HITS-CLIP mais permet de récupérer les produits de la RT que la reverse-transcriptase n'a pas pu rétro-transcrire entièrement, ces fragments d'ADNc très courts qui sont perdus durant l'analyse du HITS-CLIP. L'ajout d'une seconde amorce en 5' de l'ADNc après le décrochage de la polymérase permet de cartographier précisément le site d'interaction entre la protéine et l'ARN. En 2016, le eCLIP (**e**nhanced CLIP) [144] modifie la méthode iCLIP pour inclure des améliorations dans la préparation des bibliothèques, la ligature d'un adaptateur en 3' de l'ADNc avec un barcode spécifique à la séquence permet de reconnaître les duplicats de

PCR.

Analyse du CLIP-seq (HITS-CLIP)

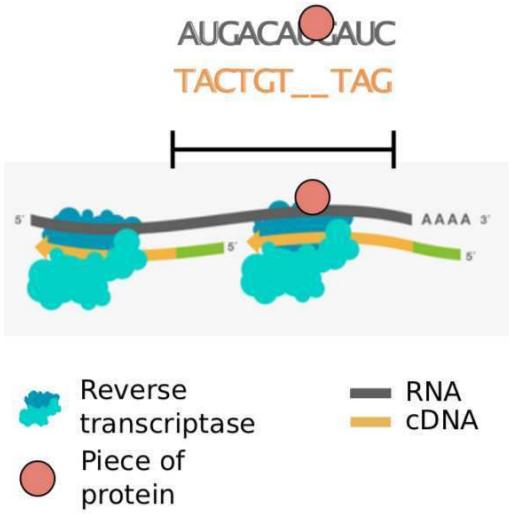


FIGURE 1.14 – **Processus de reverse transcription au cours de l'expérience de CLIP-seq.** Au cours de cette étape, la reverse transcriptase peut mal recopier la séquence d'ARN en ADNc à cause de la présence d'un résidu de protéine.

La technique de CLIP-seq a surtout été utilisée pour caractériser la liaison de facteurs d'épissage à l'ARN [145, 146], les facteurs de liaison aux ARN non codants [147], la liaison des **RBP** (**R**NA-**B**inding-**P**roteins) [148]. En revanche la liaison des facteurs de transcription à l'ARN a été peu caractérisée. De plus, il n'existe pas de pipeline bioinformatique complète pour analyser les données de CLIP-seq (de l'alignement à la recherche de motif) ni d'outil permettant d'intégrer les données de ChIP-seq et de CLIP-seq. L'étape d'alignement des lectures sur le génome est une étape très importante comme nous l'avons déjà mentionné précédemment, en particulier dans le cas des données de CLIP-seq. En effet, les lectures peuvent provenir d'ARNm prématures (constitués à la fois exons et d'introns) et d'ARNm matures (qui ont subi l'épissage et donc incluant principalement des exons). L'alignement précis de ces lectures est très important dans la mesure où la propension d'un facteur de transcription à se fixer sur l'ARNmessager mature ou prématurée est une donnée importante pour les processus biologiques qui seront suggérés suite à l'analyse du CLIP-seq. De plus, la recherche des sites de liaison de la protéine sur l'ARN est une étape au cours de laquelle plusieurs biais sont induits :

— lors de l'étape de reverse transcription (étape 8 sur la Fig. 1.13), la reverse-transcription-polymerase peut faire des erreurs si lors de l'étape réalisée par la protéinase K toutes les protéines n'ont pas été correctement éliminées. En effet des délétions, des insertions ou des mutations peuvent être introduites dans la séquence d'ADNc (Fig. 1.14).

— la quantité de transcrits dans la cellule au moment de l'immunoprécipitation peut induire un biais dans le nombre de protéines fixées sur l'ARN.

Ainsi les étapes d'alignement et de recherche des sites de liaison lors de l'analyse de données de CLIP-seq sont deux étapes qui sont déterminantes pour les analyses subséquentes.

1.3.6 Cartographie des sites d'accessibilité de la chromatine : ATAC-seq

Dans la section précédente nous avons mentionné que SPI1 était capable de modifier l'ouverture de la chromatine. Une méthode de séquençage à haut débit appelée **A**ssay for **T**ransposase-**A**ccessible **C**hromatin with highthroughput **seq**uencing (**ATAC-seq**) permet de caractériser les régions accessibles à une transposase (Tn5) dans la chromatine. Cette accessibilité dépend de la position et de la structure des nucléosomes. En effet, la transposase aura accès à l'ADN nu ce qui permettra de définir les régions sans nucléosome. L'enzyme porte avec elle deux fragments d'ADN qu'elle ajoute à chaque extrémité de l'ADN qu'elle fragmente. Elle permet donc en une réaction de fragmenter l'ADN accessible et de lier les adaptateurs aux fragments pour l'amplification par PCR. Cette méthode est assez récente puisqu'elle date de 2013. D'autres techniques étaient employées avant telles que le MNase-seq (Micrococal hypersensitive sites sequencing), le **DNase-seq** (**DNase** I hypersensitive sites **seq**uencing) qui repose sur la digestion de l'ADN par la DNase1 ou le FAIRE-seq (**F**ormaldehyde-**A**ssisted **I**solation of **R**egulatory **E**lements **seq**uencing) qui repose sur un crosslinking des protéines à l'ADN suivi d'une fragmentation et d'une purification de l'ADN. Le protocole d'ATAC-seq est très simple (Fig. 1.15), c'est pourquoi il est maintenant préféré au FAIRE-seq et DNase-seq. De plus, il nécessite peu de cellules, environ 100000 cellules. Récemment, l'ATAC-seq a été développement pour l'analyse dans une seule cellule (single-cell-ATAC-seq 10X Genomics).

Analyse ATAC-seq

Un pipeline d'analyse d'ATAC-seq est assez similaire à un pipeline d'analyse de ChIP-seq. En effet, après une étape d'alignement réalisée avec BWA [122] par exemple, le peak calling permet d'identifier les régions de chromatine ouverte. Le logiciel HMCan [124] comporte une option spécifique pour les données d'ATAC-seq. Le biais en GC n'est pas corrigé et le paysage des nucléosomes est évalué à partir de la distribution de la taille des fragments. En effet, le transposome Tn5 fragmente l'ADN autour des nucléosomes (Fig. 1.15). Ainsi les fragments dont la taille sera inférieure à 130bp sont vraisemblablement des régions inter-nucléosomiales, puisque la taille moyenne de l'ADN enroulé autour d'un nucléosome varie entre 160 et 210bp. L'analyse de données ATAC-seq permet de repérer les régions sans nucléosomes. Pour cela, il faut rendre les conditions comparables entre elles, tout comme lors de l'analyse des données de ChIP-seq mentionnée dans la partie 1.3.4. Un article récent [150] suggère que la méthode de normalisation choisie pour cela est fondamentale pour éviter des conclusions biologiques

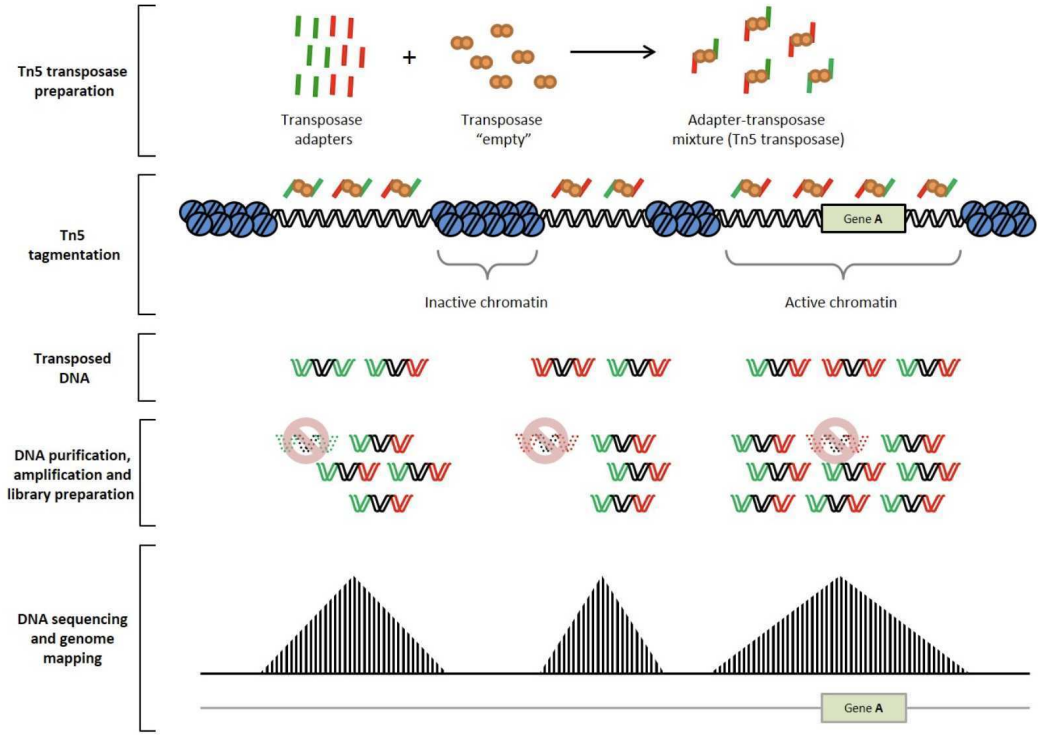


FIGURE 1.15 – **Protocole de Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq)** *Issue de [149].* La Tn5 porte avec elle deux fragments d'ADN qu'elle ajoute à chaque extrémité de l'ADN qu'elle fragmente. Seules les régions de chromatine avec un fragment différent à chaque extrémité (rouge et vert) sont conservées pour les étapes de purification, amplification et préparation de la librairie. L'ADN est ensuite séquencé et les fragments d'ADN sont alignés sur le génome lors de l'étape d'alignement détaillée ci-dessous.

fausses. Ainsi, nous avons utilisé la méthode CHIPIN que j'ai développée au cours de ma thèse pour normaliser les données de ChIP-seq entre les différentes conditions. En effet, on peut faire l'hypothèse qu'une région de chromatine sera d'autant plus relâchée que le gène associé sera hautement transcrit. Ainsi, utiliser les gènes dont l'expression ne varie pas entre les conditions pour inférer les paramètres de normalisation est raisonnable.

1.4 Objectifs de la thèse

Dans l'introduction j'ai mis en place les concepts nécessaires pour comprendre le travail de cette thèse. Dans un premier temps, la régulation de l'expression génique par les acteurs de la transcription mais aussi par l'organisation de la chromatine, organisation contrôlée par des facteurs épigénétiques et des facteurs de transcription particuliers : les facteurs pionniers dont SPI1 fait partie, ont été discutés. Puis dans un second temps, je me suis intéressée aux méthodes

de séquençage de seconde génération.
SPI1 est un facteur de transcription qui est étudié depuis 32 ans. Il appartient à la famille ETS, c'est un acteur majeur de l'hématopoïèse. Son implication dans l'ensemble des lignages hématopoïétiques fait de lui un facteur de transcription très complexe qui agit différemment selon son niveau d'expression, selon le lignage hématopoïétique et selon ses co-facteurs disponibles. Cette complexité s'observe également dans son activité leucémogène puisqu'il peut se comporter comme un oncogène ou comme un oncosuppresseur.

Les modes d'action de SPI1 connus sur la régulation des gènes ont été présentés tant lorsque SPI1 agit seul que lorsqu'il agit en synergie avec des partenaires protéiques : facteurs de transcription et facteurs épigénétiques. Le rôle de SPI1 en tant qu'activateur transcriptionnel est maintenant bien décrit dans les différentes lignées hématopoïètiques où il exerce une fonction. J'ai également décrit les fonctions oncogéniques de SPI1 dans la régulation transcriptionnelle dans l'érythroleucémie. En revanche, peu d'études se sont attachées à comprendre le mode d'action de SPI1 dans la répression génique et seules deux études récentes (2018 et 2019), décrivent le mode d'action de SPI1 dans la répression de certains de ses gènes cibles en utilisant les nouvelles méthodes de séquençage à haut débit. En effet, dans les lymphocytes T [107], SPI1 réprime l'expression génique de façon indirecte sans liaison à l'ADN. Dans les neutrophiles [108], SPI1 contrôle l'activité et la structure chromatinienne des enhancers.

L'objectif de la première partie de ma thèse est de caractériser les mécanismes de régulation de la répression génique par le facteur de transcription SPI1 dans le cas de l'érythroleucémie chez la souris en utilisant des méthodes récentes : les données de séquençage à haut débit, associées à des analyses fonctionnelles. Ces méthodes m'ont permis de décrire les sites de liaison de SPI1 à l'ADN et à l'ARN et d'autre part les modifications épigénétiques, l'accessibilité à la chromatine, la régulation des gènes en présence et en absence de SPI1.

Cette étude s'appuie sur l'ensemble des données de ChIP-seq, RNA-seq et ATAC-seq et peut être résumée en trois questions :

- Comment un gain de fonction peut engendrer la répression de certains gènes ?
- La répression se fait-elle à travers la liaison directe de SPI1 à l'ADN ou est-ce un effet indirect sans liaison à l'ADN ?
- SPI1 n'ayant pas d'activité de répression propre, avec quelle protéine interagit-il pour induire la répression ?

Pour répondre à ces problématiques, il m'a fallu analyser des données ChIP-seq pour différentes marques d'histones dans des conditions différentes. Ainsi, j'ai développé une méthode de normalisation de ChIP-seq entre différentes conditions. Cette méthode est basée sur l'hypothèse biologique selon laquelle les gènes dont l'expression est constante entre les différents échan-

tillons, ont, en moyenne, des signaux de ChIP-seq similaires. **Un chapitre complet (chapitre 3) de ce manuscrit est consacré à cette méthode qui a fait l'objet d'une publication**.

La dernière partie de ma thèse est dédiée à la compréhension du rôle de la liaison de SPI1 à l'ARN. Il a été possible grâce au développement de méthodes à haut débit (CLIP-seq) de rechercher les cibles des facteurs de transcription sur les ARNs.

Cette étude repose sur l'analyse des données de CLIP-seq et de RNA-seq et peut également être décrite en trois questions :

— Où SPI1 se fixe-t-il à l'ARN ?
— Existe-t-il un lien entre la liaison de SPI1 à l'ADN et à l'ARN ?
— Quelles sont les conséquences de la liaison de SPI1 à l'ARN sur l'expression génique ?

Ce travail de thèse est le fruit d'une collaboration entre deux disciplines et deux équipes que sont la biologie et la bioinformatique. Ma thèse est co-dirigée par Valentina Boeva, directrice de l'équipe dans laquelle j'ai commencé ma thèse et Christel Guillouf qui dirige une équipe à l'Institut Gustave Roussy. L'ensemble des données biologiques a été généré par l'équipe de Christel Guillouf et en particulier par Sebastian Gregoricchio et Michaela Esposito. En particulier tous les résultats de biologie qui sont présentés dans cette thèse et dont les protocoles sont détaillés en annexe sont le fruit de leur travail.

Chapitre 2

Modèle biologique et données disponibles

2.1 Le modèle biologique

Le modèle utilisé pour étudier le rôle oncogénique de SPI1 dans la lignée érythroïde pour les expériences est constitué des souris transgéniques qui surexpriment le gène *spi1* (TgSpi1). Les cellules pré-leucémiques issues de ces souris ont été utilisées, ces cellules sont des progéniteurs érythroïdes appelés CFU-E ou proérythroblastes dont la différenciation est bloquée [66]. Elles sont indépendantes de l'Epo pour leur prolifération et leur survie [66] (pour détails, voir section 1.2.5).

A partir des cellules pré-leucémiques (appelées aussi HS1), les chercheurs du laboratoire ont généré des lignées dans lesquelles l'expression de *spi1* peut être réprimée de façon inductible *in vitro*. La répression est basée sur l'introduction de deux vecteurs d'expression dans les cellules HS1 : un vecteur exprimant le répresseur à la tétracycline (TetR) et un vecteur exprimant un shRNA anti-spi1. L'expression du shRNA anti-spi1 est bloquée par le TetR. L'addition de la doxycycline (dox), un analogue de la tétracycline, dans le milieu de culture permet l'inactivation du TetR et induit ainsi la production du shRNA anti-spi1. La dox induit une diminution de l'expression de la protéine SPI1 de 90% après 48H de traitement. Deux types de cellules sont alors disponibles :

— des cellules qui surexpriment *Spi1* (**Spi1+**)

— des cellules dont l'expression de *Spi1* est réduite par ajout de la dox (**Spi1-**).

On observe sur la Fig. 2.1 que les cellules dans lesquelles l'expression de *Spi1* est réprimée ont un aspect plus rouge ce qui est dû au fait que les cellules produisent de l'hémoglobine ; elles ont donc récupéré leur capacité à se différencier. D'autres cellules vont mourir par apoptose.

C'est sur ces deux types de cellules qu'ont été réalisées les expériences de séquençage à haut débit qui ont été présentées dans les sections précédentes et dont la liste va être donnée dans la section suivante.

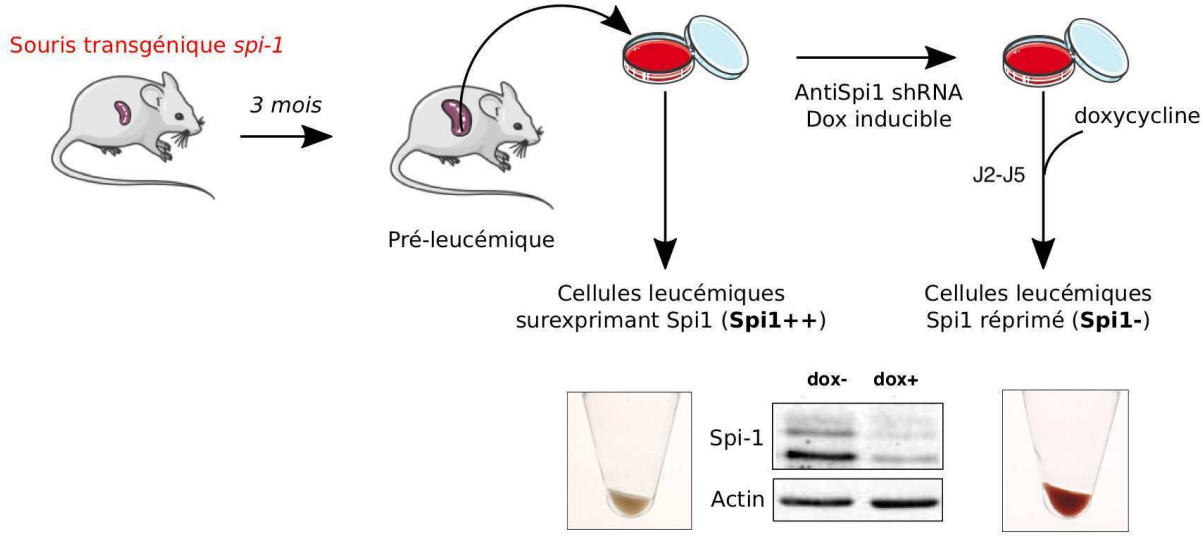


FIGURE 2.1 – **Modèle biologique** [67]

2.2 Les données disponibles

Dans la Tab. 2.1 sont récapitulées les données de séquençage à haut débit que j'ai utilisées au cours de ma thèse. Plusieurs marques d'histones ont été caractérisées par des expériences de ChIP-seq dans les deux conditions, *Spi1* surexprimé (Spi1+) ainsi que *Spi1* réprimé (Spi1-). Deux jeux de données pour l'expérience de ChIP-seq ciblant SPI1 sont disponibles, celles de la publication [37] (condition Spi1+ uniquement) ainsi qu'un jeu de données avec les deux conditions Spi1+ et Spi1-. Je me suis également appuyée sur des données de RNA-seq afin de déterminer les gènes activés ou réprimés par SPI1, des données de CLIP-seq pour caractériser la fixation de SPI1 à l'ARN lorsque *Spi1* est surexprimé et des données d'ATAC-seq révélant les régions de chromatine ouverte en présence de SPI1 surexprimé et en absence de SPI1. Toutes les expériences ont été effectuées au moins deux fois.

Expérience	Protéine ciblée	Condition	Séquençage	Librairie	Nbr de réplicats	Taille des reads	Numéros d'accession
ChIP-seq	H3K27me3	Spi1+/Spi1-	Illumina	Paired-end	2	50bp	non-publié
ChIP-seq	H3K27ac	Spi1+/Spi1-	Illumina	Paired-end	2	50bp	non-publié
ChIP-seq	H3K27ac	Spi1+ +/-Entinostat	Illumina	Paired-end	2	100bp	non-publié
ChIP-seq	H3K4me3	Spi1+/Spi1-	Illumina	Paired-end	2	50bp	non-publié
ChIP-seq	H3K4me1	Spi1+/Spi1-	Illumina	Paired-end	2	100bp	non-publié
ChIP-seq	H3K36me3	Spi1+/Spi1-	Illumina	Single-end	2	100bp	non-publié
ChIP-seq	RNAPolII	Spi1+/Spi1-	Illumina	Single-end	2	100bp	non-publié
ChIP-seq	SPI1 [37]	Spi1+	SOLiD	Single-end	2	50bp	GSE33611
ChIP-seq	SPI1	Spi1+/Spi1-	Illumina	Single-end	2	100bp	non-publié
ChIP-seq	GATA1 [151]	Spi1+/Spi1-	Illumina	Single-end	2	100bp	ERA000161
CLIP-seq	SPI1	Spi1+	Illumina	Single-end	4	50bp	non-publié
RNA-seq	-	Spi1+/Spi1-	Illumina	Paired-end	3	50bp	non-publié
ATAC-seq	-	Spi1+/Spi1-	Illumina	Paired-end	2	42bp	non-publié

TABLE 2.1 – **Données disponibles**

Chapitre 3

Développement d'une méthode de normalisation inter-conditions de ChIP-seq

Dans l'introduction, plus précisément dans la section 1.3.4, j'ai mentionné le fait que lorsque l'on veut comparer le comportement d'une marque d'histone ou d'une protéine dans une condition versus une autre à l'aide de données de ChIP-seq il faut rendre les conditions comparables. Pour cela, il faut normaliser les données entre elles. Dans cette section je vais présenter une méthode de normalisation que j'ai développée au cours de ma thèse qui a déjà été mentionnée au cours de l'introduction et dont le nom est **CHIPIN** pour **ChIP**-seq **I**ntersamples **N**ormalisation. Cette méthode a été soumise en tant qu'article de recherche dans le journal *BMC Bioinformatics* en Octobre 2020.

CHIPIN est un package R, disponible en ligne sur GitHub (`https://github.com/BoevaLab/CHIPIN`), une vignette y est aussi accessible. Plusieurs exemples sont fournis et des données tests sont téléchargeables. La méthode est basée sur l'hypothèse suivante : en moyenne, aucune différence dans les vrais signaux de ChIP-seq ne doit être observée dans les régions régulatrices de gènes dont l'expression est constante entre les échantillons ou conditions. Ces gènes sont utilisés pour déterminer les paramètres de normalisation. Notre méthode est séparée en trois étapes (Fig. 3.1A). Une étape optionnelle permet à l'utilisateur de vérifier la spécificité de l'anticorps utilisé. Les trois premières étapes correspondent à la fonction *CHIPIN_normalize* et l'étape optionnelle est implémentée dans une autre fonction, *plot_expression*, celle-ci peut être exécutée indépendamment de la première.

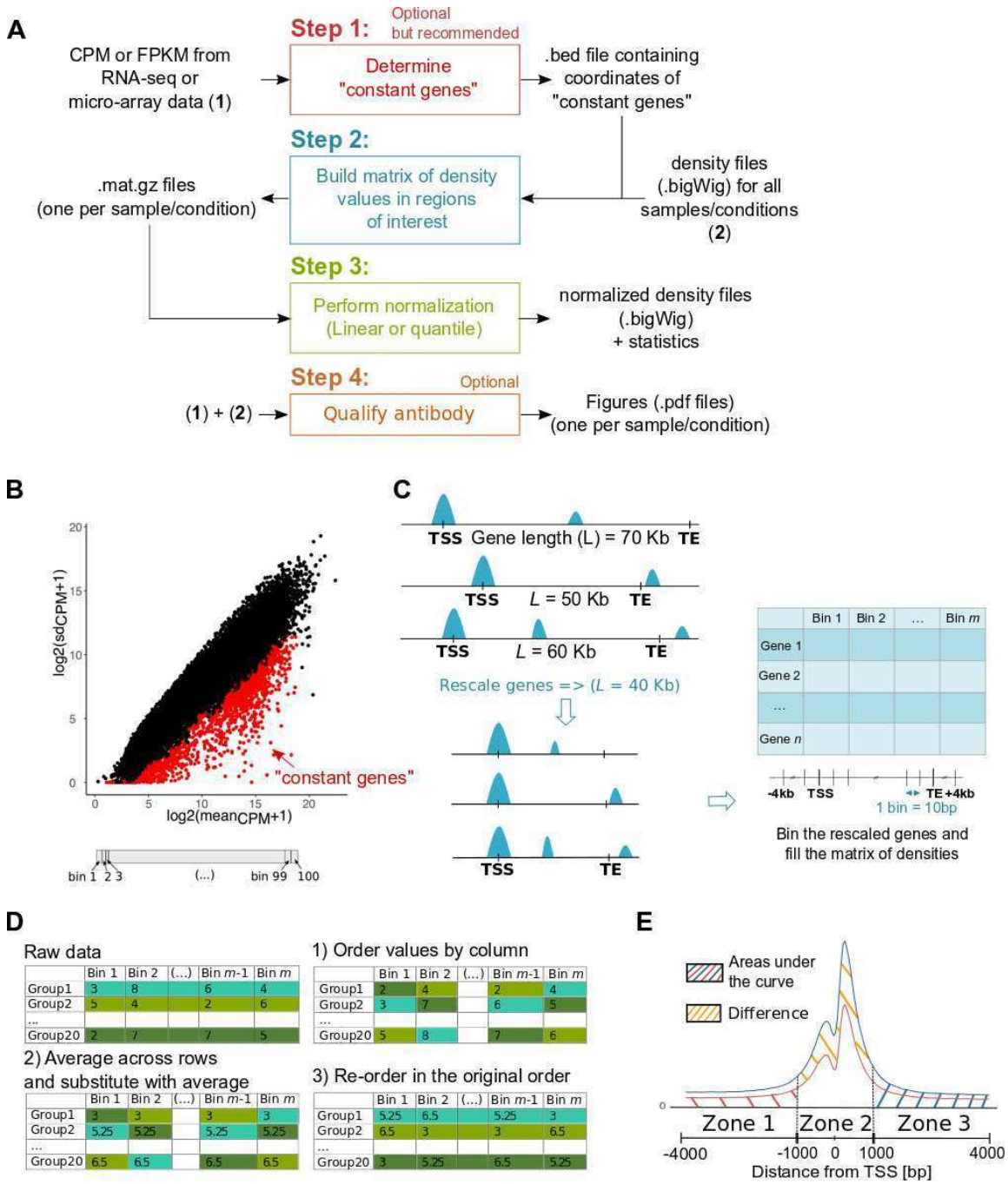


FIGURE 3.1 – **Aperçu des méthodes développées dans CHIPIN. A.** Approche en quatre étapes implémentée dans CHIPIN. **B.** Définition des "gènes constants" (indiqués en rouge) utilisés pour inférer les paramètres de normalisation : les "gènes constants" représentent 10% (par défaut) des gènes avec l'écart-type le plus faible du nombre de lectures. Ils sont sélectionnés dans différentes gammes d'expression génique (100 groupes par défaut). **C.** Les régions corps de gène (gene body) des "gènes constants" et les régions flanquantes (+/- n kb, n=4 par défault) sont redimensionnés (taille de la région redimensionnée par défaut : 40kb) et segmentées en bin (taille des bins par défaut : 10bp). **D.** Principales étapes de la normalisation quantile. Dans CHIPIN, la normalisation quantile est appliquée séparément à des sous-ensembles de gènes dans une certaine plage de signal d'intensité ChIP-seq (k groupes, k = 20 par défaut). **E.** Différence en terme d'aire sous les courbes de densité utilisée pour l'évaluation du succès du processus de normalisation.

3.1 Méthodes

3.1.1 Détermination des gènes pour établir les paramètres de normalisation

La méthode est basée sur l'hypothèse selon laquelle, en moyenne, aucune différence dans les signaux de ChIP-seq ne doit être observée dans les régions régulatrices de gènes dont l'expression est constante entre les échantillons ou conditions. Ces gènes peuvent être déterminés par le package en utilisant des données d'expression (RNA-seq ou micro array) (Fig. 3.1A). Pour cela, on calcule la moyenne et l'écart-type des valeurs de Count Per Million (CPM) pour chaque gène à partir des valeurs de RNA-seq ou des valeurs de microarrays à la puissance 2. Nous supposons que les gènes qui ont les plus petites valeurs d'écart type sont les gènes dont l'expression varie le moins.

Pour extraire un pourcentage égal de gènes "constants" dans chaque plage de valeurs d'expression, nous divisons tous les gènes en 100 groupes de taille égale selon les valeurs moyennes de l'expression des gènes. Nous définissons les gènes les moins variables, "constants", comme les gènes qui présentent les plus petites valeurs d'écart-type entre les échantillons/conditions (par défaut : 10%) dans chaque groupe d'expression (Fig. 3.1B les "genes constants" sont indiqués en rouge). Le résultat de cette étape est un fichier .bed standard avec les coordonnées des gènes. Outre le format CPM, l'utilisateur peut fournir des données d'expression sous la forme de Fragments par kilobase par million de lectures (FPKM) ; dans ce cas, notre méthode transformera les comptages FPKM en CPM en utilisant les annotations de transcription disponibles (longueur exons).

3.1.2 Calcul de la densité dans les régions régulatrices

Les fichiers .bigWig, fournis par l'utilisateur, qui contiennent l'information d'intensité de signal pour chaque échantillon et le fichier .bed contenant les coordonnées des "gènes constants" sont utilisés. CHIPIN calcule la densité du signal à travers les "gènes constants" et leur régions flanquantes (+/- n Kb, n=4 par défaut) pour chaque échantillon (Fig. 3.1C). Pour cette étape, nous utilisons la fonction *computematrix* fournie par le package *deeptools* [134]. La fonction *computematrix* offre la possibilité de redimensionner tous les gènes à la même longueur (par défault : 40kb). Les régions du corps de gène redimensionnées et leurs régions flanquantes en amont et en aval du corps de gène sont ensuite segmentées en bin (taille de bin par défault : 10bp). Enfin, le signal moyen par bin pour chaque gène est calculé. Le résultat de cette étape est une matrice par échantillon/condition avec l'élément correspondant au signal de densité cumulé par bin et par gène (Fig. 3.1C).

3.1.3 Normalisation

Pour chaque échantillon/condition, la matrice obtenue par Deeptools [134] est utilisée pour déterminer les paramètres de normalisation. L'utilisateur a le choix entre deux stratégies de normalisation : la normalisation quantile et la régression linéaire, décrites ci-dessous. CHIPIN fournit également des indicateurs permettant d'illustrer le succès de la procédure de normalisation : (i) des statistiques montrant la différence relative entre les courbes moyennes du signal avant et après la procédure de normalisation ; (ii) la visualisation du signal autour des TSSs des gènes, dont l'expression ne varie pas, avec et sans normalisation.

Régression linéaire

Le but de la normalisation est de rendre tous les échantillons comparables entre eux, ainsi s'il y a plus de deux échantillons, l'un des échantillons est choisi comme référence (par défaut : échantillon dont la valeur moyenne est la plus proche de la médiane de toutes les valeurs moyennes de chacun des échantillons). Nous calculons séparément les coefficients de régression pour chaque échantillon en utilisant le profil de densité de référence, puis nous normalisons les profils de densité à l'aide de ces coefficients. Plus précisément, pour chaque échantillon et pour la référence, nous calculons d'abord le signal moyen par bin à partir des matrices de densité calculées dans la section 3.1.2 (Fig. 3.2A). Ensuite, nous effectuons la régression linéaire avec interception non nulle sur ces valeurs moyennes d'intensité du signal (3.1) versus l'échantillon de référence (3.2).

$$\bar{S}_{current\ sample} \tag{3.1}$$

$$\bar{S}_{reference\ sample} \tag{3.2}$$

Pour chaque échantillon, une régression linéaire avec interception non nulle fournit α et β qui minimisent la somme des moindres carrés (équation 3.3). Etant donné α et β, le profil de densité pour chaque condition est alors corrigé (équation 3.4).

$$\bar{S}_{current\ sample} = \alpha\ \bar{S}_{reference\ sample} + \beta + \epsilon \tag{3.3}$$

$$S_{current\ sample\ corrected} = \frac{S_{current\ sample} - \beta}{\alpha} \tag{3.4}$$

Cette transformation est appliquée sur les valeurs d'intensité du signal de ChIP-seq orginales disponibles dans le fichier .bigWig fourni par l'utilisateur. Un nouveau fichier .bigWig, normalisé par rapport à tous les autres échantillons, est alors écrit pour chaque échantillon.

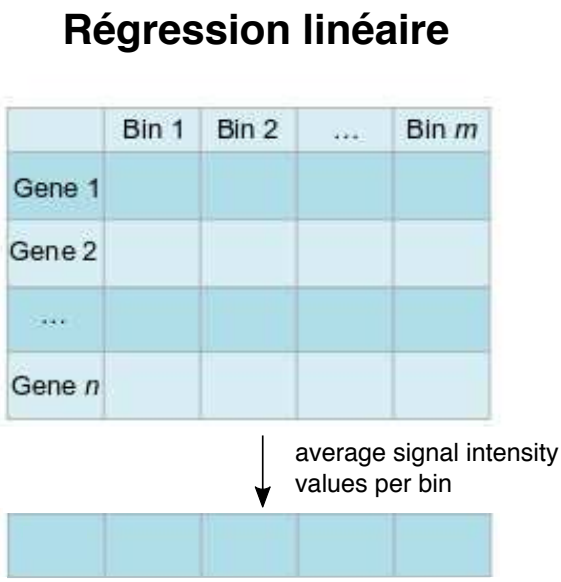


FIGURE 3.2 – **Illustration du processus de normalisation par régression linéaire.** Calcul des valeurs moyennes d'intensité du signal par bin sur lesquelles sera effectuée la régression linéaire.

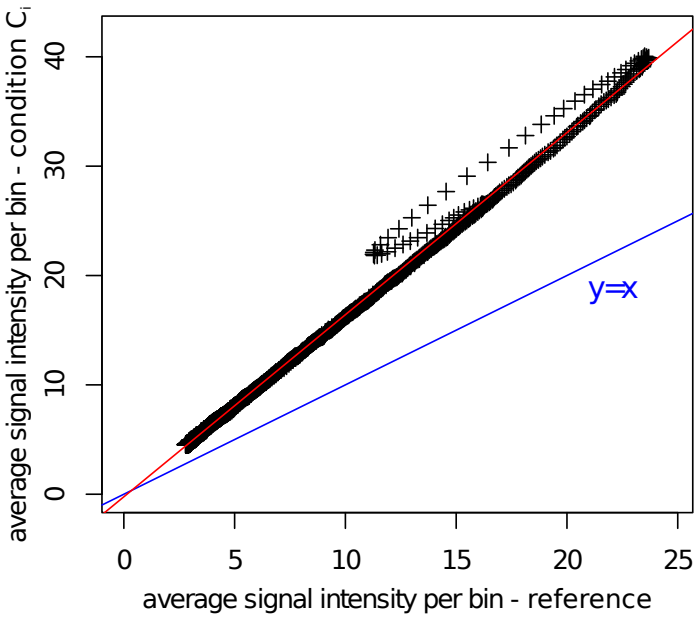


FIGURE 3.3 – **Résultat de la régression linéaire par CHIPIN.** Résultat de la régression linéaire réalisée par CHIPIN sur les valeurs moyennes de l'intensité du signal pour la condition C_i par rapport à la référence définie comme l'échantillon dont la valeur moyenne est la plus proche de de la médiane de toutes les valeurs moyennes des autres échantillons. En rouge est représentée la droite de régression linéaire construite à partir de la fonction *lm* du package stats de R.

Normalisation quantile

Pour cette stratégie de normalisation, chaque matrice d'intensité du signal de ChIP-seq calculée par deeptools dans la section 3.1.2 est triée en fonction de la valeur totale du signal de chaque gène dans les bins (Fig. 3.1C). Comme tous les niveaux d'intensité de fixation présents dans la matrice doivent être représentés de façon équivalente dans la procédure de normalisation, pour chaque échantillon/condition nous construisons k groupes de gènes (défaut : k=20) correspondant à k intensités de signal de ChIP-seq différentes, (Fig. 3.4B). La normalisation quantile (Fig. 3.1D) est effectuée sur les valeurs moyennes de ces k groupes (Fig. 3.4B) et la fonction de normalisation est appliquée aux valeurs d'intensité du signal de ChIP-seq originales disponibles dans le fichier .bigWig fourni par l'utilisateur. Un nouveau fichier .bigWig, normalisé par rapport à tous les autres échantillons, est alors écrit pour chaque échantillon.

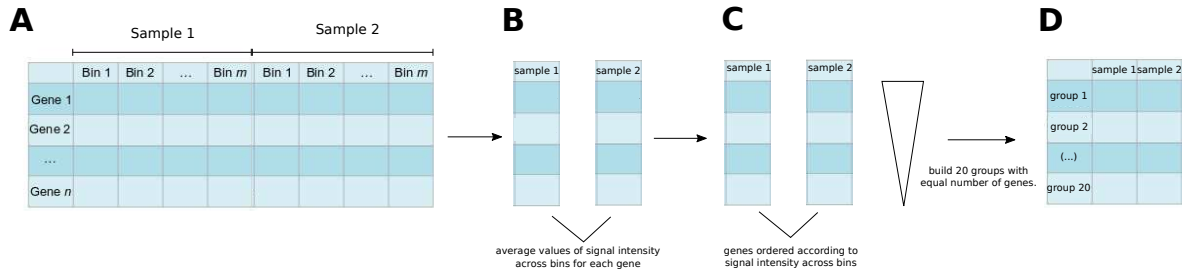


FIGURE 3.4 – **Illustration du processus précédant la normalisation quantile sur deux échantillons A.** Matrices issues de l'étape de calcul de la densité dans les régions régulatrices pour les deux échantillons. **B.** Calcul du signal moyen de l'intensité du signal dans les bins pour chaque gène. **C.** Les gène sont ordonnés de façon décroissante en fonction de leur valeur moyenne d'intensité du signal dans les bins. k groupes sont construits avec le même nombre de gènes dans chacun des groupes (k=20 par défault). **D.** Matrice contenant autant de colonnes qu'il y a d'échantillons (ici 2) et autant de lignes qu'il y a de groupes (ici 20) sur laquelle sera effectuée la normalisation quantile.

3.1.4 Intensité ChIP-seq en fonction de l'expression génique

Le package CHIPIN offre la possibilité de visualiser l'intensité du signal de ChIP-seq autour des TSSs des gènes en fonction de l'expression de ces derniers. Les données d'expression de tous les gènes (valeurs RPKM ou valeurs de microarray) sont séparés en trois groupes en utilisant l'algorithme de clustering des k-means : gènes hautement, moyennement et faiblement exprimés. Le signal ChIP-seq autour des TSS de ces trois groupes de gènes est visualisé à l'aide d'un profil de densité (voir 3.2.3 pour un exemple d'application).

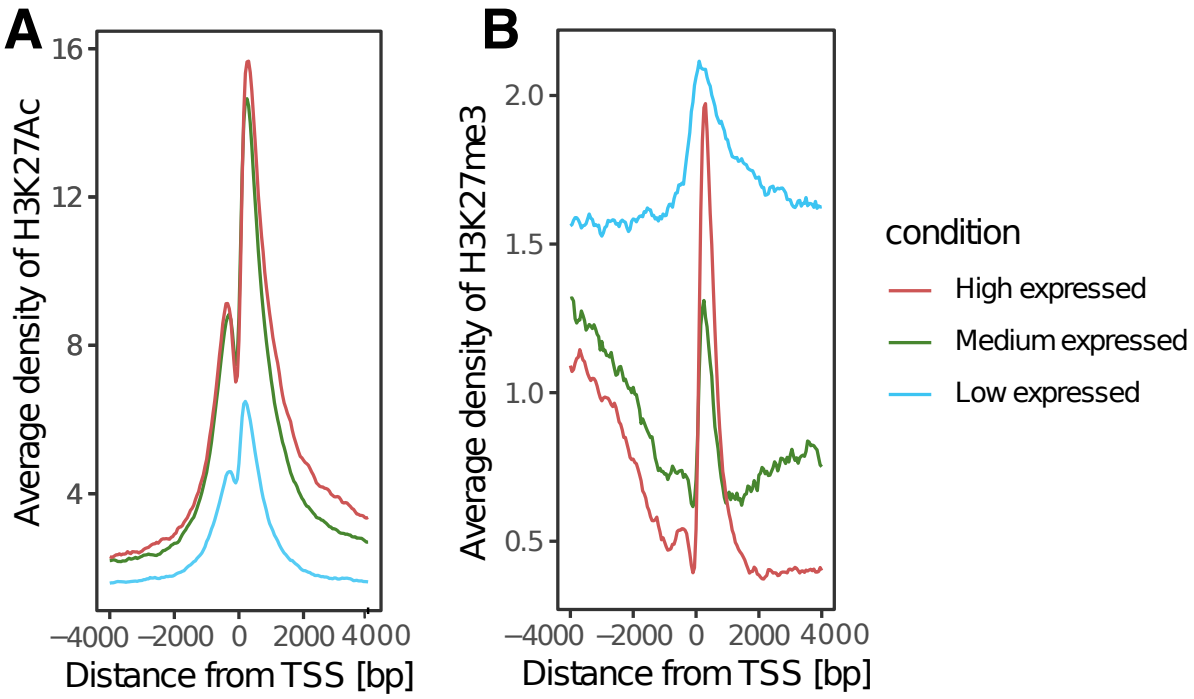


FIGURE 3.5 – Profiles générés par CHIPIN pour évaluer la spécificité de deux anticorps : **A.** anticorps ab4729 pour H3K27Ac dans un échantillon de carcinome adrénocortical (non publié), et **B.** anticorps ACM39155 pour H3K27me3 dans des cellules shSpi1-A2B (non publié). Pour ab4729 **(A)**, comme attendu, les gènes hautement exprimés montrent la plus haute densité de H3K27ac ; tandis que pour ACM39155 **(B)**, les gènes hautement exprimés présentent une augmentation de H3K27me3 après la position TSS. Cela suggère une fixation potentiellement non spécifique de l'anticorps contre H3K27me3 documenté également par [152]

3.2 Résultats

J'ai comparé CHIPIN à trois autres méthodes : (i) normalisation au même nombre de lectures, (ii) la méthode LILY basée sur l'appariement du signal à l'intérieur des pics communs les plus forts [140] et (iii) ChIPSeqSpikeInFree, une méthode de normalisation très récente basée sur le calcul de la pente de la courbe du nombre cumulé de lectures [139]. La normalisation au même nombre de lectures a été choisie car c'est la méthode la plus utilisée dans les analyses de ChIP-seq. Les deux autres méthodes, bien qu'elles ne couvrent pas tout le spectre des approches de normalisation sont représentatives des deux hypothèses utilisées dans les stratégies de normalisation : la première basée sur les régions où le signal est le plus fort, l'autre basée sur le signal global. Les procédures de normalisation de chacune des méthodes ont été appliquées sur les densités de ChIP-seq, contenues dans les fichiers .bigwig obtenus en utilisant HMCan [124].

3.2.1 Validation de l'efficacité de la méthode pour la correspondance de profils de densité

Nous avons évalué la performance de CHIPIN sur deux jeux de données :

— ChIP-seq H3K27ac cinq échantillons de carcinome cortico-surrénaliens humains.

— ChIP-seq H3K27ac pour la lignée cellulaire de souris transgénique pour le gène *Spi1* (shSpi1-A2B) dans deux conditions : (i) Spi1 surexprimé (Spi1+) ; (ii) Spi1 réprimé (Spi1-) (non publié).

Echantillons de carcinomes cortico-surrénaliens humains

J'ai appliqué les deux normalisations de CHIPIN (normalisation quantile et régression linéaire) et les trois autres méthodes aux densités de ChIP-seq pour les cinq échantillons de carcinome cortico-surrénaliens humains. Pour évaluer l'efficacité de la procédure de normalisation, nous avons calculé et comparé les valeurs moyennes de la densité de ChIP-seq autour des TSS des "gènes constants" avant et après la normalisation. Contrairement aux trois autres méthodes, la normalisation linéaire et la normalisation quantile de CHIPIN ont permis de supprimer les différences entre les valeurs de densité moyenne de H3K27ac autour des TSSs de ces gènes dont l'expression varie peu (Fig. 3.6).

J'ai quantifié les différences entre les courbes des cinq échantillons avant et après chaque procédure de normalisation (Tab. 3.1). Pour cela, la région de 8kb autour des TSSs des "gènes constants" a été divisée en trois zones : (i) zone 1 : [-4Kb, -1Kb[, (ii) zone 2 :[-1Kb, +1Kb], (iii) zone 3 :]+1Kb, +4Kb] (Fig. 3.1E) et j'ai calculé l'aire sous chaque courbe dans les trois différentes zones (Fig. 3.1E aires hachurées en rouge et bleu). L'écart entre chaque échantillon/condition a été évalué par la différence moyenne en pourcentage de surface sous les courbes de densité (Fig. 3.1E aire hachurée en orange, Tab. 3.1). Dans la région centrale (zone 2) correspondant aux régions promotrices et participant à la régulation de l'expression des gènes, CHIPIN a obtenu de bien meilleures performances que les autres méthodes. En effet, la différence initiale a été réduite par CHIPIN à seulement 1.8% (normalisation quantile), alors que la meilleure parmi les trois méthodes testées permet de réduire à 18% de différence moyenne entre les profils normalisés.

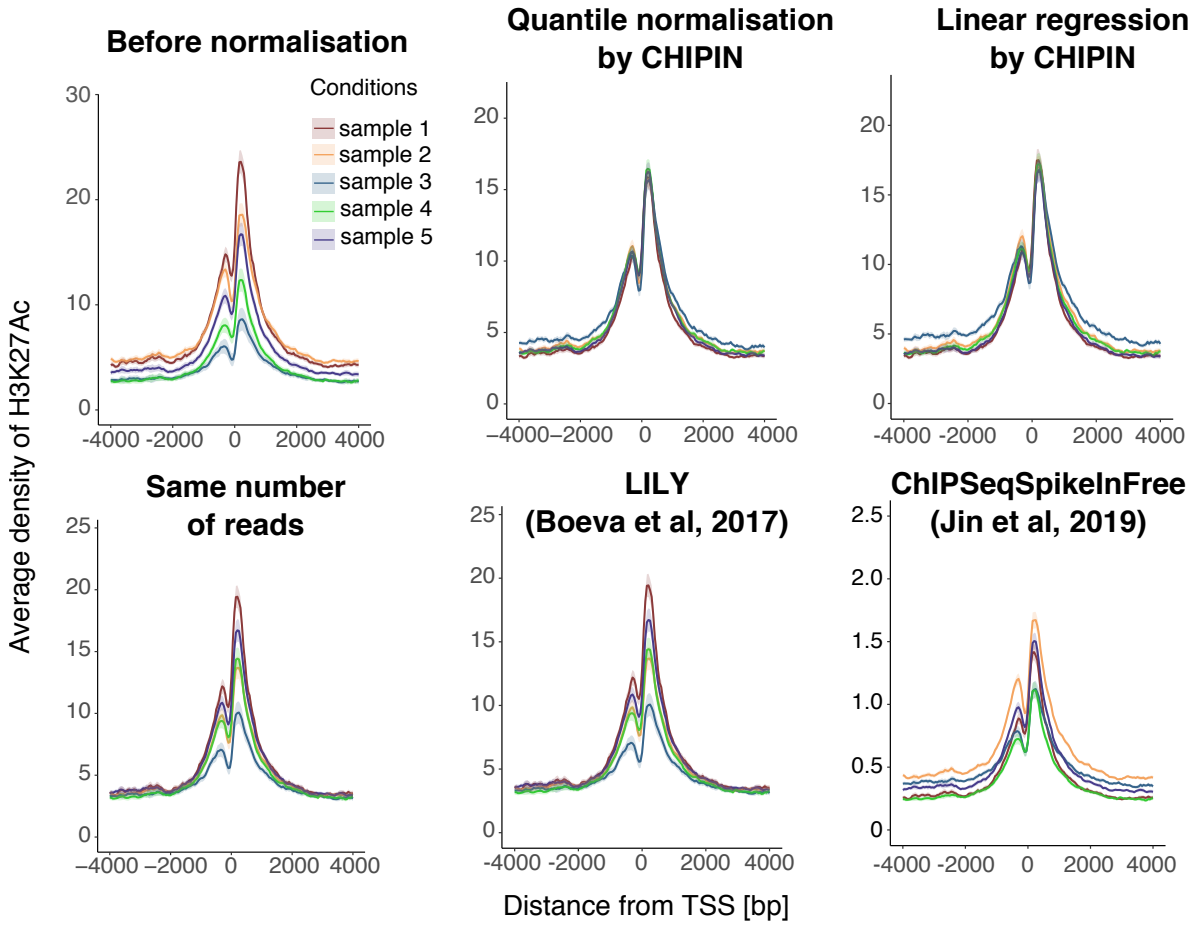


FIGURE 3.6 – **Comparaison de l'efficacité de la normalisation sur cinq échantillons de carcinomes cortico-surrénaliens pour la marque H3K27ac.** Densité moyenne de signal autour des TSSs des gènes dont l'expression ne varie pas. Le premier profil de densité correspond au signal moyen sans normalisation.

	Zone 1	**Zone 2**	**Zone 3**
Before normalisation	27.5%	34.75%	28.25%
Quantile Normalisation by CHIPIN	15.3%	1.8%	7.9%
Linear Regression by CHIPIN	21%	4.5%	19.2%
Same number of reads	7.5%	23%	6.5%
LILY [140]	6.5%	18%	6.7%
ChIPSeqSpikeInFree [139]	29%	27%	29.5%

TABLE 3.1 – **Pourcentage de différence entre les courbes de densité moyenne (Fig. 3.6) dans les trois zones autour des TSS des "gènes constants" (Fig. 3.1).**

Lignée cellulaire de cellules de souris shSpi1-A2B

Pour valider davantage les performances de CHIPIN, nous avons analysé le second jeu de données consistant en la lignée cellulaire de souris shSpi1-A2B pour laquelle la marque d'his-

tone H3K27ac a été profilée dans deux conditions : (i) Spi1 surexprimé : Spi1+ et (ii) Spi1- réprimé : Spi1-. Pour chaque condition, nous avons deux réplicats techniques : réplicat 1 et 2. Pour montrer l'importance de la correction de l'effet de lot ("batch effect"), nous avons utilisé la combinaison : réplicat 1 pour condition Spi1+ et réplicat 2 pour condition Spi1-. Ainsi, j'ai appliqué les deux méthodes CHIPIN (régression linéaire et normalisation quantile) et les trois autres méthodes aux densités de H3K27ac calculées dans les deux conditions Spi1+ et Spi1- (Fig. 3.7, Tab. 3.2).

CHIPIN permet de faire correspondre les deux courbes de densité moyenne autour des TSS des "gènes constants" presque parfaitement en utilisant à la fois la normalisation quantile et la régression linéaire (0.06% et 0.1% de différence après la normalisation contre 38% de différence avant la normalisation pour la zone centrale 2). L'application des trois autres méthodes a également permis de réduire la différence des signaux ChIP-seq entre les deux conditions (de 38% à 7.6%, 17.2% et 11.9% pour même nombre de lectures, LILY et ChIPSeqSpikeInFree respectivement), sans toutefois atteindre les performances de CHIPIN.

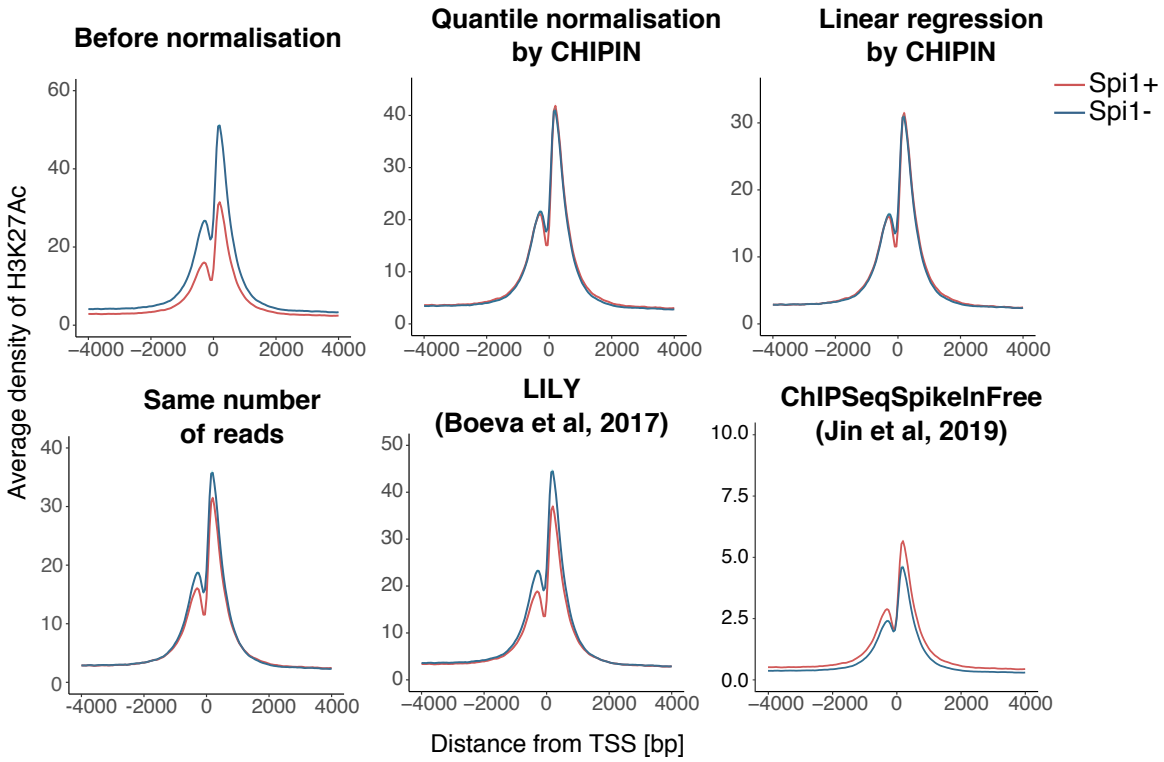


FIGURE 3.7 – **Comparaison de l'efficacité de la normalisation sur la lignée cellulaire de souris shSpi1-A2B (Réplicats croisés).** Densité moyenne de signal autour des TSSs des "gènes constants". Le premier plot correspond au signal moyen sans normalisation. Anticorps : Abcam 4729 (ab4729) contre H3K27ac ; conditions : "Spi1+" - *Spi1* sur-exprimé, "Spi1-" - *Spi1* réprimé par un shRNA inducible par la doxycycline (non publié).

	Zone 1	Zone 2	Zone 3
Before normalisation	32%	38%	30%
Quantile Normalisation by CHIPIN	4.6%	0.06%	8.6%
Linear Regression by CHIPIN	1.5%	0.1%	5.3%
Same number of reads	2.8%	7.6%	2.3%
LILY [140]	6.8%	17.2%	1.9%
ChIPSeqSpikeInFree [139]	15.6%	11.9%	18.2%

TABLE 3.2 – **Pourcentage de différence en terme d'aire hachurée entre les courbes de la Fig. 3.7 dans les trois zones différentes (Fig. 3.1E).**

Nous avons également comparé CHIPIN aux trois autres méthodes en utilisant le même réplicat pour les deux conditions (Fig. 3.8A, Tab. 3.3). CHIPIN surpasse également les autres méthodes de normalisation (de 40% à 0.7%, 0.6%). L'application des trois autres méthodes a également permis de réduire la différence des signaux ChIP-seq entre les deux conditions (de 38% à 3.6%, 15.6% et 11.9% pour même nombre de lectures, LILY et ChIPSeqSpikeInFree respectivement), sans atteindre les performances de CHIPIN dans ce cas non plus.

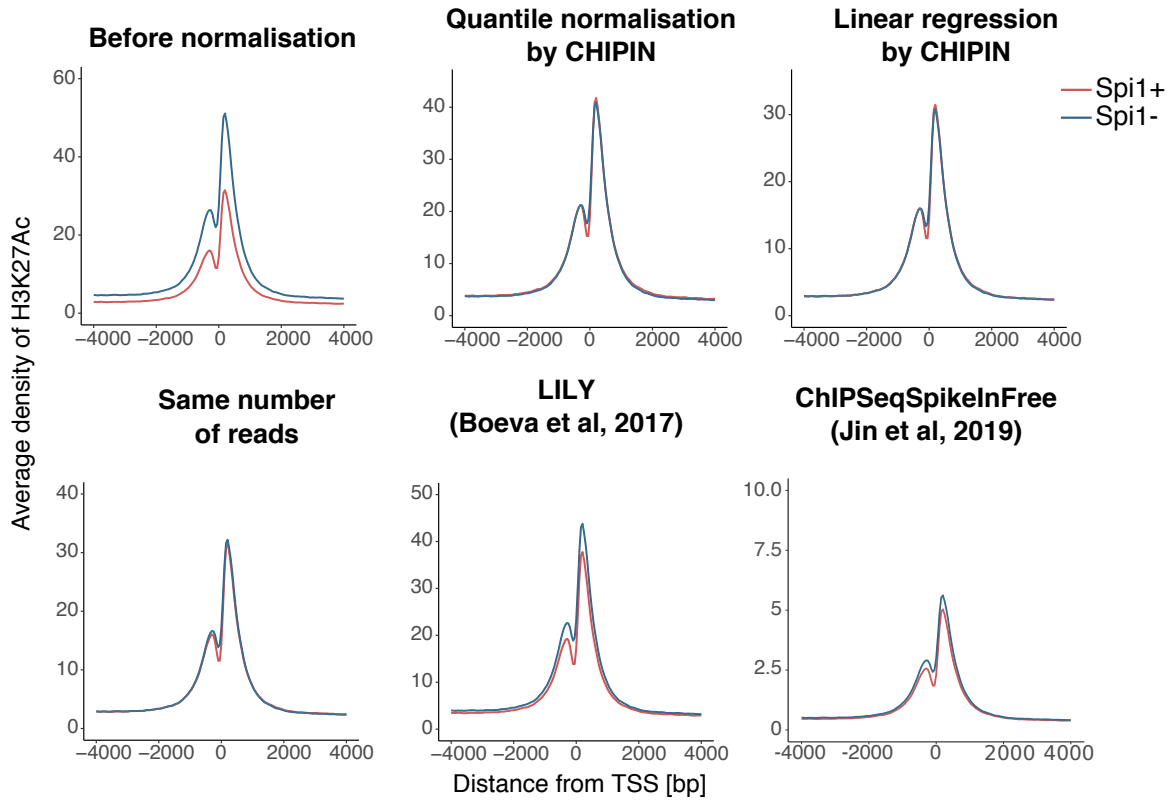


FIGURE 3.8 – **Visualisation de l'efficacité de la normalisation par CHIPIN sur la lignée cellulaire shSpi1-A2B (Réplicat 1).** Anticorps : Abcam 4729 (ab4729) pour H3K27ac ; conditions : "Spi1+" - *Spi1* sur-exprimé, "Spi1-" - *Spi1* réprimé par un shRNA (non publié). Densité moyenne de signal autour des TSSs des gènes dont l'expression ne varie pas. Le premier profil de densité correspond au signal moyen sans normalisation.

	Zone 1	**Zone 2**	**Zone 3**
Before normalisation	37%	40%	35%
Quantile Normalisation by CHIPIN	3.6%	0.7%	6.5%
Linear Regression by CHIPIN	0.09%	0.6%	2.6%
Same number of reads	0.7%	3.6%	3.4%
LILY [140]	12.2%	15.6%	9.6%
ChIPSeqSpikeInFree [139]	15.7%	11.9%	18.2%

TABLE 3.3 – **Pourcentage de différence en terme d'aire hachurée entre les courbes de la Fig. 6.1 dans les trois zones différentes (Fig. 3.1E).**

3.2.2 CHIPIN préserve les différences biologiques des signaux de densité de ChIP-seq entre les conditions

Dans la section précédente, j'ai montré que la normalisation CHIPIN surpasse les trois procédures de normalisation testées. En effet, les autres méthodes ne parviennent pas à faire correspondre les courbes Spi1+ et Spi1- autour des TSSs des "gènes constants". Toutefois comme la procédure de normalisation de CHIPIN utilise exclusivement les régions génomiques entourant les "gènes constants", nous avons évalué l'efficacité de la méthode pour maintenir les différences biologiques dans le signal ChIP-seq pour les gènes qui sont exprimés de manière différentielles entre les conditions. Nous avons extrait les gènes différentiellement exprimés entre les conditions Spi1+ et Spi1- dans la lignée cellulaire de souris shSPi1-A2B (gènes activés par SPI1 : $FC > 1.5$, $pval_{ajuste} < 0.05$, gènes réprimés par SPI1 : $FC < 0.6666$, $pval_{ajuste} < 0.05$) puis nous avons analysé les profils de densité de la marque d'histone activatrice H3K27ac au voisinage des TSSs de ces gènes activés ("UpReg") ou réprimés ("DownReg") par SPI1 après application des cinq procédures de normalisation testées (Fig. 3.9).

Nous pensions que pour les gènes UpReg par SPI1 la normalisation correcte se traduirait par un signal H3K27ac plus fort dans la condition Spi1+ par rapport au signal dans la condition Spi1-. De même, pour les gènes DownReg par SPI1, nous attendions un signal H3K27ac plus fort dans la condition Spi1-. De plus, nous nous attendions à ce que les différences absolues entre les courbes de densités moyennes entre les deux conditions seraient comparables pour les gènes UpReg et pour les gènes DownReg.

Parmi les cinq méthodes testées, seules les deux méthodes implémentées dans CHIPIN fournissent les résultats escomptés compte tenu de la signification du signal ChIP-seq analysé (Fig. 3.9A - réplicats croisés, Fig. 3.10A - réplicat 1). Au contraire, la normalisation par ChIPSeqSpikeInFree a entraîné des différences de densité entre les deux conditions beaucoup plus importantes dans les gènes UpReg que dans les gènes DownReg. Nous avons quantifié nos observations sur les différences entre les profils de densité H3K27ac normalisés par les cinq techniques en représentant la différence relative pour les gènes DownReg sur l'axe des x et pour les gènes UpReg sur l'axe des y (Fig. 3.9B - réplicats croisés, Fig. 3.10B - réplicat 1). Plus chaque

observation est proche de la diagonale (y=x), plus les différences absolues entre les conditions Spi1+ et Spi1- sont similaires pour les gènes UpReg et DownReg et plus les profils de densité résultants sont proches de nos attentes biologiques. Les deux approches de normalisation par CHIPIN correspondent aux points les plus proches de la diagonale (Fig. 3.9B). Nous avons donc conclu que CHIPIN est plus performant que les trois autres méthodes testées. Toutefois, lorsque nous avons utilisé le même réplicat dans les deux conditions (Spi1+ et Spi1-) nous constatons que la normalisation au même nombre de reads a également donné des résultats presque parfaits (Fig. 3.10).

Si la normalisation a été bien effectuée, la différence entre les deux courbes autour des TSSs des gènes Up-régulés et Down-régulés devrait être similaire, sinon cela signifie que la normalisation n'a pas été bien effectuée. Premièrement, pour la méthode LILY et la normalisation au même nombre de reads, la différence entre les courbes Dox0 et Dox1 est beaucoup plus élevée autour des TSSs des gènes DownReg par SPI1 qu'autour des TSSs des gènes UpReg par SPI1 (Fig. 3.9A). Deuxièmement, pour la méthode ChIPSeqSpikeInFree, la normalisation entraîne une différence beaucoup plus importante autour des TSSs des gènes UpReg par SPI1 qu'autour des TSSs des gènes DownReg par Spi-1 (figure 5A). Enfin, les deux méthodes développées dans CHIPIN donnent les meilleurs résultats, car autour des TSSs des gènes UpReg et DownReg par SPI1, la différence entre les courbes Dox0 et Dox1 est équivalente (Fig. 3.9A). J'ai quantifié la différence entre ces deux courbes sur un biplot montrant sur l'axe des x la différence pour les gènes DownReg et sur l'axe des y la différence pour les gènes UpReg (Fig. 3.9B). Pour quantifier les différences, j'ai calculé la différence des moyennes des deux courbes dans chacune des trois zones (voir la Fig. 3.1E). En gris pointillé est représentée la ligne y = x, la différence moyenne des gènes UpReg et DownReg pour chacune des cinq méthodes testées était représentée par un point coloré (Fig. 3.9B). Plus un point est proche de la droite y = x, plus la différence entre les courbes Dox0 et Dox1 est similaire dans les deux cas, UpReg et DownReg. Ainsi, les points représentant les meilleures méthodes doivent être sur ou très proches de la droite y = x. Les deux méthodes de normalisation de CHIPIN sont celles pour lesquelles les points étaient les plus proches de la droite y = x ; pour les zones 1 et 2, pour la zone 3 ces deux points tombent sur la courbe y = x. Ainsi, CHIPIN a largement surpassé les trois autres méthodes testées.

L'efficacité de CHIPIN a été démontrée sur deux ensembles de données différents. Sur ces deux ensembles de données différents, CHIPIN a fait correspondre les courbes des différentes conditions / échantillons autour des TSSs des "gènes constants" utilisés pour inférer les paramètres de normalisation, alors que les trois autres méthodes n'ont pas réussi à le faire. En particulier, nous avons utilisé les données de la lignée cellulaire shSpi-1 A2B, dans laquelle des réplicats différents pour les deux conditions ont été utilisés afin de tenir compte de l'effet de lot, pour comparer la densité aux TSSs des gènes Up-régulés et Down-régulés par SPI1, non utilisés pour entraîner la méthode. Nous avons montré que la normalisation effectuée par CHIPIN était la seule méthode qui permet d'obtenir la différence attendue au niveau des TSSs des gènes

UpReg et DownReg par SPI1.

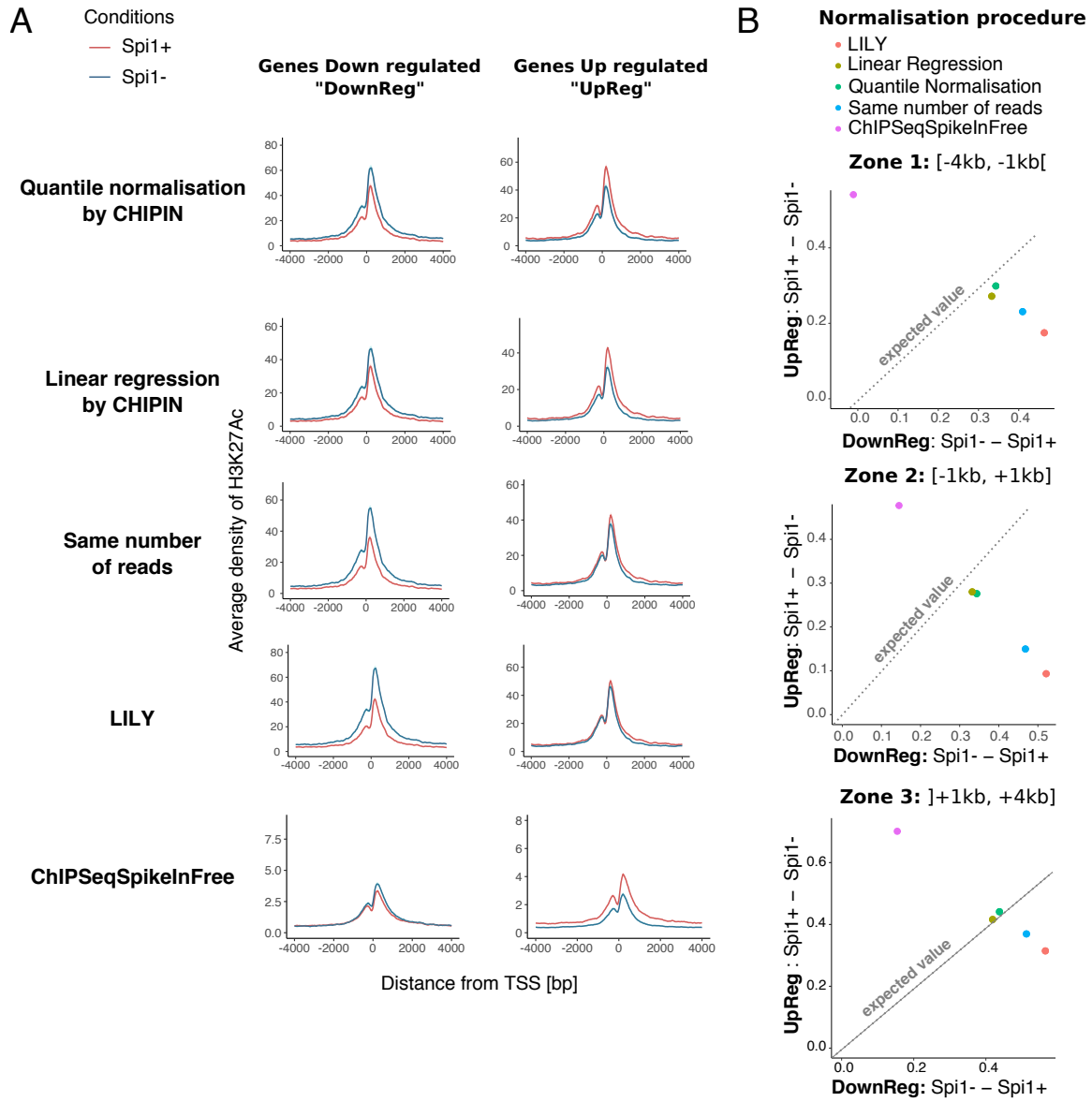


FIGURE 3.9 – **Comparaison de l'efficacité de la normalisation par CHIPIN sur les gènes différentiellement exprimés dans les cellules shSpi1-A2B (Réplicats croisés). A** Profils de densité de H3K27ac autour des TSS des gènes Down-régulés (DownReg) et Up-régulés (UpReg) par SPI1 dans deux conditions : "Spi1+" - *Spi1* surexprimé, "Spi1-" : *Spi1* réprimé. **B** Différence dans le signal H3K27ac pour les zones 1-3 (Fig. 3.1E) dans les régions promotrices des gènes up- et down-régulés par SPI1. Les axes x et y montrent la différence entre les conditions Spi1+ et Spi1-. Des procédures de normalisation correctes doivent donner des observations proches de la diagonale y=x en pointillés gris.

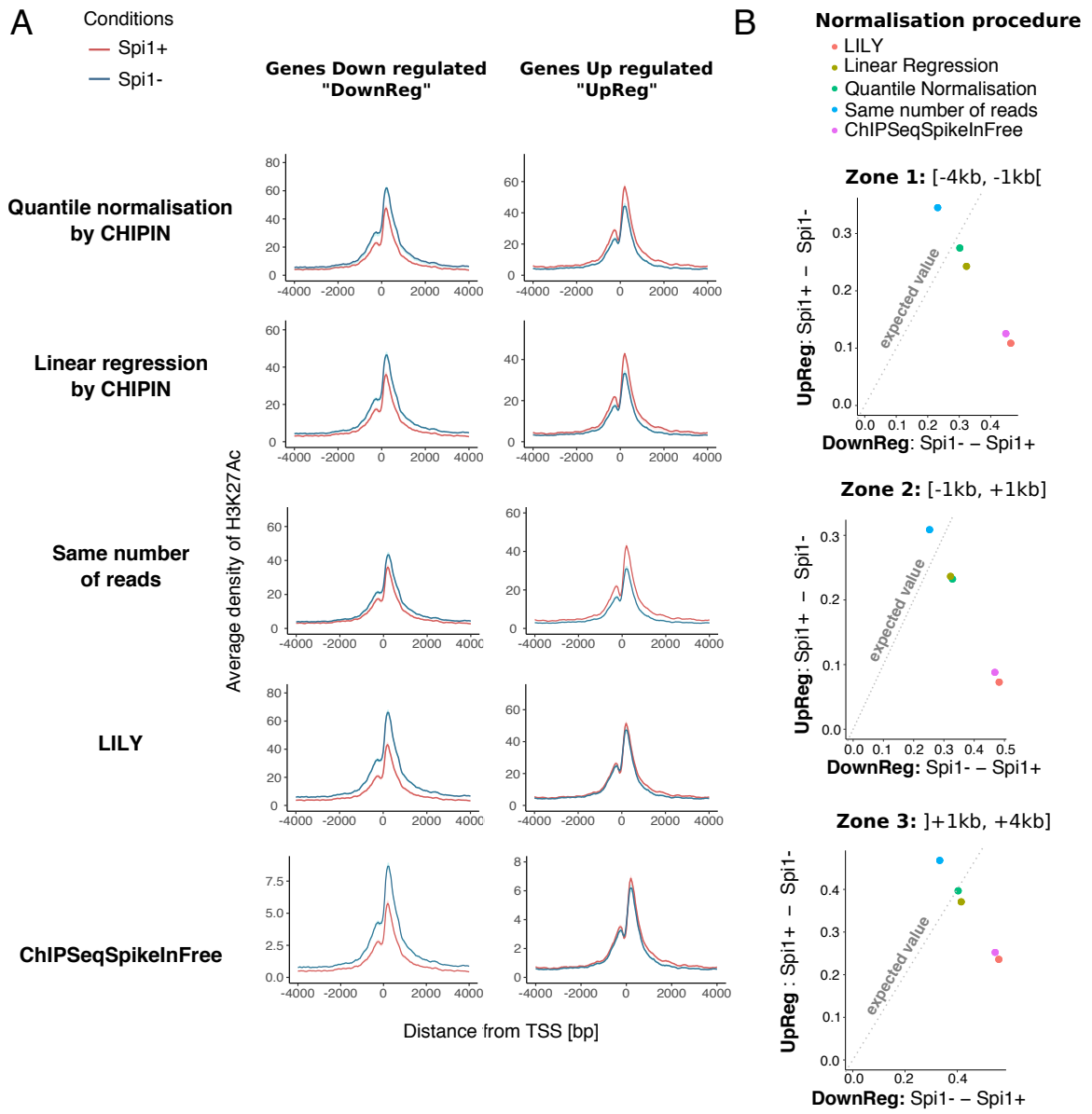


FIGURE 3.10 – **Comparaison de l'efficacité de la normalisation par CHIPIN sur les gènes différentiellement exprimés dans les cellules shSpi1-A2B (Réplicat 1). A** Profils de densité de H3K27ac autour des TSS des gènes Down-régulés (DownReg) et Up-régulés (UpReg) par SPI-1 dans deux conditions : "Spi1+" - *Spi1* surexprimé, "Spi1-" : *Spi1* réprimé. **B** Différence dans le signal H3K27ac pour les zones 1-3 (Fig. 3.1E) dans les régions promotrices des gènes up- et down-régulés par SPI1. Les axes x et y montrent la différence entre les conditions Spi1+ et Spi1-. Des procédures de normalisation correctes doivent donner des observations proches de la diagonale y=x en pointillé gris.

3.2.3 Qualification de la spécificité de l'anticorps utilisé

La qualité d'une expérience de ChIP-seq repose beaucoup sur l'efficacité et la spécificité de l'anticorps utilisé. En effet, l'utilisation d'anticorps non spécifiques peut conduire à des

conclusions biologiques erronées. Ainsi, évaluer la densité moyenne du signal de ChIP-seq autour des TSSs des gènes en fonction de l'expression de ces gènes est un bon moyen de vérifier que les tendances biologiques connues sont respectées (Fig. 3.5). Pour cela, on peut utiliser des informations apriori connues sur les effets de la liaison de la protéine d'intérêt sur l'expression des gènes : les gènes fortement exprimés ont tendance à avoir un signal plus fort des marques d'histones activatrices (par exemple, H3K27ac et H3K4me3) et du facteur de transcription se liant dans leurs promoteurs. H3K27ac est une marque d'histone activatrice qui a un signal élevé aux TSSs des gènes hautement exprimés, tandis que les gènes éteints ou faiblement exprimés ont tendance à avoir un signal de marques d'histones répressives plus fort (par exemple : H3K27me3 et H3K9me3) à proximité de leurs TSSs [84].

On fournit ici un exemple de détection de la fixation non-spécifique d'un anticorps par la fonction *plot_expression* du package CHIPIN (Fig. 3.5).

Nous avons appliqué CHIPIN à deux expériences de ChIP-seq : H3K27ac dans des échantillons de carcinome cortico-surrénaliens (anticorps Abcam ab4729) et H3K27me3 dans les cellules shSpi1-A2B (anticorps Active Motif ACM39155). Les données de ChIP-seq ont été couplées à des expériences de RNA-seq pour évaluer l'expression des gènes. CHIPIN a permis la génération de profils de densité moyenne autour des TSSs pour trois groupes de gènes : gènes faiblement, moyennement et hautement exprimés (Fig. 3.5).

Comme prévu, pour les échantillons de carcinomes cortico-surrénaliens la densité de la marque H3K27ac autour des TSSs corrèle avec l'expression des gènes, confirmant la spécificité et la haute qualité de l'anticorps utilisé (Fig. 3.5A). Il est toutefois surprenant de constater que dans les cellules shSpi1-A2B, les gènes hautement exprimés (Fig. 3.5B en rouge) montrent un signal de densité moyenne très élevée pour H3K27me3 dans la région du gène immédiatement en aval du TSS (1kb après la position TSS). Cette augmentation est également observée dans une certaine mesure pour les gènes moyennement et faiblement exprimés (Fig. 3.5BB en vert en bleu respectivement) et était probablement liée à une liaison non spécifique de l'anticorps ACM29155 attendu pour cibler la méthylation de H3K27 dans cette expérience. En effet, une étude récente des spécificités des anticorps a montré qu'en plus de la marque répressive ciblée H3K27me3, l'anticorps ACM29155 pouvait également se lier à plusieurs marques d'activateur telles que les lysines acétylées [152]. Cet exemple a démontré l'utilité de CHIPIN dans l'évaluation des spécificités des anticorps et dans l'évaluation de la qualité d'un jeu de données de ChIP-seq.

Chapitre 4

Mécanismes moléculaires pour la répression génique par SPI1 dans la différenciation érythroïde

4.1 Méthodes

L'analyse du RNA-seq +/- dox a été réalisée par M'boyba Khadija Diop (Institut Gustave Roussy). J'ai obtenu les fichiers .bam, les valeurs de RPKM et de DESeq2 pour chaque gène ainsi que l'information sur la régulation des gènes par SPI1 grâce au calcul du FoldChange (FC). Ainsi, la liste des gènes activés par SPI1 (FC $\geqslant$ 1.5 et pvalue $\leq$ 0.05), c'est à dire les gènes dont l'expression est augmentée lorsque SPI1 est surexprimé, la liste des gènes réprimés par SPI1 (FC < 0.67 et pvalue $\leq$ 0.05), c'est à dire ceux dont l'expression est diminuée lorsque SPI1 est surexprimé est disponible. J'ai également obtenu la liste des gènes NoResp (0.9 $\geqslant$ FC $\geqslant$ 1.1 et pvalue > 0.05), ce sont les gènes dont l'expression ne change pas en fonction de la présence ou de l'absence de SPI1.

L'analyse du RNA-seq +/- Entinostat a été réalisée par Stéphanie Legras (Plateforme GenomEast, IGBMC). Les mêmes seuils que pour le RNA-seq +/- dox ont été appliqués pour déterminer les gènes dont l'expression était activée ou réprimée par HDAC1.

4.1.1 Analyse des données de ChIP-seq (protéines histones et RNAPolII) et des données d'ATAC-seq

Alignement des lectures sur le génome

Pour les données de ChIP-seq des différentes marques d'histone ainsi que pour les données d'ATAC-seq (Tab. 2.1), les lectures ont été alignées sur un assemblage du génome de la souris (GRCm38/mm10) en utilisant l'algorithme *mem* du logiciel BWA [122]. L'option -t concernant

le nombre de coeurs utilisés a été mise à 7 et le mode paired-end fut utilisé quand cela était nécessaire, pour les données d'ATAC-seq et pour toutes les ChIP-seq des marques d'histone sauf H3K36me3 et RNAPolII. Un contrôle qualité de l'alignement des lectures a été effectué, pour cela un seuil de 20 a été appliqué sur le champ MAPQ des fichiers bam. Les duplicats (de PCR) ont été retirés.
La commande utilisée dans le cas de données paired-end est la suivante :

```
> bwa mem  -t 7 Mouse/mm10/BWA_indexes/mm10.fa
rep_histoneMark_condition_R1.fastq histoneMark_R2.fastq |
samtools view - -S -b > rep_histoneMark_condition.bam
```

Recherche des sites de liaison et normalisation des signaux de densité

La recherche des sites de liaison (peak calling) a été effectué en utilisant le logiciel HMCan [124] qui permet de corriger pour le biais en GC ainsi que pour le biais en nombre de copies. Certains paramètres du logiciel HMCan sont communs à toutes les protéines étudiées (minLength=150, medLength=350, maxLength=600, smallBinLength=50bp, largeBinLength=100000, pvalueThreshold=0.05, blackListFile fourni). En revanche, le paramètre mergeDistance est différent suivant la protéine analysée. Ce paramètre permet de fusionner des pics qui sont proches (dont la distance est inférieure à une certaine valeur). En effet, les marques d'histone H3K4me1, H3K27ac, H3K4me3 et la protéine RNAPolII peuvent couvrir 1 à 10 nucléosomes consécutifs (mergeDistance égal à 200) tandis que des marques d'histone telles que H3K27me3 ou H3K36me3 peuvent couvrir de grandes régions génomiques, de dix à plusieurs centaines de kilobases (mergeDistance égal à 1000 pour H3K36me3 et à 3000 pour H3K27me3). La valeur de mergeDistance pour la marque d'histone H3K27me3 a fait l'objet d'une fine régulation car les données de H3K27me3 chez la souris sont généralement très compliquées à analyser. En effet, le signal n'est pas aussi net que celui de la marque d'histone H3K27ac. Le seuil de p-valeur a été fixé à 0.05 pour l'ensemble des marques d'histone.
En plus des fichiers contenant les coordonnées génomiques des pics (.bed), HMCan construit des fichiers .bigWig représentant le niveau de la protéine d'intérêt sur l'ensemble du génome. L'un des objectifs de l'analyse des données de ChIP-seq est de comparer le comportement des différentes marques d'histone et de RNAPolII entre la condition SPI1 surexprimé et SPI1 réprimé, pour cela il faut normaliser les fichiers .bigWig entre les deux conditions. J'ai utilisé CHIPIN (section 3). Cette méthode repose sur l'hypothèse selon laquelle les gènes dont l'expression ne varie pas entre les conditions Spi1+ et Spi1- ne doivent pas présenter de différence dans leurs régions régulatrices dans le signal des différentes marques d'histone. Ainsi, pour inférer les paramètres de normalisation, les gènes NoResp, déterminés en utilisant les données de RNA-seq Spi1+/Spi1-, ont été utilisés.

Les fichiers .bigWig d'ATAC-seq ont été obtenus par HMCan [124] en utilisant des paramètres spécifiques aux données d'ATAC-seq. En effet, la correction du GC-content n'est pas réalisée et le paramètre smallBinLength est égal à 10bp de façon à augmenter la résolution, cette valeur est également utilisée pour les facteurs de transcription. Le package CHIPIN (section 3) est également utilisé pour la procédure de normalisation des fichiers .bigWig d'ATAC-seq.

4.1.2 Analyse des données de ChIP-seq SPI1+

Les données de ChIP-seq pour la protéine SPI1 que j'ai utilisées sont des données déjà publiées [37]. Les sites de liaison de SPI1 avaient déjà été définis précédemment sur le génome de la souris (assemblage NCBI37/mm9, n=17781) ont été transférés sur un assemblage plus récent (GRCm38/mm10, n=17778) en utilisant l'outil *liftOver* [153]. Les pics de SPI1 ont ensuite été annotés en utilisant la fonction *annotatePeak* du package R ChIPSeeker [154]. La commande utilisée est la suivante :

```
> annotatedPeaks=annotatePeak(ToAnnotate, tssRegion=c(-2000, +1000),
TxDb=txdb, annoDb="org.Mm.eg.db", sameStrand=FALSE, overlap="all")
```

Les différentes régions d'annotation sont représentées sur la Fig. 4.1 et détaillées ci-dessous :

- **région intergénique** : de -50kb avant le TSS à -2kb avant le TSS.
- **région promotrice** *(paramètre tssRegion)* : de -2kb avant le TSS à +1kb après le TSS, découpée en trois sous régions différentes (Fig. 4.1).
- **région corps de gène (gene body)** : de +1kb après le TSS jusqu'au TES [1]. Bien que la terminologie biologique définit le corps de gène (gene body) du TSS au TE, on utilise ici une définition modifiée car la région qui va du TSS à +1kb après le TSS correspond à la région ImmediateDownStreamTSS.
- **région DownStream** : du TES jusqu'à +3kb après le TES.

La classe TxDb permet, sous R, de stocker des informations de transcrits. Sa création passe par l'utilisation de la fonction *makeTxDbFromGFF* du package GenomicFeatures [155]. Pour l'annotation on utilise l'assemblage gencode VM16 [156], la plus récente lorsque j'ai commencé ma thèse. La position utilisée pour annoter les pics de SPI1 est la position **peakmax** définie comme la position génomique où le signal est le plus haut pour chacun des pics de SPI1. Elle a été déterminée par FindPeaks dans [37] et nous l'avons conservée. Ainsi, à chacun des pics de SPI1 est associé une position **peakmax** que l'on peut définir comme la position la plus vraisemblable pour SPI1 de se situer sur chacun de ses pics. Lorsque plusieurs isoformes étaient disponibles pour annoter un pic, l'isoforme était choisi par ChIPseeker selon la hiérarchie suivante : Promoter > Gene body > DownStream > Intergenic. Dans le cas où il existe plusieurs

1. **TES :** **T**ranscription **E**nd **S**ite

isoformes avec des TSS différents pour une même annotation, le TSS choisi est celui dont la distance à la position peakmax de SPI1 est la plus faible.

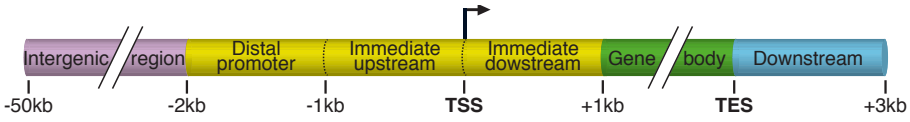


FIGURE 4.1 – **Modèle de gène utilisé pour l'annotation des pics de SPI1** Régions génomiques d'annotation : (i) Intergénique de -50kb avant le TSS à -2kb avant le TSS (violet) ; (ii) Promoteur de -2kb avant le TSS à +1kb après le TSS (jaune) ; (iii) Corps de gène de +1kb après le TSS jusqu'au TES (bleu) ; (iv) Downstream de -3kb avant le TES jusqU'au TES (bleu).

4.1.3 Définition des régions distales cis-régulatrices (enhancers)

La régulation de l'expression génique se fait par la liaison de protéines régulatrices sur les promoteurs et les régions enhancers. Afin de définir ces régions, j'ai utilisé la distribution de la marque activatrice H3K27ac et de la marque H3K4me1 (en dehors des régions promotrices) et le logiciel ChromHMM [157]. Le logiciel ChromHMM [157] fait appel à un modèle de Markov caché multivarié (HMM) permettant la modélisation explicite de la présence ou de l'absence combinatoire de chaque marque. Les états cachés du modèle correspondent à l'état de la chromatine. Ce terme englobe à la fois la nature probabiliste d'un modèle à plusieurs états et la nature biologique de l'état de la chromatine à chaque position génomique en dehors des régions promotrices. ChromHMM prend en entrée les coordonnées des pics des marques d'histone que l'utilisateur veut utiliser pour construire son paysage d'états épigénétiques. Je n'ai utilisé que H3K27ac et H3K4me1 qui sont suffisantes pour définir les régions cis-régulatrices en dehors des régions promotrices. Les pics de H3K27ac et H3K4me1 ont été déterminés en utilisant le logiciel HMCan [124], à chaque pic est associée une valeur de FDR, celle-ci est calculée en utilisant la largeur des pics appelés dans l'échantillon Input par rapport à la largeur des pics déterminés dans l'échantillon ChIP. Je n'ai conservé que les pics dont le FDR était inférieur à 0,01 afin d'obtenir majoritairement des pics qui sont de vrais positifs.
Le nombre d'états cachés du modèle de Markov caché est l'un des paramètres de ChromHMM. Nous avons demandé trois états car la marque H3K27ac est très rarement trouvé sans H3K4me1.

- co-occurence de H3K27ac et H3K4me1 : enhancers actifs (**E3**).
- présence de H3K4me1 seulement : enhancers inactifs (**E2**).
- absence des deux marques suggérant ainsi que ces régions ne sont pas des régions enhancers (**E1**).

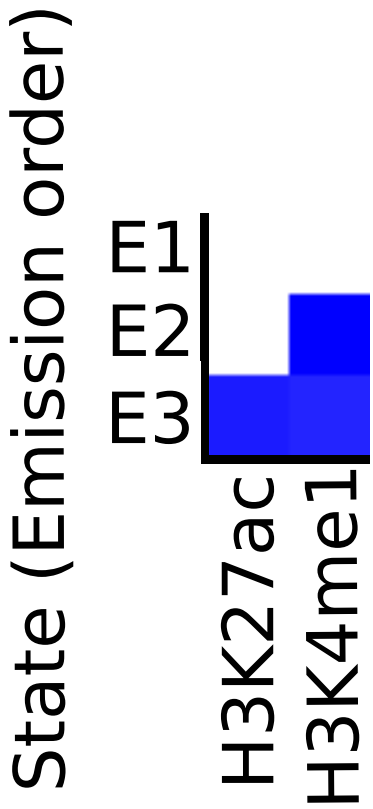


FIGURE 4.2 – Emissions états ChromHMM

L'ensemble du génome (hors régions promotrices) est couvert par les émissions des états E1, E2, E3 (Fig. 4.2). Cette segmentation du génome permet d'analyser la répartition des pics de SPI1 (en dehors de ceux annotés promoteurs) dans ces régions E1, E2 et E3. Pour cela, j'utilise la position peakmax de chacun des pics de SPI1 à laquelle 1000bp sont ajoutées en amont et en aval et je regarde quels sont les états de la chromatine (E1, E2, E3) qui chevauchent cette région de 2000bp. Chaque pic de SPI1 est alors défini dans un état chromatinien particulier. Il est possible qu'une région de 2000bp représentant un pic de SPI1 chevauche plusieurs états différents, ainsi les règles suivantes ont été définies :

— co-occurence des états E3 et E2 $\rightarrow$ E3
— co-occurence des états E3 et E1 $\rightarrow$ E3
— co-occurence des états E2 et E1 $\rightarrow$ E2
— co-occurence des états E3 et E2 et E1 $\rightarrow$ E3

4.1.4 Recherche de motifs

Le motif de SPI1 est déjà connu, en effet il a été déterminé par Ridinger-Saison et al [37], la recherche de potentiels co-facteurs déjà connus est réalisée en utilisant l'outil *AME* de la suite

MEME [132]. J'ai utilisé l'option par défaut de AME pour les motifs en input, cela consiste à mélanger les lettres des séquences données en entrée. La collection de motifs "MOUSE DNA" de la database "HOCOMOCCO MOUSE (v11 CORE)" a été utilisée. Cet outil identifie des motifs connus qui sont relativement enrichis dans les séquences que l'utilisateur donne en entrée. Dans la suite MEME [132] il y a également l'outil CentriMo qui permet d'identifier des motifs connus qui sont relativement enrichis dans les séquences que l'utilisateur donne en entrée et qui présentent une préférence significative pour une position particulière dans l'ensemble des séquences d'entrée.
Pour certaines analyses, j'ai eu besoin de définir toutes les occurrences du motif de liaison du facteur de transcription GATA1 dans certaines régions génomiques. Pour cela j'ai utilisé l'outil FIMO (**F**ind **I**ndividual **M**otif **O**ccurences) de la suite MEME [132], celui-ci permet à partir d'un fichier contenant le motif au format MEME de déterminer toutes les occurrences de ce motif dans les régions génomiques données sous la forme d'un fichier .fasta.
Pour faire la recherche de motifs de facteurs de transcription déjà connus dans les pics de SPI1, la position peakmax de SPI1 est utilisée et 1000bp sont rajoutées en amont et en aval de cette position peakmax. En réalisant la recherche de motifs de cette façon, je fais l'hypothèse que le potentiel co-facteur de SPI1 est dans son voisinage proche.

4.1.5 Profils de densité

Certains résultats sont présentés sous la forme de profils de densité dans des régions génomiques d'intêret (exemples : autour des TSS des gènes réprimés par SPI1, autour des pics de SPI1 dans les gènes réprimés ...). Ces profils de densité ont été réalisés en utilisant le language (R version 3.6.2). Les fichiers .bigWig contenant l'information de densité de ChIP sont chargés sous R sous la forme de GenomicRanges en utilisant le package *rtracklayer*. Les fichiers .bed contenant les régions d'intérêt dans lesquelles construire les profils de densité sont également chargés sous R sous la forme de GenomicRanges. La densité de ChIP dans les régions d'intérêt est obtenue en utilisant la fonction *findOverlaps* entre les deux objets GRanges. La visualisation graphique est réalisée en utilisant les fonctions *ggplot*, *geom_line* et *geom_ribbon* du package *ggplot2*. Les packages *plotrix*, *tidyr* sont également utilisés pour la visualisation des variances autour des courbes.

Les régions d'intérêts sont construites autour d'une position particulière : la position peakmax de SPI1 lorsqu'on s'intéresse aux densités autour des pics de SPI1, la position TSS lorsqu'on s'intéresse aux densités autour des TSSs, autour de ces positions les régions sont ensuite étendues de +/- x bp (en général x=2000 ou x=4000 bp). Les positions TSSs des gènes sur lesquels SPI1 se fixe sont directement extraites de l'annotation réalisée à l'aide de ChIPseeker [154]. En effet, à chaque pic est associé l'information "distance au TSS" grâce à laquelle on retrouve la position du TSS de l'isoforme du gène sur lequel le pic a été annoté. Si parmi les

régions d'intérêt, plusieurs pics se trouvent sur le même gène alors seuls les isoformes différents sont conservés lors de la construction du profil de densité aux TSSs associés.

Lors de la construction des profils de densité au niveau des pics de SPI1 se trouvant dans des régions enhancers inactifs (E2), les profils de densité aux enhancers et aux TSSs associés ne contient que les positions TSSs des gènes pour lesquels il n'existe pas de pic de SPI1 dans une région enhancer actif (E3) afin de s'extraire de l'influence d'un pic de SPI1 dans une région enhancer actif sur la densité au TSS associé.

4.1.6 Définition des régions différentielles pour H3K27ac

Au cours de ce travail, la marque d'histone H3K27ac a beaucoup suscité mon intérêt, en effet autour des pics de SPI1 et aux TSSs associés des variations dans la densité de cette marque en fonction de la présence ou de l'absence de SPI1 ont été repérées. Ainsi, nous avons été amenés à définir l'état d'acétylation de certains pics de SPI1. En particulier tous les pics de SPI1 qui se trouvaient dans un état E3, c'est à dire dans un environnement avec les marques d'histone H3K4me1 et H3K27ac, ont été séparés en fonction du différentiel de H3K27ac dans leur voisinage entre les conditions Spi1+(dox-) et Spi1-(dox+). Pour cela, j'ai utilisé la position peakmax de SPI1 +/- 500bp et le différentiel de H3K27ac a été évalué sur ces régions de 1000bp. Cette évaluation a été réalisée à partir des fichiers .bigWig normalisés par la méthode CHIPIN (chapitre 3).

Les pics de SPI1 se trouvant dans les gènes réprimés par SPI1 dans la région corps de gène ou intergénique sont considérés comme les plus différentiels et seront nommés $H3K27ac^{Diff}$ si ils respectent les conditions suivantes (Pour la définition des pics de SPI1 se trouvant dans les gènes activés par SPI1, on utilisera l'équation 4.1 sur le niveau moyen de H3K27ac en condition SPI1+ et non SPI1- ainsi que l'équation 4.2 avec le rapport inversé.) :

$$\frac{\sum_{i=1}^{n} \left(H3K27ac^{Spi1-}\right)_i}{n} > 5 \tag{4.1}$$

$$\frac{\frac{\sum_{i=1}^{n} \left(H3K27ac^{Spi1-}\right)_i}{n}}{\frac{\sum_{i=1}^{n} (H3K27ac^{Spi1+})_i}{n}} > 1.5 \tag{4.2}$$

avec :

- *i* représentant l'ensemble des *n* bins de 50bp couverts par la région de 1000bp.
- $(H3K27ac^{Spi1+,Spi1-})_i$ représentant le densité de H3K27ac pour une condition donnée dans le bin *i*.

Les pics de SPI1 se trouvant dans les gènes réprimés par SPI1 dans les régions corps de gène ou intergénique sont considérés comme non-différentiels et nommés $H3K27ac^{NoDiff}$ si ils respectent les conditions données par les équations 4.1 ainsi que les conditions suivantes (pour la

définition des pics de SPI1 se trouvant dans les gènes activés par SPI1, on utilisera les équations 4.1, 4.3 ainsi que l'équation 4.4 avec le rapport inversé.) :

$$\frac{\sum_{i=1}^{n} \left(H3K27ac^{Spi1+}\right)_i}{n} > 5 \tag{4.3}$$

$$0.83 \leq \frac{\frac{\sum_{i=1}^{n} \left(H3K27ac^{Spi1+}\right)_i}{n}}{\frac{\sum_{i=1}^{n} (H3K27ac^{Spi1-})_i}{n}} \leq 1.2 \tag{4.4}$$

Les seuils dans les équations 4.1, 4.3 ont été définis à partir de la distribution des niveaux moyens de H3K27ac dans les régions de +/-500bp autour de la position peakmax de SPI1 dans la condition Spi1+ (Fig. 4.3A). En effet, la distribution est bi-modale et l'on choisit un seuil égal à 5 afin de garder tous les pics de SPI1 qui se situent dans la deuxième partie de la distribution. On illustre également les valeurs choisies pour définir les pics de SPI1 se situant dans des régions différentielles ($H3K27ac^{Diff}$, équation 4.2) ou non ($H3K27ac^{NoDiff}$, 4.4) pour H3K27ac (Fig. 4.3B). Sur ce MA-plot, ne sont représentées que les gènes réprimés dans un souci de lisibilité. La droite verte en pointillés (Fig. 4.3B) correspond à la valeur $log2(1.5)$ (équation 4.2), tous les points au-dessus de cette droite sont considérés comme correspondant à des pics de SPI1 dans un contexte différentiel pour H3K27ac. Les droites rouges pointillés (Fig. 4.3B) encadrent les valeurs $log2(0.83)$ et $log2(1.2)$ et les points se trouvant entre ces deux droites sont considérés comme correspondant à des pics de SPI1 dans un contexte non différentiel pour H3K27ac.

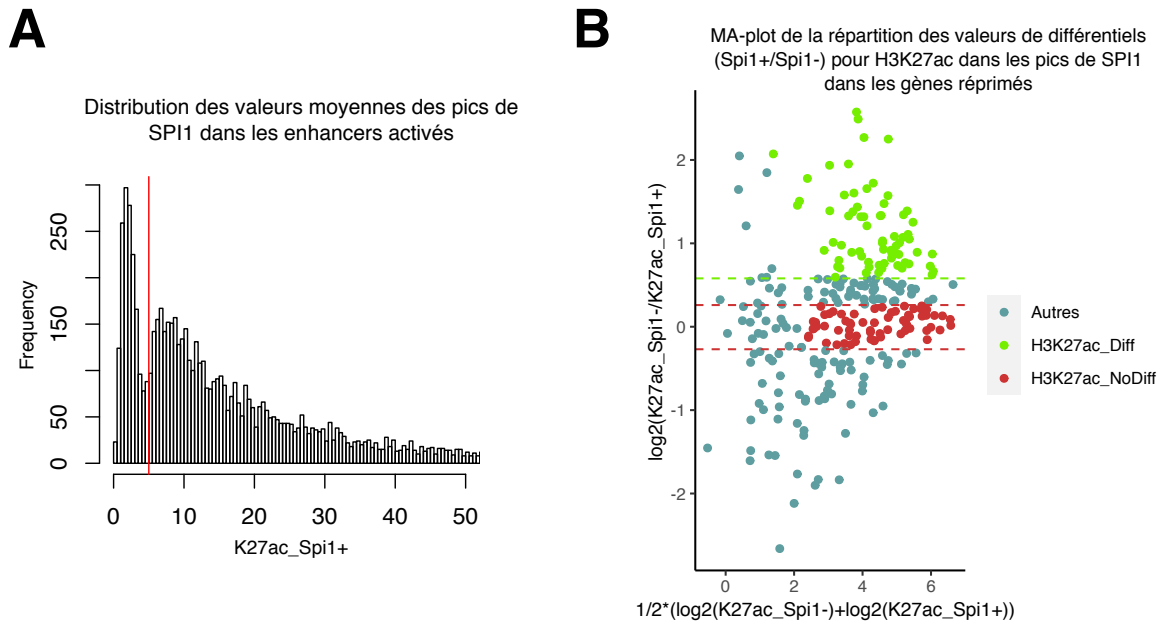


FIGURE 4.3 – **Illustration de la localisation des seuils choisis dans les équations sur les niveaux de H3K27ac. A.** Distribution des valeurs moyennes des pics de SPI1 dans les enhancers activés. **B.** MA-plot de la répartition des valeurs des différentiels (conditions Spi1+/Spi1-) pour H3K27ac dans les pics de SPI1 dans les enhancers activés des gènes réprimés.

4.2 Résultats

Ce travail, pour lequel j'ai effectué l'analyse bioinformatique, fera l'objet d'un manuscrit en cours d'écriture. Je présente ici l'ensemble des résultats afin de replacer mon travail dans la problématique posée. Dans la section qui suit je présenterai l'ensemble des résultats de validation des analyses bioinformatiques que j'ai effectuées. Les sections qui suivent présentent les résultats sous la forme d'une histoire pour définir un mécanisme de répression génique par le facteur de transcription SPI1.

4.2.1 Validation des analyses bioinformatiques

Afin de caractériser le ou les mécanismes de répression exercé(s) par SPI1 dans les cellules pré-leucémiques pour lesquelles SPI1 est dans un cas surexprimé, et dans l'autre cas très faiblement exprimé, nous avons eu besoin de comparer ces deux conditions et intégrer des expériences de ChIP-seq, ATAC-seq et RNA-seq ensembles. Pour cela, une analyse bioinformatique initiale a été réalisée pour les expériences de ChIP-seq et ATAC-seq, celle ci comprend plusieurs étapes : alignement, peak-calling, normalisation. Dans un premier temps je présenterai les résultats de l'alignement et dans un second temps les résultats de la normalisation utilisant CHIPIN pour les différentes marques d'histones et pour l'ATAC-seq.

Le résultat de l'alignement des lectures des différentes expériences de ChIP-seq (marques d'histones et pour RNAPolII) ainsi que de l'expérience d'ATAC-seq est donné par le pourcentage de lectures alignées après filtrage sur la qualité et duplicats de PCR éliminés par rapport au nombre total de lectures initales Tab. 4.1. Certains pourcentages peuvent paraître faibles. Toutefois, ces pourcentages sont calculés à partir du nombre de lectures alignées après filtrage sur la qualité (Q20) et élimination des duplicats de PCR qui peuvent être relativement élévés et contribuer à la baisse de ces pourcentages.

Target protein	Condition	Inital number of reads	Percentage mapped reads after quality filtering
H3K27me3	Spi1+	169553232	67.31%
H3K27me3	Spi1-	132293302	72.76%
H3K27ac	Spi1+	70047996	76.75%
H3K27ac	Spi1-	63724152	76.44%
H3K27ac	DMSO	95753640	90.71%
H3K27ac	Entinostat	102826188	94.58%
H3K4me1	Spi1+	56597404	58.48%
H3K4me1	Spi1-	51404914	70.58%
H3K4me3	Spi1+	89075276	80.87%
H3K4me3	Spi1-	98235658	83.30%
H3K36me3	Spi1+	25712905	85.19%
H3K36me3	Spi1-	38177973	85.03%
RNAPolII	Spi1+	36953766	75.81%
RNAPolII	Spi1-	38592961	80.73%
ATAC-seq	Spi1+	88134206	80.58%
ATAC-seq	Spi1-	98751136	80.10%

TABLE 4.1 – **Résultats des alignements des données de ChIP-seq et ATAC-seq utilisées pour ce travail.**

Après l'étape d'alignement, le peak-calling réalisé à l'aide du logiciel HMCan [124], qui permet de s'extraire du biais du nombre de copies ainsi que du biais en contenu en GC, donne en sortie des fichiers .bigWig qui contiennent pour tout le génome, la densité de la protéine d'intêret dans le cas des expériences de ChIP-seq, la densité des régions d'ADN non enroulées autour des nucléosomes dans le cas de l'expérience d'ATAC-seq. Une étape de normalisation, en utilisant le package CHIPIN que j'ai développé au cours de ma thèse, permet de rendre les conditions comparables entre elles. Cela est indispensable dans le cadre de notre étude car nous voulons comparer le paysage épigénétique autour des régions de liaison de SPI1 ainsi qu'aux TSSs associés en absence et en présence de SPI1. Pour cela nous devons comparer les densités des modifications des protéines histones (acétylation, méthylation), encodées par les fichiers .bigwig, entre les conditions Spi1+ et Spi1-. Pour normaliser les données de ChIP-seq ainsi que les données d'ATAC-seq nous utilisons les gènes NoResp, c'est à dire les gènes dont l'expression ne varie pas entre les conditions Spi1+ et Spi1-, déterminés à l'aide des données de

RNA-seq. Pour nous assurer de l'efficacité de la normalisation, nous avons calculé et comparé les valeurs moyennes de densité de ChIP-seq (Fig. 4.4) et d'ATAC-seq (Fig. 4.5) autour des TSS des gènes NoResp avant et après la normalisation. La quantification des différences entre les courbes des deux conditions avant et après normalisation (comme décrit sur la figure Fig. 3.1) décrit une diminution moyenne de la différence entre les deux courbes de 25% (données non montrées).

De plus, le package CHIPIN permet de vérifier que l'anti-corps utilisé lors de l'expérience de ChIP-seq est bien spécifique de la protéine d'intêret. Pour cela, on peut utiliser des informations connues à priori sur le lien entre la liaison de la protéine d'intêret et l'expression des gènes. En effet, les gènes fortement exprimés auront tendance à avoir un signal plus fort de la marque d'histone activatrice H3K27ac mais aussi d'enzymes liées à la transcription (RNAPolII) tandis que les gènes faiblement exprimés auront tendance à avoir un signal plus fort de la marque d'histone répressive H3K27me3. Ainsi j'ai quantifié le niveau de RNAPolII, H3K27me3, H3K27ac en fonction de l'expression des gènes. Les gènes ont été séparés en utilisant l'algorithme des k-means (k=3) : gènes hautement exprimés (rouge), gènes moyennement exprimés (orange), gènes faiblement exprimés (bleu). Les densités de la marque activatrice H3K27ac et de l'enzyme RNAPolII sont bien corrélées avec les gènes hautement exprimés et sont proportionnelles au niveau d'expression des gènes (Fig. 4.6). La densité de la marque répressive est élevée au niveau des TSSs des gènes faiblement exprimés (Fig. 4.6). On remarque que cette marque est faible autour des TSS des gènes moyennement et fortement exprimés, sans disctinction pour ces deux catégories (Fig. 4.6).

L'expérience d'ATAC-seq permet de détecter les zones de chromatine décondensée, qui est, aux TSSs caractéristique des gènes exprimés. Un fort signal d'ATAC-seq sera associé à une région de chromatine décondensée. J'ai quantifié le niveau d'ATAC-seq autour des TSS des trois groupes de gènes (hautement, moyennement, faiblement exprimés) et l'on obtient, comme attendu, une différence de densité de signal entre les gènes les plus fortement exprimés et ceux les plus faiblement exprimés (Fig. 4.5). Néanmoins, ce résultat indique que la technique ne permet pas de distinguer les gènes hautement exprimés des gènes moyennement exprimés sur la base de la structure de la chromatine. Plusieurs raisons pourraient expliquer ce résultat :

— une limite de sensibilité de la technique d'ATAC-seq
— si le gène est transcriptionnellement actif, le niveau de décondensation autour des TSSs n'est pas proportionnel au niveau d'expression des gènes exprimés.

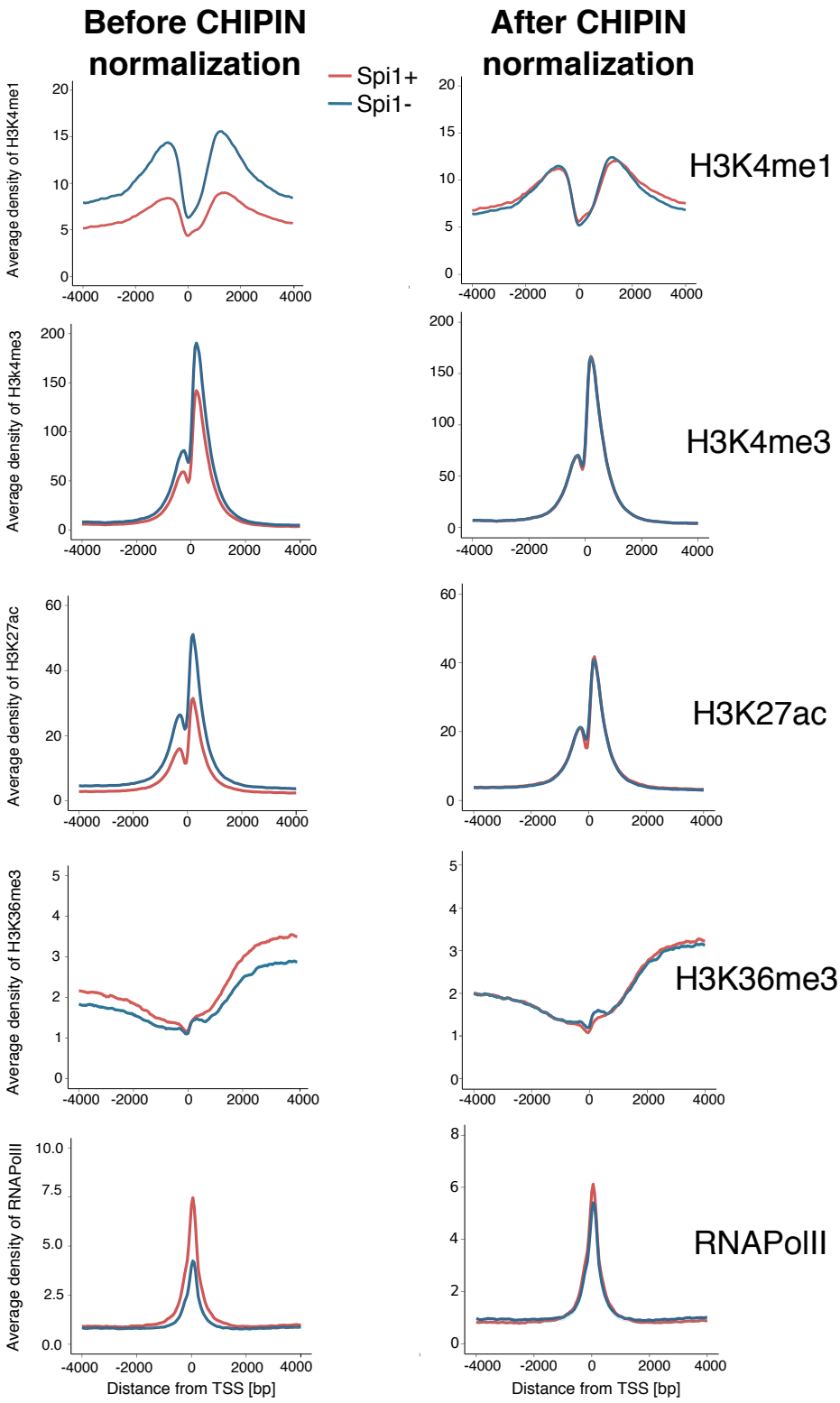


FIGURE 4.4 – **Validation de la normalisation par CHIPIN sur les ChIP-seq histones et RNAPolII.** Profils de densité des différentes marques d'histones et de RNAPolII dans les cellules pré-leucémiques dans les conditions *Spi1* surexprimé (Spi1+) et *Spi1* réprimé (Spi1-) autour des TSS des gènes NoResp, gènes dont l'expression ne varie pas en fonction de la présence ou de l'absence de SPI1 avec et sans normalisation par CHIPIN.

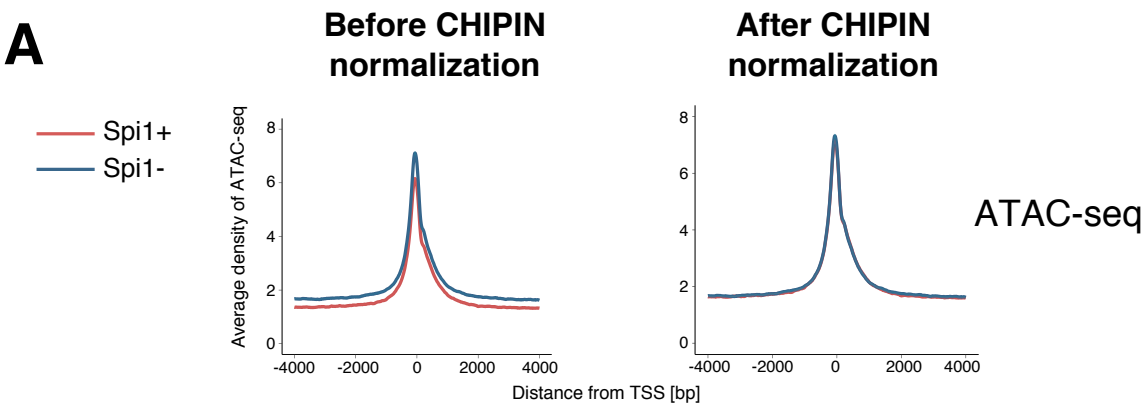


FIGURE 4.5 – **Validation de la normalisation par CHIPIN sur les données d'ATAC-seq.** Profils de densité d'ATAC-seq dans les cellules pré-leucémiques dans les conditions *spi1* surexprimé (Spi1+) et *spi1* réprimé (Spi1-) autour des TSS des gènes NoResp, gènes dont l'expression ne varie pas en fonction de la présence ou de l'absence de SPI1 avec et sans normalisation par CHIPIN.

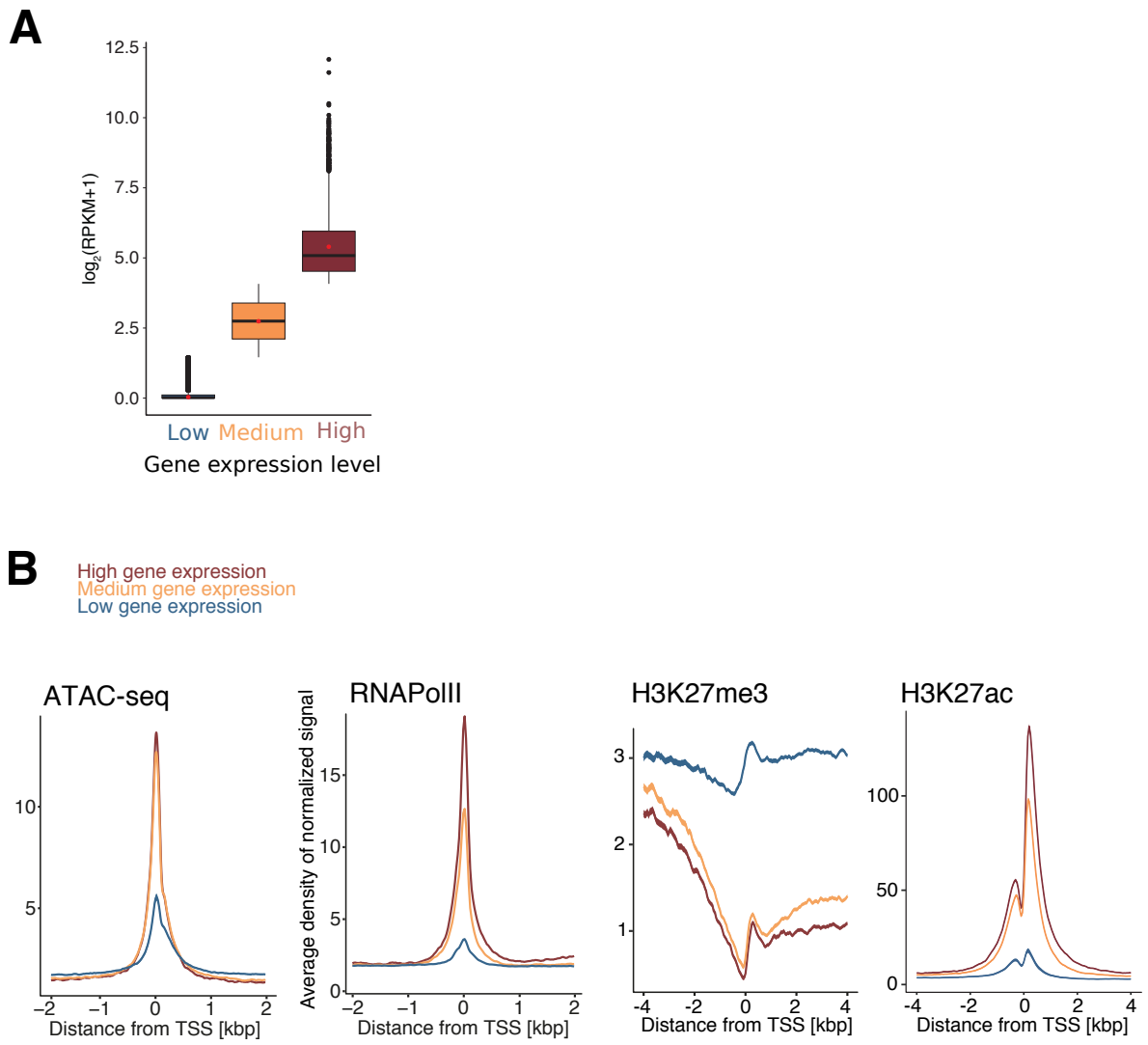


FIGURE 4.6 – **Validation des données de ChIP-seq et ATAC-seq avec les données d'expression. A.** Boxplot des valeurs d'expression ($log_2(RPKM + 1)$) séparés selon l'algorithme des k-means. **B.** Profils de densité pour les données d'ATAC-seq de ChIP-seq RNAPolII et certaines marques d'histones en fonction des données d'expression. Les différents gènes ont été clusterisés en trois catégories en utilisant l'algorithme des k-means (k=3), la densité moyenne normalisée du ChIP-seq/ATAC-seq est visualisée autour des TSSs des gènes.

4.2.2 SPI1 réprime des gènes du réseau érythroïde principal ou core erythroid network en se fixant à la chromatine

Les conséquences de la diminution de SPI1 sur les profils transcriptomiques dans les cellules pré-leucémiques ont été analysées après 63H de traitement ou non à la doxycycline par RNA-seq. La diminution de l'expression de SPI1 par la doxycycline a été validée par Western blot (Fig. 4.7A), et la diminution de sa présence à la chromatine a été analysée par ChIP-seq de la

protéine SPI1 (Fig. 4.7A). Les données de RNA-seq Spi1+/Spi1- ont permis de définir les gènes différentiellement exprimés en fonction du niveau d'expression de SPI1. Le vulcano-plot (Fig. 4.7Bb) présente :

— en abscisse : la valeur $log2(FoldChangeexpression)$ (FC) qui correspond au rapport de l'expression du gène dans la condition Spi1- (*spi-1* réprimé) sur l'expression du gène dans la condition Spi1+ (*spi-1* surexprimé).

— en ordonnée : la valeur $-log10(pvalue_{ajd})$ associée.

La diminution de l'expression de *spi1* entraîne :

— la répression de l'expression de **972 gènes** qui correspondent aux **gènes activés** par SPI1 ($FC < 0.667$, $pvalue_{adj} < 0.05$; en vert sur la Fig. 4.7A).

— l'augmentation de **607 gènes** correspondant aux **gènes réprimés** par SPI1 ($FC > 1.5$, $pvalue_{adj} < 0.05$; en rouge sur la Fig. 4.7A).

16681 gènes ne sont pas régulés par SPI1 ($0.91 < FC < 1.1$, $pvalue_{adj} \geq 0.05$; en bleu sur la Fig. 4.7A), ils seront nommés NoResp dans la suite. La catégorie Null correspond aux gènes qui ne remplissent pas les conditions précédentes. Afin de définir les gènes cibles de SPI1, c'est à dire les gènes dont la transcription est régulée par SPI1 et sur lesquels SPI1 se fixe, les données de ChIP-seq SPI1 issues des cellules pré-leucémiques surexprimant SPI1 ont été intégrées aux données sur le statut transcriptionnel des gènes (Fig. 4.7C). Sur cette figure, on représente le pourcentage de gènes activés, réprimés et NoResp qui sont fixés par SPI1. Nous n'avons considéré que les pics de SPI1 annotés entre -50kb en amont du TSS jusqu'à +3kb en aval du TE. 56.51% des gènes réprimés par SPI1 présentent au moins un pic de SPI1, cette valeur est plus de deux fois supérieure à celles des gènes non régulés par SPI1 (24.94% pour les gènes NoResp).

Nous nous sommes intéressés à la nature des gènes cibles de SPI1. Pour cela, une analyse fonctionnelle en utilisant la **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (DAVID) [158] a été réalisée sur les gènes réprimés (Fig. 4.7D) et sur les gènes activés (Fig. 4.7E). Pour les gènes activés par SPI1, on retrouve un enrichissement de la catégorie "immune system process" (Fig. 4.7E) qui correspond aux gènes connus pour être activés par SPI1 dans la lignée des macrophages ou des lymphocytes B. Pour les gènes réprimés par SPI1, cette analyse montre l'enrichissement des catégories liées à la différenciation érythroïde : erythrocyte differentiation, erythrocyte development, heme biosynthetic process (Fig. 4.7D). Ces catégories sont enrichies uniquement pour les gènes réprimés qui présentent une liaison de SPI1 (Fig. 4.7D panel du haut), cela suggère une fonction directe de SPI1 dans la répression des gènes de la différenciation érythroïde.

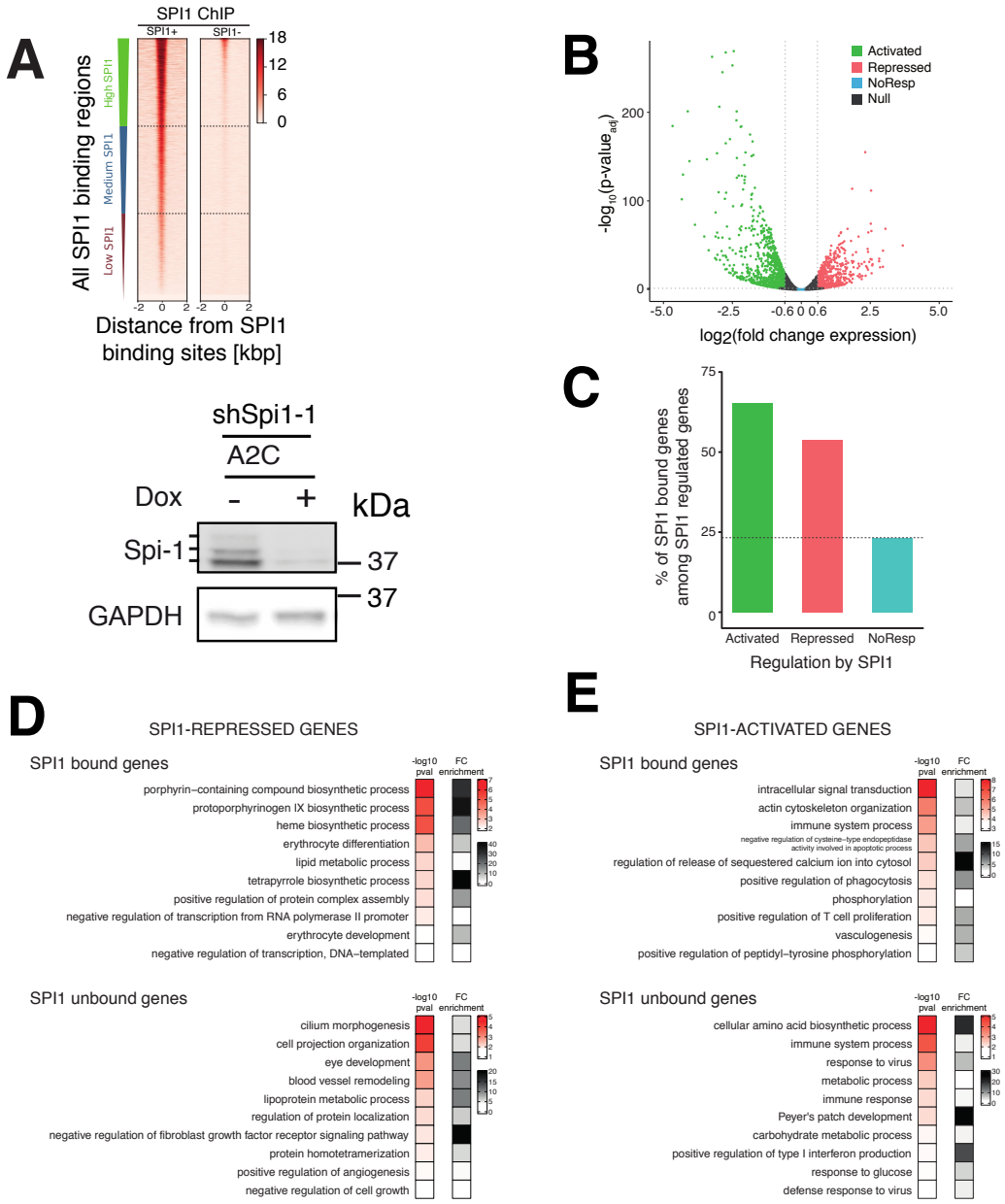


FIGURE 4.7 – **SPI1 réprime l'expression des gènes de la différenciation érythroïde en se fixant à la chromatine de ces gènes.** Les cellules ont été traitées pendant 63h par la doxycycline pour activer le shRNA anti-Spi1 pour réprimer l'expression de *spi1*. **A.** Haut : La densité du signal de SPI1 à la chromatine dans la condition Spi1+ est classée selon la force de liaison de SPI1. Le signal à la chromatine dans la condition Spi1- est reporté à droite. Bas : Image représentative d'un western blot dirigé contre la protéine SPI1 des extraits de cellules pré-leucémiques traitées ou non par la dox. **B.** Volcano plot mettant en évidence les gènes activés (vert), réprimés (rouge) par SPI1, NoResp (bleu) et Null (gris). **C.** Croisement des données de ChIP-seq Spi1+ avec les données de RNA-seq pour identifier les gènes régulés et fixés par SPI1. **D-E.** Top 10 des processus biologiques identifiés par l'analyse fonctionnelle DAVID et classés par p-value en utilisant les gènes réprimes (D) ou activés (E) qui sont fixés (panel du haut) ou non (panel du bas) par SPI1.

Un ensemble de gènes érythroïdes, appelé "core erythroid gene network" (CEGN) en anglais [159], sont fixés et régulés par les facteurs de transcription GATA1, KLF1 et TAL1 au cours de la différenciation des progéniteurs érythroides murins dérivés des cellules souches embryonnaires (ES) [159]. Afin de déterminer si c'est également le cas dans les cellules progénitrices adulte pré-leucémiques, nous avons recherché si la liaison de SPI1 à ce réseau de gènes est enrichie par rapport au hasard (i. e. l'ensemble des gènes) et si cette liaison est associée à une répression génique dans les cellules pré-leucémiques de notre étude. En utilisant les données de ChIP-seq publiées de GATA1 [151, 159], KLF1 [160], TAL1/SCL [161] dans les cellules qui développent une érythroleucémie de Friend, les cellules MEL, surexprimant également SPI1, nous avons reconstruit le CEGN des cellules érythroleucémiques (Fig. 4.8A). Celui-ci se compose de 236 gènes sur lesquels se fixent ces trois facteurs de transcription (GATA1, KFL1 et TAL1/SCL). Parmi ces 236 gènes, 183 sont également fixés par SPI1 (77.5%) dans les cellules pré-leucémiques transgéniques pour SPI1. Ainsi, il y a quatre fois plus de chance que le hasard que SPI1 se fixe sur les gènes du CEGN ($oddsratio = 4.3; pval < 2.2e-16$). De plus, l'analyse des profils d'expression de ces gènes sur lesquels SPI1 se fixe dans des cellules pré-leucémiques montre que 82% des gènes du CEGN différentiellement exprimés sont réprimés par SPI1 (Fig. 4.8B - majorité de points roses).

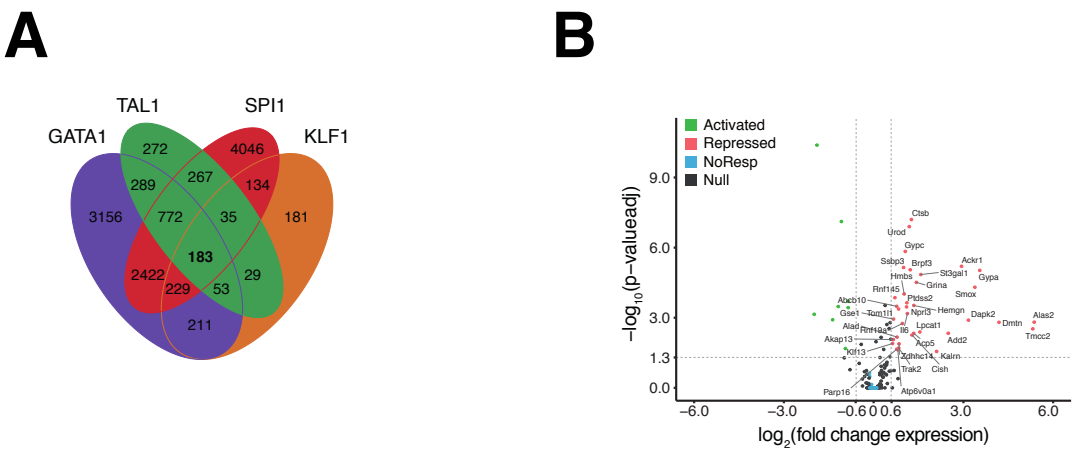


FIGURE 4.8 – **SPI1 réprime l'expression des gènes du réseau principal de gènes érythroïdes. A.** Diagramme de Venn mettant en évidence l'intersection entre les gènes fixés par trois facteurs de transcription majeurs de l'érythropoïèse (Données publiques de ChIP-seq pour GATA1, TAL1/SCL, KLF1) et les gènes fixés par SPI1. **B.** Les cellules ont été traitées pendant 72h à la doxycycline pour activer le shRNA pour réprimer l'expression de SPI1. Le volcano plot permet la projection en abcisse : log2(foldchangeexpression) et en ordonnée : -log10(adjusted pvalue)) des valeurs d'expression pour les 183 gènes fixés par GATA1, TAL1/SCL, KLF1 et SPI1. Les valeurs d'expression correspondent à celles des cellules cultivées en présence ou non de dox pendant 72h.

L'ensemble de ces données montre que SPI1 réprime une partie de ses gènes cibles en se

fixant sur la chromatine de ces gènes, incluant des gènes de la différenciation érythroïde qui est un mécanisme bloqué par SPI1.

4.2.3 SPI1 se lie dans des régions distales cis-régulatrices pour réprimer l'expression de ses gènes cibles

Pour caractériser le mode d'action de SPI1 dans la répression des gènes, nous avons identifié les régions génomiques de liaison de SPI1. Cette analyse a été réalisée en effectuant l'annotation des pics de SPI1 par ChIPseeker [154]. Chaque pic a été annoté à une région génomique selon la hiérarchie suivante : Promoter > Gene body > DownstreamTE > Intergenic. Dans le cas où il existe plusieurs isoformes avec des TSS différents pour une même annotation c'est celui dont la distance au pic de SPI1 est la plus faible qui est utilisé. La région promotrice a ensuite été divisée en trois sous-régions (Fig. 4.9A) :

— Distal Promoter = -2kb à -1kb avant le TSS
— Immediate Upstream TSS = -1kb avant le TSS au TSS
— Immediate Downstream TSS = TSS à +1kb après le TSS

La question posée ici est : **les gènes ont-ils plus de chance d'être réprimés par SPI1 s'il se fixe dans une région génomique particulière ?**
Pour répondre à cette question, nous avons évalué si les gènes régulés et fixés par SPI1 présentaient un enrichissement de la liaison de SPI1 pour une certaine sous-région génomique. L'enrichissement est calculé par la méthode du odds-ratio et la significativité est évaluée à l'aide du test de Fisher (Fig. 4.9A). La méthode du odds ratio permet de quantifier l'association entre plusieurs évènements, ici deux évènements que sont l'état transcriptionnel du gène et la région de liaison de SPI1 dans celui-ci (Fig. 4.9A). La liaison de SPI1 dans les gènes non-modulés (en bleu) n'est enrichie dans aucune sous région génomique spécifique (Fig. 4.9A). De façon intéressante, les gènes réprimés par SPI1 (en rouge) présentent un fort enrichissement en pics de SPI1 dans les régions distal promoter, corps de gène et intergenic. La liaison de SPI1 dans les régions "ImmediateUpStream" et "ImmediateDownStream" n'est pas enrichie pour les gènes réprimés ; indiquant que la liaison de SPI1 dans les promoteurs des gènes réprimés n'est pas le mécanisme préférentiel de la répression transcriptionnelle. Il n'y a pas de région spécifique pour les gènes activés.

En conclusion, bien que la liaison de SPI1 à l'ADN ne soit pas prédictive de son activité transcriptionnelle (moins de 2000 gènes sont modulés transcriptionnellement sur un total de plus de 16000 gènes fixés), nous montrons que sa liaison dans certaines sous-régions génomiques est liée à son activité de répresseur transcriptionnel. Les gènes réprimés par SPI1 sont majoritairement associés à une liaison de SPI1 dans des régions éloignées du TSS, à savoir intergénique, promoteur distal et corps de gène.

L'analyse de la localisation des pics de SPI1 dans les gènes réprimés montre que 226 gènes

des 325 gènes réprimés fixés par SPI1 ont un pic de SPI1 dans la région corps de gène dont 72% (163/226) sont uniques et 13% (30/226) sont associés à un pic de SPI1 dans la région intergénique (Fig. 4.9B). 69 gènes réprimés présentent un pic de SPI1 dans la région intergénique, ce pic est unique pour 45% (35/69) des gènes et 43% (30/69) présentent également un pic dans la région corps de gène (Fig. 4.9B).

La distribution des pics de SPI1 en pourcentage de la taille de la région intergénique (48kb) montre que les pics de SPI1 sont principalement distribués proche de la région promotrice (Fig. 4.10A). Au contraire, les pics localisés dans le corps de gène sont distribué de façon relativement homogène sur toute la longueur des gènes (Fig. 4.10A). En raison du faible nombre de gènes réprimés et fixés par SPI1 dans la région promoteur distal (14/325), nous avons décidé de ne pas prendre en compte cette région et de focaliser la suite de nos analyses sur les régions intergéniques et corps de gène.

Considérant le nombre de gènes et l'enrichissement de la liaison de SPI1 dans des régions génomiques, ces résultats suggèrent que les corps de gène et les régions intergéniques des gènes réprimés sont deux régions qui jouent un rôle clef dans la répression transcriptionnelle exercée par SPI1.

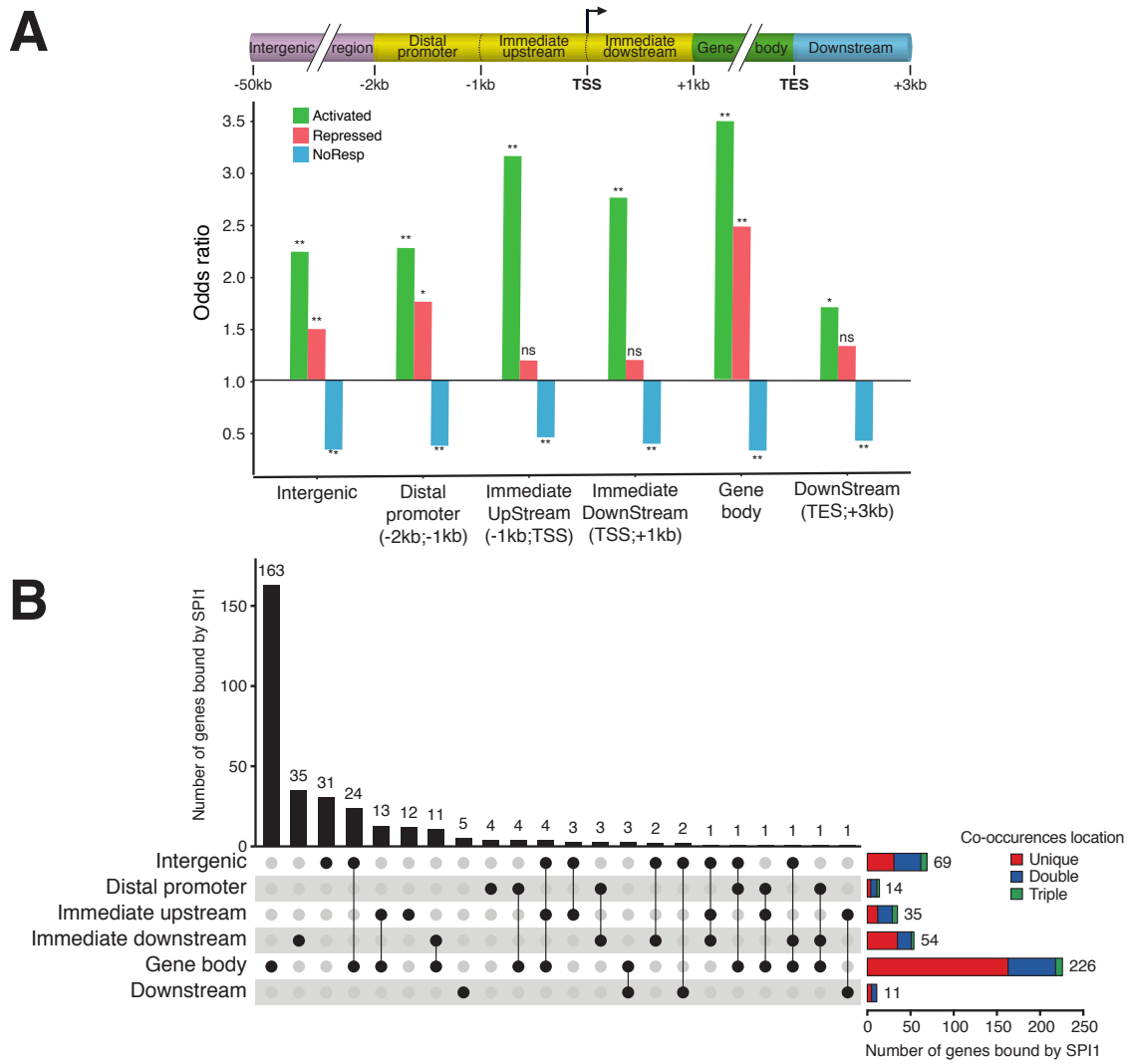


FIGURE 4.9 – **SPI1 se fixe dans des régions cis-régulatrices distales pour réprimer l'expression de ses gènes cibles. A.** Enrichissement en gènes fixés par SPI1 pour chaque catégorie de régulation transcriptionnelle dans les différentes régions génomiques ; (*) $pval < 10^{-2}$, (**) $pval < 10^{-3}$. **B.** Upset plot représentant le nombre de gènes réprimés liés par SPI1 dans les différentes régions génomiques. Les points noirs pleins représentent les régions génomiques liées par SPI1, les liaisons multiples de SPI1 dans différentes régions du même gènes sont indiquées par deux points ou plus reliés par une ligne. Dans le bar plot vertical sont indiqués le nombre de pics de SPI1 représentant chaque combinaison, tandis que dans le stack plot horizontal sont indiqués le nombre de gènes dans lesquels SPI1 se lie dans un, deux ou trois endroits génomiques différents.

Nous avons donc cherché à comprendre comment la liaison de SPI1 dans ces deux types de régions intervient pour réprimer les gènes cibles de SPI1. Ceci concerne 69 gènes avec 92 pics dans la région intergénique et 226 gènes avec 382 pics dans la région corps de gène (Tab. 4.2). Pour cela, nous avons exploré l'hypothèse selon laquelle la liaison de SPI1 dans les régions corps de gène et intergéniques des gènes réprimés pourraient être des éléments distaux cis-

	Intergenic	Distal Promoter	Immediate UpStream TSS	Immediate DownStream TSS	Gene body	Downstream TE
nPeaks	92	14	35	55	382	11
nGenes	69	14	35	54	226	11

TABLE 4.2 – **Tableau présentant le nombre de pics et le nombre de gènes différents fixés par SPI1 dans chacune des six régions génomiques pour les gènes réprimés.**

régulateurs. Le terme distal est employé ici en opposition au terme proximal qui concerne les régions autour du TSS et englobe donc à la fois les régions de corps de gène et les régions intergéniques. De plus, nous employerons le terme enhancer dans la suite du manuscrit car il est beaucoup utilisé dans la littérature biologique pour définir les modules cis-régulateurs (CRM). Par conséquent, nous avons établi trois états chromatiniens sur les sites de liaison de SPI1 dans les gènes réprimés en utilisant ChromHMM [157]. Pour cela, deux marques épigénétiques ont été utilisées : H3K4me1 indiquant la présence d'une région enhancer et H3K27ac qui indique que la région est active transcriptionnellement. Les pics de ces marques épigénétiques, utilisés par ChromHMM pour établir le paysage des états chromatiniens ont été obtenus par l'analyse des ChIP-seq dans les cellules pré-leucémiques exprimant ou non SPI1 (*spi1+*, *spi1-*). Nous avons défini les régions présentant uniquement la marque H3K4me1 comme des régions enhancers inactives, les régions présentant à la fois les marques H3K27ac et H3K4me1 comme des régions actives et celles ne présentent aucune des deux marques sont appelées "none". Ces analyses révèlent que 91.3% et 90% des régions intergéniques et corps de gène, respectivement, liées par SPI1 dans les gènes réprimés sont des enhancers, dont 77% et 75%, respectivement, sont des enhancers actifs (Fig. 4.10B). Ceci est en accord avec le fait que SPI1 réprime mais n'éteint pas les gènes (Fig. 4.10C).

En conclusion, SPI1 réprime les gènes principalement en se fixant sur des régions de régulation distales actives appelées enhancers actifs qui se trouvent dans les régions intergéniques ou dans les régions de corps de gène des gènes réprimés.

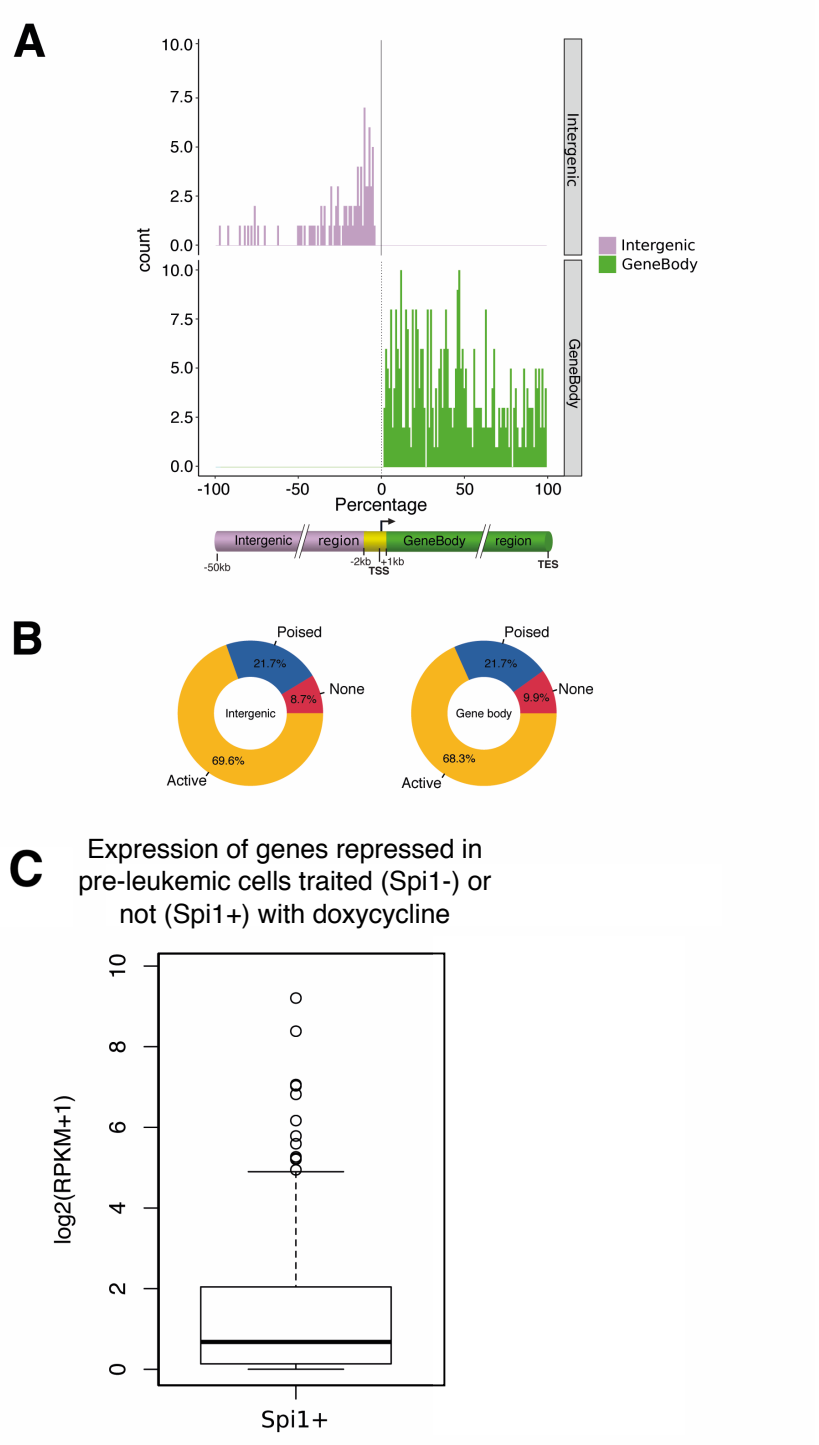


FIGURE 4.10 – **SPI1 se fixe dans des régions d'enhancer actif pour réprimer l'expression de ses gènes cibles. A.** Distribution des pics de SPI1 en région gene body et en région intergénique dans les gènes réprimés en pourcentage de la longueur totale des gènes pour la région gene body et en pourcentage de la taille de la région intergénique (48kB) pour la région intergénique. **B.** Doughnut-chart présentant le pourcentage de peak de SPI1 dans des régions portant H3K27ac et H3K4me1 (jaune), ou uniquement H3K4me1 (bleu) ou portant ni l'une ni l'autre de ces deux marques (rouge) parmi les pics de SPI1 localisés dans des régions intergénique (gauche) ou gene body (droite) dans les gènes réprimés. **C.** Boxplot représentant l'expression des gènes réprimés par SPI1 ($log_2(RPKM + 1)$) dans les cellules pré-leucémiques non traitées (Spi1+) à la doxycycline.

4.2.4 SPI1 interagit avec la déacétylase HDAC1 qui représente un candidat pour servir de médiateur de l'activité transcriptionnelle répressive de SPI1

SPI1 n'a pas d'activité répressive propre mais il possède un domaine d'activation transcriptionnelle, nous avons donc fait l'hypothèse qu'il réprime les gènes avec des partenaires tels que des facteurs épigénétiques. Afin de définir les partenaires de SPI1 à la chromatine, nous avons effectué une expérience appelée RIME pour **R**apid **I**mmuno-precipitation **M**ass spectometry of **E**ndogenous protéin [162] qui combine un ChIP SPI1 et l'identification des protéines associées à SPI1 par spectrométrie de masse. Parmi les protéines identifiées (Fig. 4.11A), on trouve des facteurs d'épissage (en bleu) connus pour interagir avec SPI1 (dont FUS), des protéines de liaison à l'ARN (en bleu), des protéines du complexe SWI/SNF (en vert) dont les protéines SMARC et des protéines du complexe répresseur de la transcription HDAC1/mSIN3A et HDAC1/NURD (HDAC1 et CHD4) (en rose). Nous nous sommes intéressés au rôle de HDAC1 qui est une histone dé-acétylase impliquée dans la répression transcriptionnelle. L'interaction entre SPI1 et HDAC1 a été validée par co-immunoprécipitation (Fig. 4.11B).

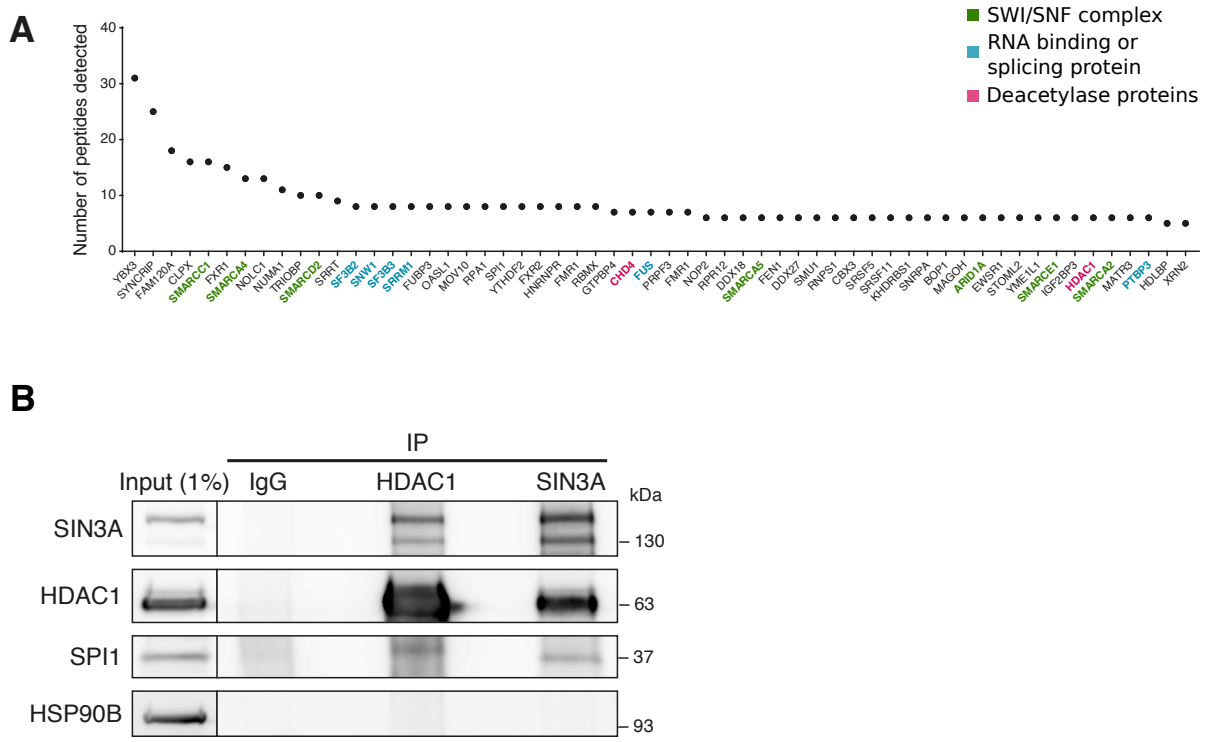


FIGURE 4.11 – **Partenaires de SPI1 à la chromatine. A.** Résultat de l'expérience de RIME permettant l'identification des protéines associées à SPI1 (**R**apid **I**mmunoprecipitation **M**ass spectrometry of **E**ndogenous proteins [162]). Les protéines avec une activité déacétylase sont indiqués en rouge foncé, les protéines du complexe SWI/SNF en bleu et les protéines de liaison à l'ARN ou impliqués dans des évènements d'épissage en vert. **B.** Immunoprécipitation de HDAC1 ou de Sin3A et révélation par Western blot avec les anticorps anti-HDAC1, anti-SIN3A et anti-SPI1. HSP90B est utilisé comme contrôle négatif de l'immunoprécipitation.

4.2.5 SPI1 et HDAC1 agissent ensemble pour la répression des gènes cibles de SPI1

Afin de déterminer le rôle de HDAC1 dans la répression exercée par SPI1 sur ses gènes cibles, nous avons inhibé l'activité de la protéine HDAC1 en utilisant un inhibiteur pharmacologique : l'entinostat. L'entinostat inhibe également la protéine HDAC3. L'efficacité de l'inhibiteur entinostat est démontrée par l'augmentation du niveau de H3K27ac, attestant que l'inhibiteur a bloqué l'activité dé-acétylase de HDAC1/3 (Fig. 4.12A). Les expériences suivantes ont pour but de définir si SPI1 requiert HDAC1 pour réprimer les gènes et si HDAC1 requiert SPI1 pour réprimer les gènes cibles de SPI1. Pour cela, les ARNs des cellules pré-leucémiques traitées ou non par l'entinostat associé ou non à la doxycycline pour réduire l'expression de SPI1 ont été extraites et des expériences de RNA-seq ont été pratiquées. Le diagramme de Venn présente l'intersection entre les gènes réprimés par SPI1 en présence de HDAC1 (en rose) et les gènes réprimés par HDAC1 en présence de SPI1 (en vert) (Fig. 4.12B). Plus de 60%

des gènes réprimés par SPI1 sont également réprimés par HDAC1 (test de fisher donne une valeur de odds-ratio supérieure à 2 et une pvaleur inférieure à $2.2e-16$), cela suggère une possible coopération fonctionnelle entre ces deux facteurs. Nous avons ensuite analysé les niveaux d'expression globale des gènes réprimés (Fig. 4.12C) et des gènes activés (Fig. 4.12E) selon l'activité de HDAC1 et le niveau d'expression de SPI1. Les violin plots indiquent les valeurs de $log_2(FoldChangeExpression)$ des gènes des cellules traitées versus non traitées à la doxycycline et/ou à l'entinostat permettant d'analyser :

— les effets de l'inhibition de SPI1 (bleu) et de l'inhibition de HDAC1 (vert) **séparément (1)**.

— les effets de l'inhibition de SPI1 en présence (bleu) et en absence (violet) de HDAC1 **(2)**.

— les effets de l'inhibition de HDAC1 en présence (vert) et en absence (violet) de SPI1 **(3)**.

Dans le cas des gènes réprimés, on observe qu'en absence de l'expression de *spi1* ou de l'activité de HDAC1, les expressions augmentent comme attendu (Fig. 4.12C **(1)**). Au contraire pour les gènes activés (Fig. 4.12E **(1)**), la perte de SPI1 et de l'activité de HDAC1 ont des conséquences opposées, en accord avec l'indépendance de SPI1 avec HDAC1 dans son activité activateur de la transcription. Les violin plots centraux présentent les conséquences de l'absence de SPI1 en présence (bleu) ou en absence (violet) d'une activité HDAC1 dans les cellules. Les résultats montrent qu'en absence de HDAC1, SPI1 réprime moins fortement les gènes (Fig. 4.12C **(2) violet**). Les violin plots de droite montrent l'effet de l'inhibition de HDAC1 sur les gènes réprimés par SPI1 en présence ou en absence de SPI1. On observe qu'en absence de SPI1, HDAC1 réprime moins les gènes qu'en présence de SPI1 (Fig. 4.12C **(3) violet**). En ce qui concerne les gènes activés par SPI1 (Fig. 4.12E), l'absence de SPI1 en présence ou en absence de l'activité de HDAC1 et l'inhibition de HDAC1 en présence ou en absence de SPI1 changent peu les niveaux d'expression des gènes (Fig. 4.12C **(2) et (3)**). Les résultats de l'expérience décrite ci-dessus sont présentés pour deux gènes réprimés (Fig. 4.12D) et deux gènes activés (Fig. 4.12F) par SPI1.

Ces données montrent que SPI1 a besoin de HDAC1 pour réprimer ses gènes cibles et que HDAC1 a besoin de SPI1 pour réprimer les gènes cibles de SPI1. Au contraire l'activation transcriptionnelle par SPI1 est indépendante de HDAC1. Ainsi SPI1 et HDAC1 coopèrent pour réprimer l'expression des gènes cibles de SPI1.

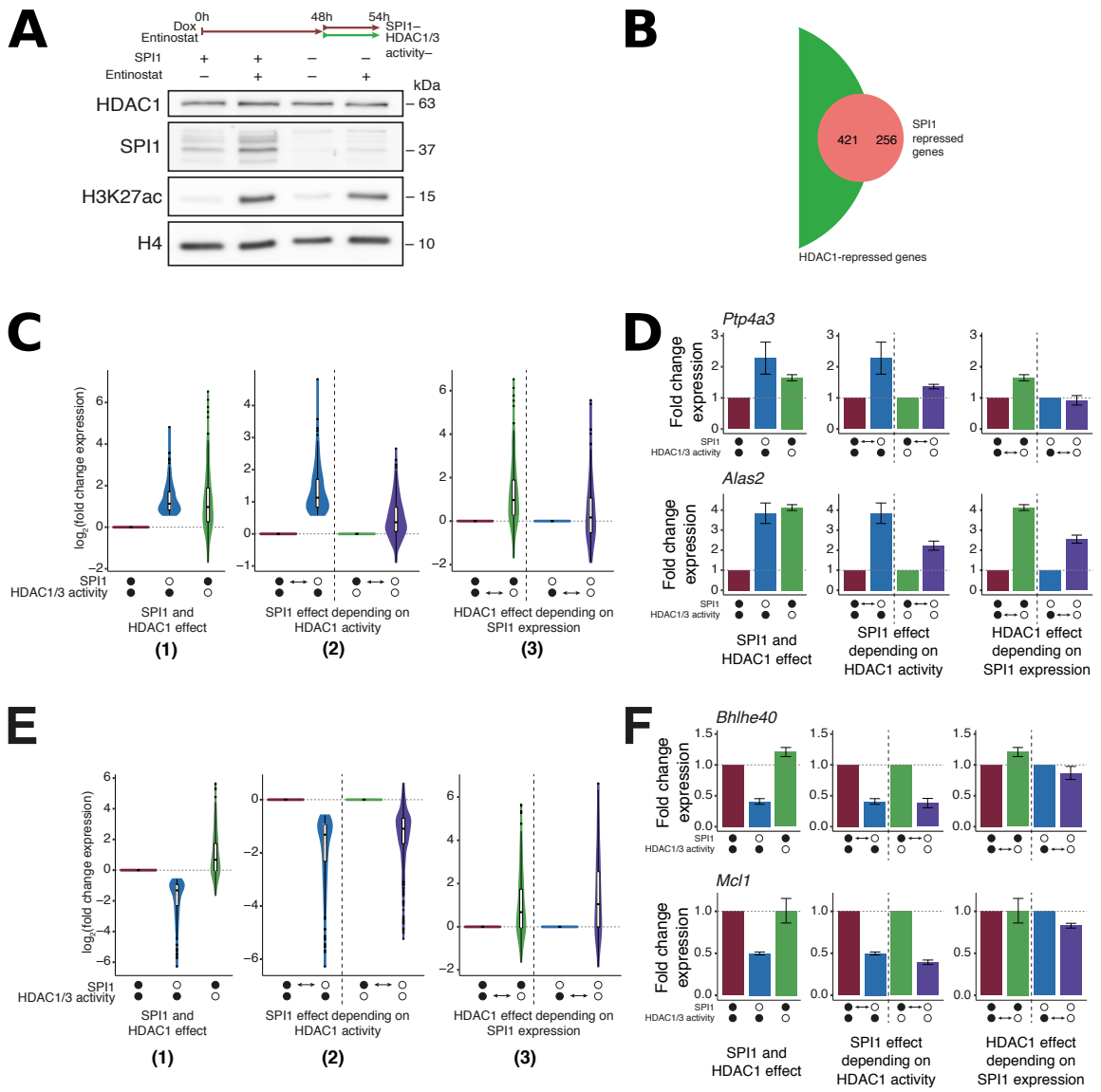


FIGURE 4.12 – **SPI1 et HDAC1 sont tous les deux nécessaires pour la répression des gènes cibles de SPI1. A.** Les cellules ont été traitées pendant 48h avec ou sans doxycycline pour induire l'expression du shRNA contre SPI1 puis combinées avec un inhibiteur de HDAC1/3 (Entinostat) ou au DMSO pendant 6h supplémentaires, des extraits de cellules entières ont été analysés par immunoblotting avec les anticorps contre HDAC1, SPI1, H3K27ac et H4. **B.** Diagramme de Venn montrant l'intersection entre les gènes réprimés par SPI1 (rose) et les gènes réprimés par HDAC1 (vert) de l'analyse du RNA-seq comparant les cellules traitées ou non avec la doxycycline et les cellules traitées ou non avec l'entinostat. **C, E.** Violin plots des valeurs de log2 de fold change d'expression des gènes réprimés (C) ou activés (E) par SPI1. **(1)** représente les effets de l'inhibition de SPI1 (bleu) et de HDAC1 (vert) séparément. **(2)** représente les effets de l'inhibition de SPI1 en présence (bleu) et en absence (violet) de HDAC1. **(3)** représente les effets de l'inhibition de HDAC1 en présence (vert) et en absence (violet) de SPI1. **D, F.** Barplots de la valeur moyenne de foldchange d'expression de l'expérience de RNA-seq des tréplicats +/- SEM pour des gènes réprimés (D) ou activés (F).

4.2.6 SPI1 induit localement une dé-acétylation de l'histone H3 dans les enhancers où il se fixe et aux TSSs associés dans les gènes réprimés

HDAC1 étant capable de dé-acétyler les lysines de l'histone H3 et H3K27ac étant une marque de l'activation transcriptionnelle, nous avons fait l'hypothèse que SPI1 est capable d'induire une dé-acétylation locale des histones au niveau de ses sites de liaison par un mécanisme impliquant l'activité de HDAC1. Ce mécanisme pourrait être réponsable de l'effet répressif de SPI1 sur certains de ses gènes cibles. Pour cela, nous avons comparé le niveau de H3K27ac par ChIP-seq entre les cellules pré-leucémiques traitées ou non par dox pendant 63h. En particulier, nous nous sommes intéressés aux sites de liaison de SPI1 (+/- 2kb), dans les gènes réprimés, situés dans les enhancers actifs et inactifs, en région intergénique (Fig. 4.13A) ou en région corps de gène (Fig. 4.13B) ainsi qu'aux TSS associés (associated TSSs) dans les deux cas. Les profils de densité moyenne de H3K27ac montrent que la diminution de SPI1 (condition SPI1-) entraîne une augmentation du signal H3K27ac au niveau des sites de liaison de SPI1 sur les enhancers actifs en région intergénique et en région corps de gène de ses gènes cibles réprimés. Dans les deux cas, l'acétylation des TSS augmente également lorsque SPI1 diminue (Fig. 4.13A et B panels de gauche). En revanche, la diminution de SPI1 ne semble pas réactiver le signal de H3K27ac lorsque SPI1 est fixé dans les enhancers inactifs (Fig. 4.13A et B panels de droite). De plus, dans les régions null qui ont été définies précédemment (ne portant ni H3K27ac ni H3K4me1) la diminution de SPI1 ne permet pas non plus la réactivation du signal H3K27ac (données non montrées). De plus, il est intéressant de noter que même si SPI1 ne change pas l'acétylation des enhancers intergénique inactifs (Fig. 4.13A panel de droite), il est capable de réduire l'activité du TSS associé. De façon intéressante, dans les gènes pour lesquels SPI1 ne modifie pas la transcription (gènes NoResp) mais se fixe en région intergénique ou corps de gène, il n'existe pas de différence dans le niveau de H3K27ac aux TSSs associés et une diminution de H3K27ac est observée lorsque SPI1 est diminuée à l'inverse des gènes réprimés par SPI1, cela indique donc une réponse spécifique de l'acétylation dans le cas des gènes réprimés. Les régions intergéniques et corps de gène fixées par SPI1 dans les gènes réprimés ont été rassemblées pour les enhancers actifs et inactifs (Fig. 4.14A - panel du haut) ainsi qu'aux TSSs associés (Fig. 4.14A - panel du haut).

Ces résultats suggèrent que SPI1 est capable de réprimer un groupe de gènes en se fixant dans des enhancers actifs et de diminuer l'acétylation des enhancers localement autour de lui et également distalement aux TSSs de ces gènes.

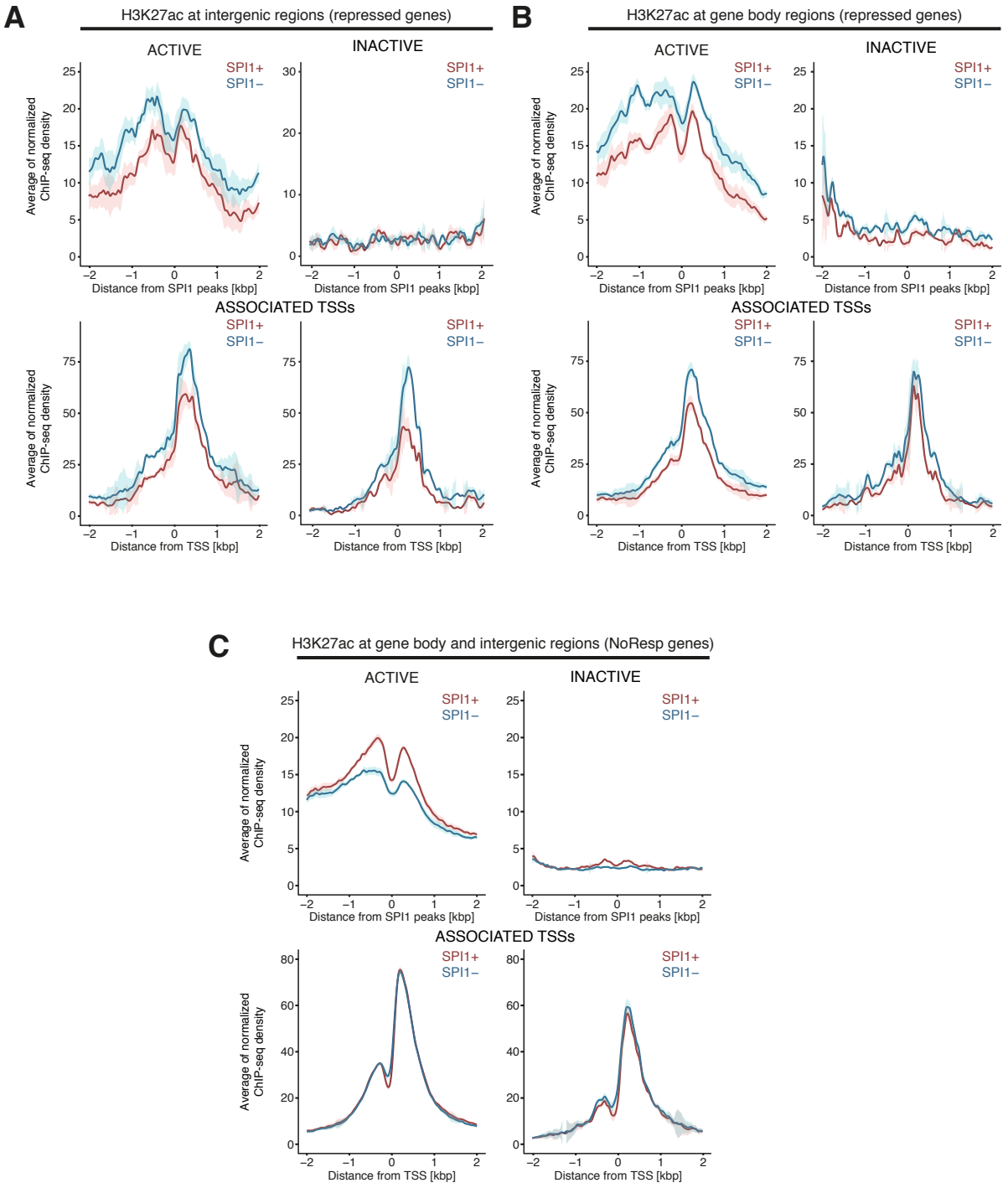


FIGURE 4.13 – **SPI1 diminue l'acétylation des lysines 27 de l'histone H3 dans les enhancers actifs où il se fixe ainsi qu'aux TSSs des gènes.** Profils de densité moyenne de H3K27ac dans les conditions SPI1+ et SPI1- centrés sur la position peakmax de SPI1 situés **A.** dans les gènes réprimés dans les enhancers activés ou inactivés en région intergénique et aux TSSs des gènes correspondants ("Associated TSS"). **B.** dans les gènes réprimés dans les enhancers activés ou inactivés en région corps de gène et aux TSSs des gènes correspondants ("Associated TSS"). **C.** dans les gènes NoResp dans les enhancers activés ou inactivés (corps de gène et intergénique confondus) et aux TSSs des gènes correspondants ("Associated TSS").

Afin de mieux définir le mécanisme de répression pour les gènes fixés par SPI1 dans les enhancers actifs, nous avons affiné les groupes de gènes en séparant ceux pour lesquels SPI1 réduit fortement H3K27ac autour de lui ($FC_{H3K27ac}^{Dox1/Dox0} > 1.5$) appelés $H3K27ac^{Diff}$ et ceux pour lesquels SPI1 n'a pas d'impact sur la variation de H3K27ac ($0.83 \leq FC_{H3K27ac}^{Dox1/Dox0} \leq 1.2$) appelés $H3K27ac^{NoDiff}$. Les pics de SPI1 dans le contexte $H3K27ac^{Diff}$ sont au nombre de 71 sur 58 gènes différents, la densité moyenne de H3K27ac dans ces régions et aux TSSs associés est présentée sur la Fig. 4.14B (colonne de gauche). Les pics de SPI1 dans le contexte $H3K27ac^{NoDiff}$ sont au nombre de 63 sur 54 gènes différents, la densité moyenne de H3K27ac dans ces régions et aux TSSs associés est présentée sur la Fig. 4.14B (colonne de droite). Afin de vérifier que HDAC1 agit sur ces enhancers dont l'acétylation est réduite par SPI1, le niveau de H3K27ac a été évalué par ChIP-seq après traitement des cellules par l'inhibiteur de HDAC1/3, l'entinostat (Fig. 4.14C). La comparaison du signal ChIP-seq des enhancers dans le contexte $H3K27ac^{Diff}$ révèle que le niveau de H3K27ac dépend de l'activité de HDAC1 (Fig. 4.14C) comme de la présence de SPI1 (Fig. 4.14B colonne de gauche). Ceci est également vrai pour les TSS associés (Fig. 4.14C colonne de gauche et Fig. 4.14B colonne de gauche). Ces données renforcent l'idée selon laquelle HDAC1 et SPI1 agissent ensembles pour réprimer les gènes. La Fig. 4.14D montre des profils représentatifs de ChIP-seq de SPI1, H3K4me1, H3K27ac dans les cellules traitées ou non en présence de dox (SPI1- et SPI1+, respectivement) et de H3K27ac dans des cellules traitées ou non en présence d'entinostat (Entinostat et DMSO, respectivement) pour deux gènes réprimés par SPI1 : les gènes *Alas2* et *Ptp4a3*.

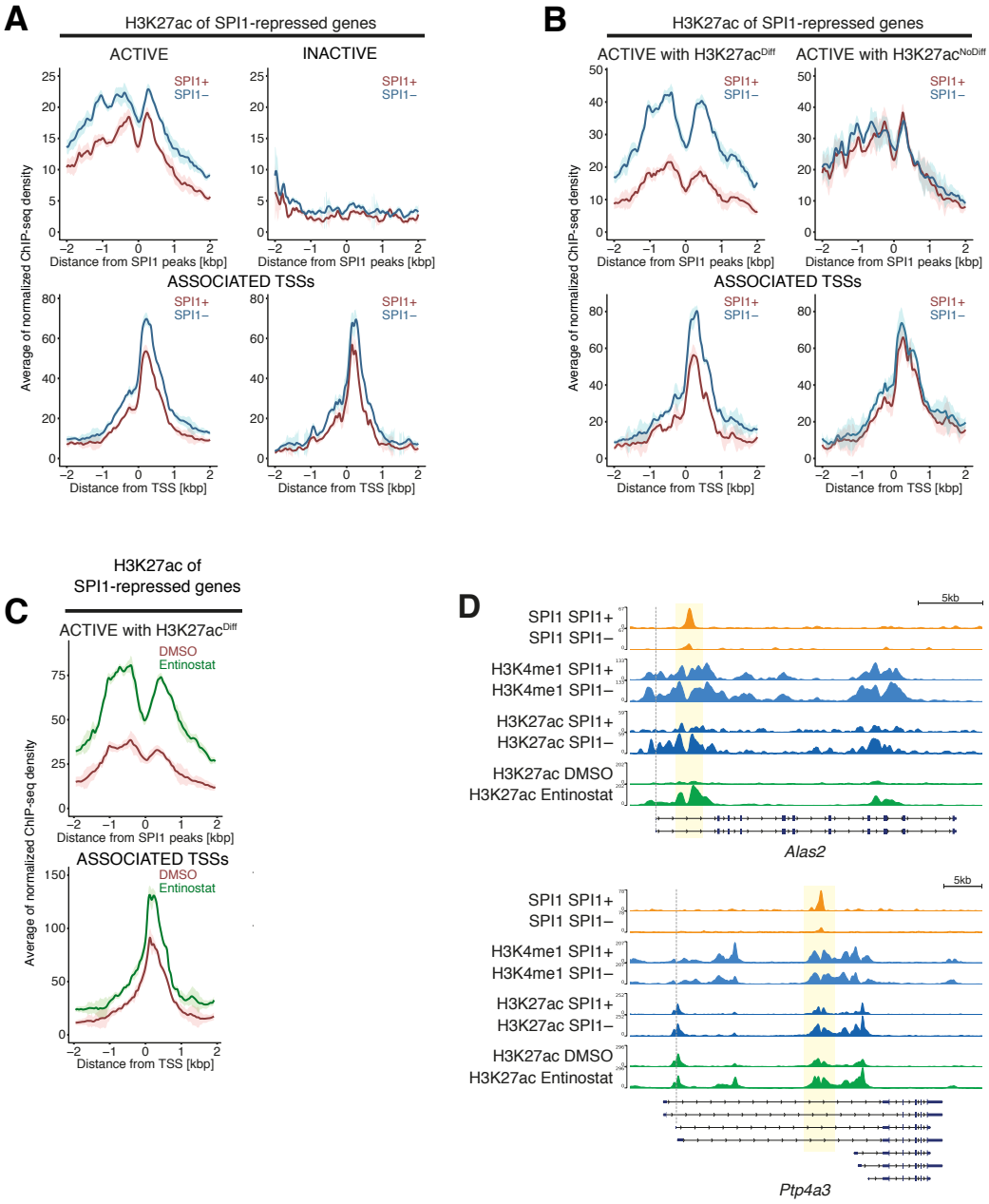


FIGURE 4.14 – **SPI1 est capable de diminuer l'acétylation dans les régions autour de lui et aux TSSs associés dans les gènes réprimés.** Profils de densité moyenne de H3K27ac dans les conditions SPI1+ et SPI1- centrés sur la position peakmax de SPI1, et aux TSSs des gènes correspondants ("Associated TSS"), situés **A.** dans les gènes réprimés dans les enhancers activés ou inactivés (régions corps de gène et intergénique confondues). **B.** dans les gènes réprimés dans les enhancers activés pour lesquels SPI1 réduit fortement H3K27ac autour de lui ($H3K27ac^{Diff}$) ou pour lesquels SPI1 n'a pas d'impact sur le niveau de H3K27ac ($H3K27ac^{NoDiff}$). **C.** Profils de densité moyenne de H3K27ac dans les conditions DMSO et Entinostat centrés sur la position peakmax de SPI1 situés dans les gènes réprimés dans les enhancers activés pour lesquels SPI1 réduit fortement H3K27ac autour de lui ($H3K27ac^{Diff}$) ou dans les enhancers inactivés (régions corps de gène et intergénique confondues). **D.** Densités de ChIP-seq pour les gènes *Alas2* et *Ptp4a3* (réprimés par SPI1) pour SPI1, H3K4me1, H3K27ac dans les conditions SPI1+ et SPI1- et pour H3K27ac dans les conditions DMSO et Entinostat.

L'ensemble des résultats obtenus jusqu'à présent montrent que :

- **SPI1 et HDAC1 coopèrent pour réprimer les gènes sur lesquels SPI1 se fixe aux enhancers localisés dans les régions intergéniques ou dans les corps de gène.**
- **SPI1 et HDAC1 réduisent l'acétylation de ces enhancers.**
- **Cette diminution d'acétylation des enhancers s'accompagne d'une diminution de l'activité des TSS, celle-ci est constatée par la diminution de H3K27ac, suggérant un cross-talk entre ces deux régions.**

4.2.7 Conséquences de la diminution de l'acétylation sur l'accessibilité à la chromatine et le niveau de RNAPolII

Les histones acétylées fonctionnent comme des plateformes pour le recrutement de protéines effectrices spécfiques, telles que des régulateurs de la transcription (ex. Mediator, BRD4) ou des remodeleurs de la chromatine (protéines à bromodomaine telles que SWI/SNF). Des données plus récentes suggèrent que les modifications des histones ont également un effet direct sur l'architecture des nucléosomes en modulant la charge des histones [163].

Dans le but de comprendre les conséquences de la diminution de l'acétylation des histones par SPI1 sur la structure de la chromatine, nous avons analysé le niveau d'accessibilité à la chromatine par ATAC-seq dans les cellules cultivées en présence ou en absence de dox (Fig. 4.15). Dans les cellules pré-leucémiques, nous observons que la liaison de SPI1 aux enhancers des gènes réprimés, bien que réduisant le signal H3K27ac (contexte $H3K27ac^{Diff}$), ne change pas l'accessibilité à la chromatine (Fig. 4.15A panel du haut). En revanche, les TSSs de ces gènes présentent une réduction de l'accessibilité à la chromatine en présence de SPI1 (Fig. 4.15A panel du milieu). Le scénario est différent pour les enhancers qui ne sont pas différentiels pour H3K27ac (contexte $H3K27ac^{NoDiff}$) (Fig. 4.15B), en effet sur ces régions, SPI1 augmente faiblement l'accessibilité à la chromatine alors que celle du TSS ne change pas. Le même type de réponse est observée pour les gènes NoResp (Fig. 4.15C). En revanche, comme déjà décrit dans les macrophages, SPI1 facilite l'accessibilité à la chromatine des gènes dont il active la transcription et qui sont différentiels pour H3K27ac au niveau des enhancers (Fig. 4.15D) [164]. Comme le TSS et son voisinage sont les régions pour lesquelles nous avons observé une réduction de l'accessibilité lorsque SPI1 se fixe aux enhancers dans les gènes réprimés, nous avons évalué les conséquences sur le niveau de RNAPolII aux TSSs (Fig. 4.15 panel du bas). Nous montrons que le niveau de RNAPolII au niveau des TSSs suit celui de l'accessibilité à la chromatine. En effet, dans le cas des gènes réprimés fixés par SPI1 dans des régions où il réduit fortement le niveau de H3K27ac, SPI1 réduit la quantité de RNAPolII (Fig. 4.15A). L'inverse est observé pour les gènes activés par SPI1, avec une augmentation de la charge de RNAPolII par SPI1 (Fig. 4.15D). Au niveau des TSSs des gènes fixés par SPI1 pour lesquels SPI1 ne modifie par le niveau de H3K27ac, aucune différence n'est observée dans le niveau de RNAPolII

(Fig. 4.15B).

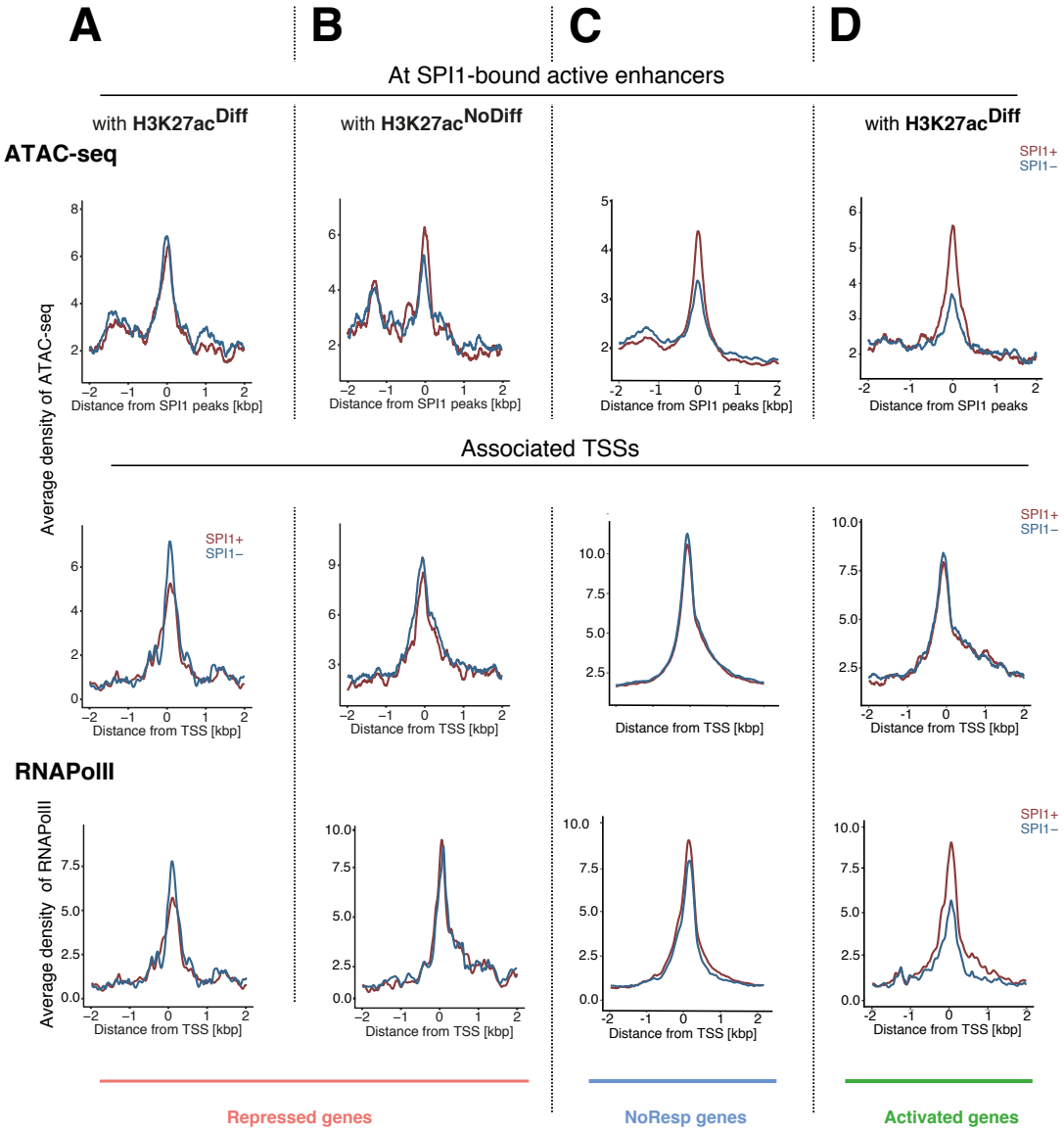


FIGURE 4.15 – **Caractérisation de l'ouverture de la chromatine et du niveau de RNAPolII dans les régions de enhancers actifs fixées par SPI1.** Profils de densité moyenne au niveau de la position peakmax de SPI1 dans les conditions SPI1+ et SPI1- pour l'ATAC-seq et aux TSSs des gènes correspondants ("Associated TSSs") pour l'ATAC-seq et RNAPolII. **A-D.** dans les gènes réprimés (**A**) et activés (**D**) fixés par SPI1 dans les enhancers activés (régions corps de gène et intergénique confondues) pour lesquels SPI1 réduit fortement H3K27ac autour de lui. **B.** dans les gènes réprimés fixés par SPI1 dans les enhancers activés (régions corps de gène et intergénique confondues) pour lesquelles SPI1 n'a pas d'impact sur le niveau de H3K27ac. **C.** dans les gènes NoResp fixés par SPI1 dans les enhancers activés (régions corps de gène et intergénique confondues).

En conclusion, SPI1 réduit l'accessibilité à la chromatine et diminue la charge de RNA-

PolII aux TSSs des gènes qu'il réprime par liaison à l'enhancer intergénique ou dans le corps de gène sans changer le niveau d'accessibilité de la chromatine localement sur ces enhancers. Ces résultats montrent l'existence d'une régulation par SPI1 fixé aux enhancers sur la dé-acétylation, la compaction et la charge de RNAPolII aux TSS, suggérant ainsi une régulation à longue distance par looping entre l'enhancer et le TSS.

4.2.8 Recherche de facteurs de transcription liés dans les enhancers fixés par SPI1 pour lesquels il réduit l'acétylation dans les gènes réprimés

Afin d'identifier quels effets pourraient avoir les changements de la structure de la chromatine et du niveau de H3K27ac, nous avons cherché quels facteurs de transcription se trouvent enrichis au niveau des enhancers des gènes réprimés. Pour cela une recherche de motifs enrichis a été réalisée autour des sites de liaison de SPI1 dans six groupes de régions en utilisant l'outil AME de la suite MEME [132] :

- Pics de SPI1 dans les gènes réprimés localisés dans les enhancers activés différentiels pour H3K27ac ($H3K27ac^{Diff}$)
- Pics de SPI1 dans les gènes réprimés localisés dans les enhancers activés non différentiels pour H3K27ac ($H3K27ac^{NoDiff}$)
- Pics de SPI1 dans les gènes réprimés localisés dans des enhancers inactifs (n=63).
- Pics de SPI1 dans les gènes réprimés localisés dans les promoteurs (n=90).
- Pics de SPI1 dans les gènes activés localisés dans des enhancers activés (n=874).
- Pics de SPI1 dans les gènes NoResp localisés dans des enhancers activés (n=3159).

La recherche a été faite autour de la position peakmax de SPI1 +/-150bp. La Fig. 4.16A présente les motifs enrichis dans les six catégories de pics de SPI1 des enhancers, les noms indiqués correspondent à des facteurs de transcription exprimés dans les cellules pré-leucémiques. L'abscisse indique le rang de chaque motif dans le classement par p-valeur ajustée et l'ordonnée correspond à la valeur $-log10(pvalue_{adjusted})$. les motifs de chaque famille sont indiqués dans une couleur différente :

- famille ETS : bleu
- famille GATA : orange
- famille IRF : rouge
- famille KLF/SP : violet
- autre famille : vert

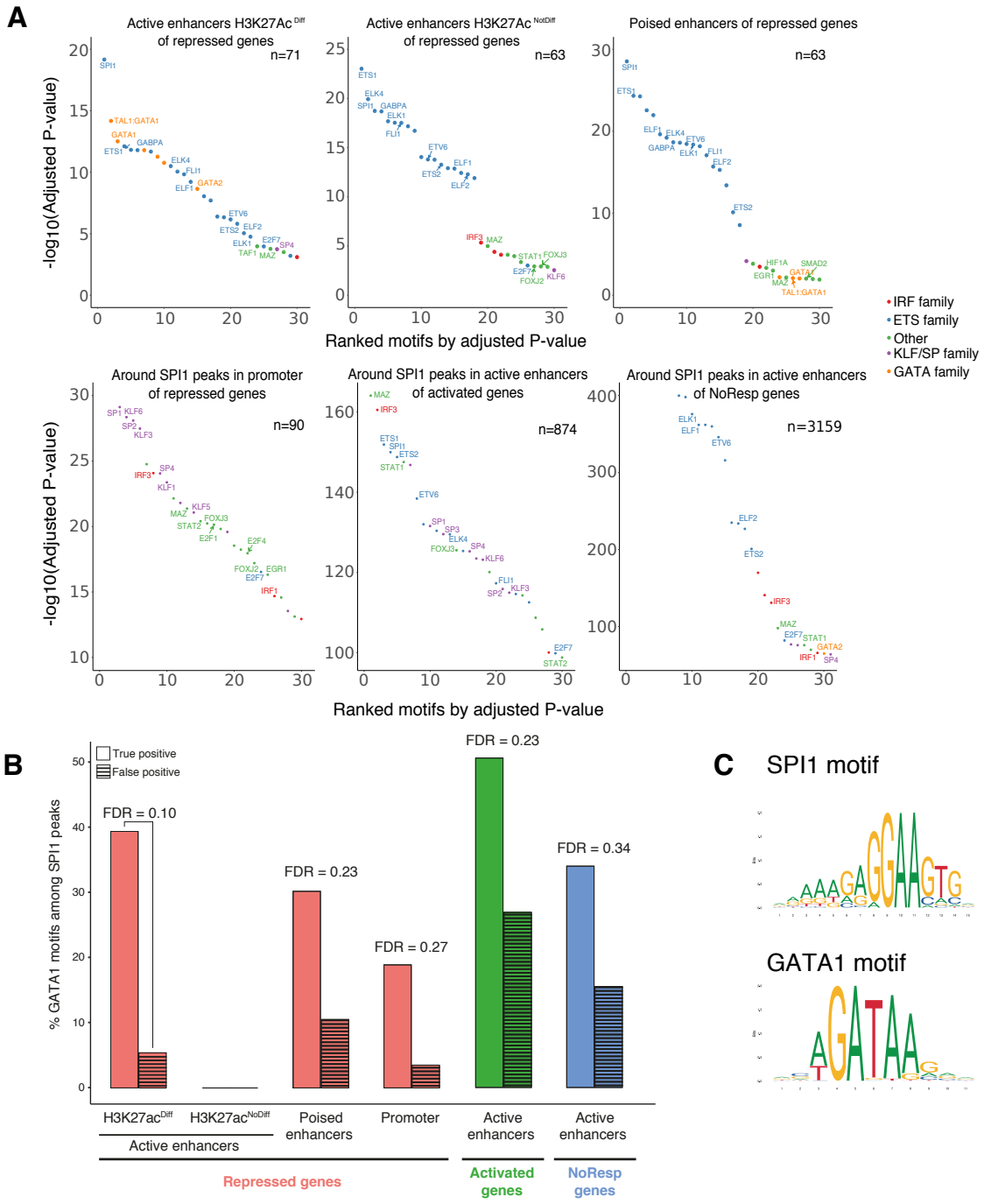


FIGURE 4.16 – **Le motif de GATA1 est spécifiquement enrichis au niveau des enhancers activés fixés par SPI1 dans les gènes réprimés présentant un différentiel pour H3K27ac. A.** Visualisation du top 30 des motifs obtenus par l'analyse d'enrichissement de motifs réalisée avec AME de la suite MEME [132] sur les régions allant de +/- 150bp autour de la position peakmax de SPI1 pour six catégories de peaks de SPI1. Les motifs des facteurs de transcription sont classés par ordre croissant de p-value ajustées. Seuls les noms des motifs correspondant aux protéines exprimées dans les érythroblastes sont indiqués. **B.** Pourcentages de séquences input (+/- 150bp depuis la position peakmax de SPI1) catégorisées comme True Positive (TP) et False Positive (FP) pour le motif de GATA1 par AME, sont indiquées les valeurs de FDR pour chacune des catégories. Formule du FDR : FDR = FP/(TP+FP). **C.** Motifs de SPI1 et GATA1 issus de la database JASPAR [165].

Les motifs impliquant GATA (TAL1 :GATA1 ou GATA1 seul) sont les motifs les plus enrichis et cela uniquement dans le voisinage immédiat des pics de SPI1 qui sont situés dans les enhancers différentiels pour H3K27ac ($H3K27ac^{Diff}$) dans les gènes réprimés. Les motifs de SPI1 et de GATA1 sont semblables dans la mesure où ils sont tous deux riches en A et G, toutefois le motif de GATA contient une thymidine en son coeur ce qui n'est pas le cas de SPI1 (Fig. 4.16C). 39% des gènes réprimés, dé-acétylés par SPI1 fixé dans les enhancers, ont un motif GATA1 (Fig. 4.16B, $FDR = 0.01$). GATA1 est le facteur de transcription majeur de la différenciation érythroïde. SPI1 et GATA1 interagissent physiquement [166]. Il a été montré que SPI1 est capable de bloquer l'activité transcriptionnelle de GATA1 en se fixant sur GATA1 à la chromatine. Deux modèles ont été proposés. Le premier propose que SPI1, fixé sur GATA1, entrainerait le dépôt de H3K9me3, marque répressive, au niveau des promoteurs ou enhancers des gènes cibles de GATA1 [92]. Le second propose que SPI1 empêcherait l'interaction de CBP/P300 avec GATA1 et inhiberait ainsi l'acétylation de GATA1 par CBP, entraînant une diminution de sa liaison à l'ADN ou de sa capacité à trans-activer [90]. De plus, certaines données indiquent que SPI1 pourrait empêcher la liaison de GATA1 à la chromatine mais ceci est controversé [90, 93, 167].

Nous sommes donc loin de comprendre l'interaction fonctionnelle, si elle existe, entre SPI1 et GATA1. Nous avons déterminé si SPI1 est effectivement fixé **sur** GATA1 comme proposé par ces études, ou si les deux protéines se fixent à l'ADN **indépendamment**.

Pour cela, j'ai analysé des données de ChIP-seq GATA1 [151] dans des cellules érythroleucémiques MEL très proches des cellules pré-leucémiques utilisées dans cette étude. La recherche des sites de liaison de la protéine GATA1 a été réalisée en utilisant HMCan [124], un seuil (>5) sur le score des pics a été appliqué. Aucune normalisation n'a été effectuée puisque nous n'avons utilisé que les coordonnées des sites de liaison de GATA1. En effet, J'ai cherché si dans les gènes avec un pic de SPI1, il existait un pic de GATA1. Nous trouvons que 39 des 71 (soit 55%) régions fixées par SPI1 dans les enhancers présentant un différentiel pour H3K27ac ($H3K27ac^{Diff}$) possèdent également, dans leur voisinage, un pic GATA1 (Tab. 4.3). Ce pourcentage est largement supérieur à celui des cinq autres catégories de régions (Tab. 4.3). Une fois les gènes présentant à la fois un pic de SPI1 et un pic de GATA1 identifiés, j'ai cherché si il existait un motif de SPI1 dans le pic de SPI1 et un motif de GATA1 dans le pic de GATA1 (Fig. 4.17A). Quatre cas de figures sont possibles (Fig. 4.17B) :

- **Cas 1 :** Motif SPI1 dans pic SPI1, Motif GATA1 dans pic GATA1.
- **Cas 2 :** Motif SPI1 dans pic SPI1, Pas de motif GATA1 dans pic GATA1.
- **Cas 3 :** Pas de motif SPI1 dans pic SPI1, Motif GATA1 dans pic GATA1.
- **Cas 4 :** Pas de motif SPI1 dans pic SPI1, Pas de motif GATA1 dans pic GATA1.

Dans le **cas 1**, si les motifs de GATA1 et de SPI1 étaient assez éloignés, nous avons conclu à une liaison indépendante des deux facteurs de transcription à l'ADN. Nous avons déterminé que parmi les 39 pics de SPI1 dans des régions différentielles pour H3K27ac ($H3K27ac^{Diff}$) dans

les gènes réprimés présentant dans leur voisinage un pic de GATA1, 33 (soit 85%) présentent à la fois un motif de SPI1 dans le pic SPI1 et un motif GATA1 dans le pic GATA1, c'est à dire le **cas 1** (Fig. 4.17B, Fig. 4.17C).

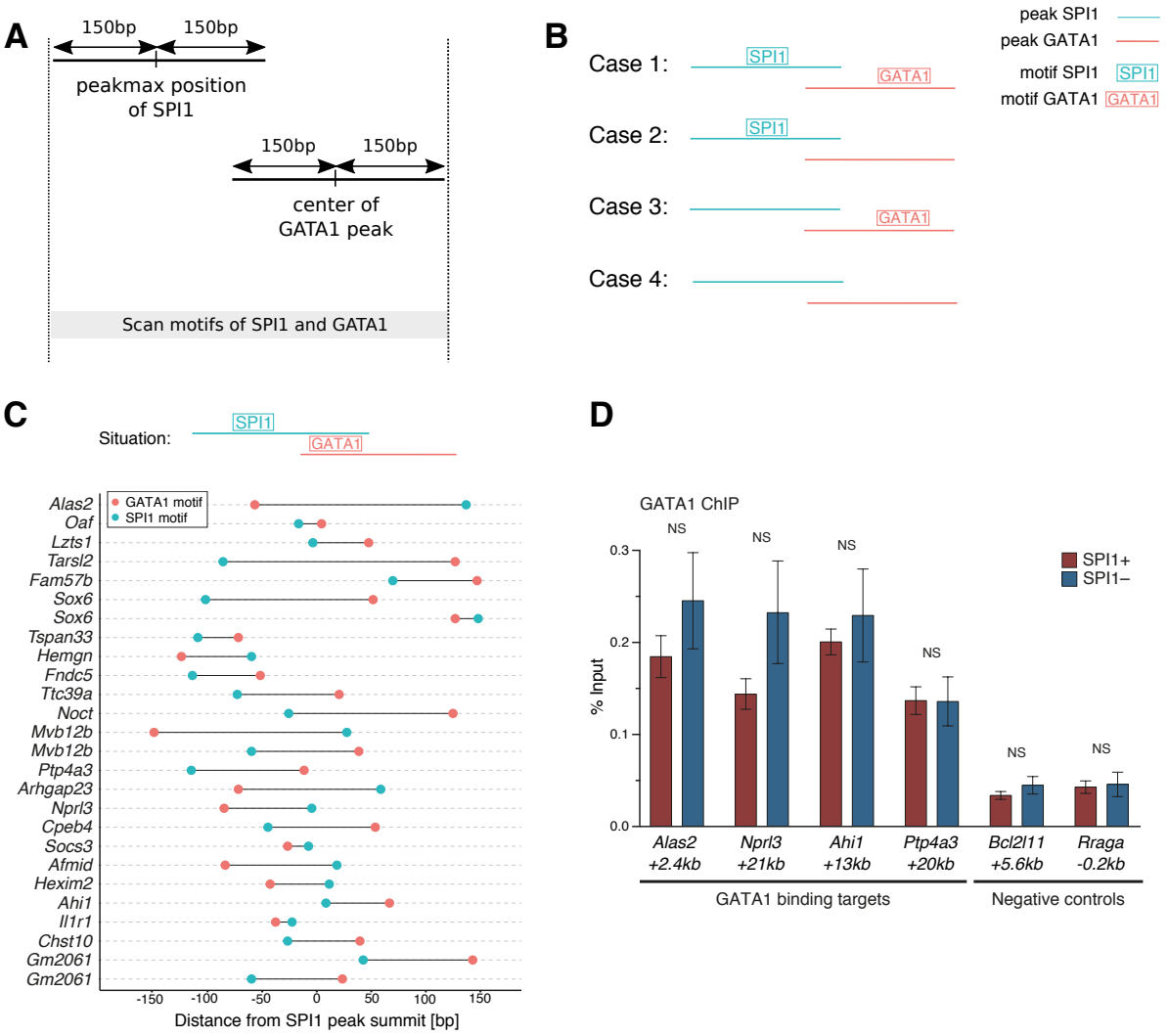


FIGURE 4.17 – **SPI1 se trouve à côté de GATA1 dans les enhancers activés fixés par SPI1 dans les gènes réprimés présentant un différentiel pour H3K27ac. A.** Schéma représentant la région dans laquelle la recherche de motifs de GATA1 et SPI1 est opérée en utilisant l'outil FIMO de la suite MEME [132]. **B.** Ensemble (non-exhaustif) des cas possibles pour les occurences des motifs de SPI1 et GATA1. **C.** Localisation des motifs de SPI1 et de GATA1 dans le cas où le motif de GATA1 se trouve sur le pic de SPI1 (n=26). **D.** ChIP GATA1 dans les conditions SPI1+/SPI1- analysé par qPCR en temps réel dans les enhancers fixés par SPI1 (la distance indiquée est la distance au TSS) sur des gènes extraits des données de la Fig. 4.17C. L'enrichissement est donné en moyenne du pourcentage input +/- SEM des triplicats.

Catégorie de pics SPI1	Pourcentage de pics SPI1 qui overlappent avec un pic GATA1
$H3K27ac^{Diff}$ - gènes réprimés	55 %
$H3K27ac^{NoDiff}$ - gènes réprimés	32 %
Enhancers Poised - gènes réprimés	11 %
Promoteurs - gènes réprimés	21 %
Enhancers actifs - gènes activés	25 %
Enhancers actifs - gènes NoResp	21 %

TABLE 4.3 – **Tableau présentant le pourcentage de pics SPI1 overlappant un peak GATA1**

L'ensemble de ces résultats suggèrent que :

— **La moitié des gènes qui sont réprimés par SPI1 pour lesquels SPI1 est capable de diminuer fortement le niveau d'acétylation ($H3K27ac^{Diff}$) présentent un pic de GATA1 proche du pic de SPI1.**

— **Dans la majorité des cas, SPI1 se fixe indépendamment de GATA1 à la chromatine (Fig. 4.16C).**

Des expériences de ChIP GATA1 dans les cellules pré-leucémiques cultivées ou non en présence de dox montrent que SPI1 ne change pas significativement le niveau de liaison de GATA1 à la chromatine (Fig. 4.16D). Ceci indique que malgré une co-localisation des pics de GATA1 et de SPI1, SPI1 ne réduit pas la stabilité de GATA1 à la chromatine. Toutefois, comme SPI1 interagit avec HDAC1, il est envisageable que SPI1 conduise indirectement à la dé-acétylation de GATA1 et à son inactivation transcriptionnelle [168].

4.2.9 Rôle de PRC2 dans la répression transcriptionnelle dûe à l'activité de SPI1

Des données récentes indiquent que la marque H3K27me3 et le complexe Polycomb qui contrôle la méthylation de la lysine 27 de l'histone H3 maintiennent l'état réprimé d'un gène [169] mais n'est pas initiateur de la répression. Bien que le mode de recrutement de PolyComb soit controversé, son modèle de recrutement le plus récent implique la bi-affinité du complexe PRC2 (Polycomb Repressive Complex 2 qui contient l'activité histone méthyl-transférase) pour l'ARN et la chromatine. Une compétition active entre la chromatine et l'ARN déterminerait les sites de liaison de PRC2. La diminution d'ARN nouvellement synthétisé lors de la répression des gènes permet le recrutement de PRC2 à la chromatine et l'amplification de la méthylation au niveau de la lysine 27 de l'histone H3 [170, 171]. Ce modèle implique donc que PRC2 n'est pas initiateur de la répression mais serait capable d'agir dans un contexte d'un gène faiblement transcrit pour maintenir la répression en déposant la marque H3K27me3. Cette marque serait associée à une répression plus forte si elle implique la position TSS [84]. L'équipe de Christel Guillouf a précédemment démontré que SPI1 et PRC2 interagissent physiquement et que la

répression de l'expression du gène *Bcl2l11* nécessite PRC2 dans les cellules pré-leucémiques TgSpi1 [66]. Afin de comprendre si l'action de PRC2 est de renforcer la répression exercée par l'interaction HDAC1 et SPI1 et si SPI1 augmente l'effet répressif de PRC2, comme le suggère leur interaction, nous avons analysé la présence de H3K27me3 sur l'ensemble des gènes réprimés pour lesquels SPI1 diminue H3K27ac. Les profils de densité de la marque d'histone H3K27me3 aux sites de liaison de SPI1 en fonction du niveau de SPI1 montre que la présence de SPI1 est anti-corrélée avec le niveau de la marque H3K27me3 (Fig. 4.18A). Toutefois, on observe qu'au niveau des TSS des gènes réprimés pour lesquels SPI1 réduit fortement H3K27ac, la présence de SPI1 augmente le niveau de H3K27me3 (Fig. 4.18B - profil de gauche) tandis que ce n'est pas le cas pour les gènes NoResp (Fig. 4.18B - profil de droite). De plus, on remarque que les gènes réprimés liés par SPI1 possèdent un niveau plus élevé de H3K27me3 dans la condition SPI1+ que les gènes réprimés qui ne sont pas fixés par SPI1 (Fig. 4.18B - profil du milieu, condition SPI1+).

L'ensemble de ces résultats montrent que :

- **La répression des gènes est associée à une augmentation de H3K27me3 de façon globale, comme attendu.**
- **La marque H3K27me3 est augmentée lorsque SPI1 est fixé à la chromatine des gènes réprimés en comparaison aux gènes réprimés non fixés par SPI1, suggèrant que SPI1 exerce un effet amplificateur de la répression en agissant sur PRC2.**

Ces données suggèrent que HDAC1 et PRC2 pourraient coopérer à la répression transcriptionnelle due à la surexpression de SPI1. Nous avons testé cette hypothèse en analysant les conséquences transcriptionnelles de l'inhibition de PRC2 et/ou HDAC1 sur quatre gènes réprimés et fixés par SPI1 à leurs enhancers actifs (*Bcl2l11, Alas2, St3gal6, Ptp4a3*). Ces gènes codent pour des protéines de la différenciation ou de la survie des cellules érythroïdes. Nous avons comparé leurs réponses avec celles de deux gènes activés et liés par SPI1 (*Cdkn2c, Spi1*). Les cellules pré-leucémiques ont été cultivées en présence de l'inhibiteur pharmacologique de EZH1/2 : l'UNC1999 et/ou l'inhibiteur de HDAC1/3 : l'Entinostat. Les doses utilisées sont très faibles afin de permettre un traitement sur 48h sans effet drastique sur la survie des cellules. Les ARN ont été extraits et leurs expressions ont été mesurées par RT-qPCR (Fig. 4.18C). Excepté le gène *Alas2*, les traitements avec l'UNC1999 **ou** l'Entinostat seuls ont montré pas ou peu d'effet sur la transcription des gènes. En revanche, le traitement combiné a induit une forte augmentation de la transcription des gènes réprimés par SPI1, alors qu'il n'a aucun effet sur les gènes activés par SPI1. Ces résultats montrent un effet synergiques des deux voies sur la répression transcriptionnelle des gènes cibles de SPI1.

Nous n'avons pas observé d'effet de la répression synergique des deux voies sur la réinduction de la différenciation érythroïde bloquée par SPI1 (données non montrées). En revanche, nous montrons clairement un effet synergique des deux inhibiteurs sur la mortalité et le nombre de cellules pré-leucémiques vivantes, mesurés à différentes doses ou sur le temps par

calcul de synergie comme décrit dans [172] (Fig. 4.18D et données non montrées).

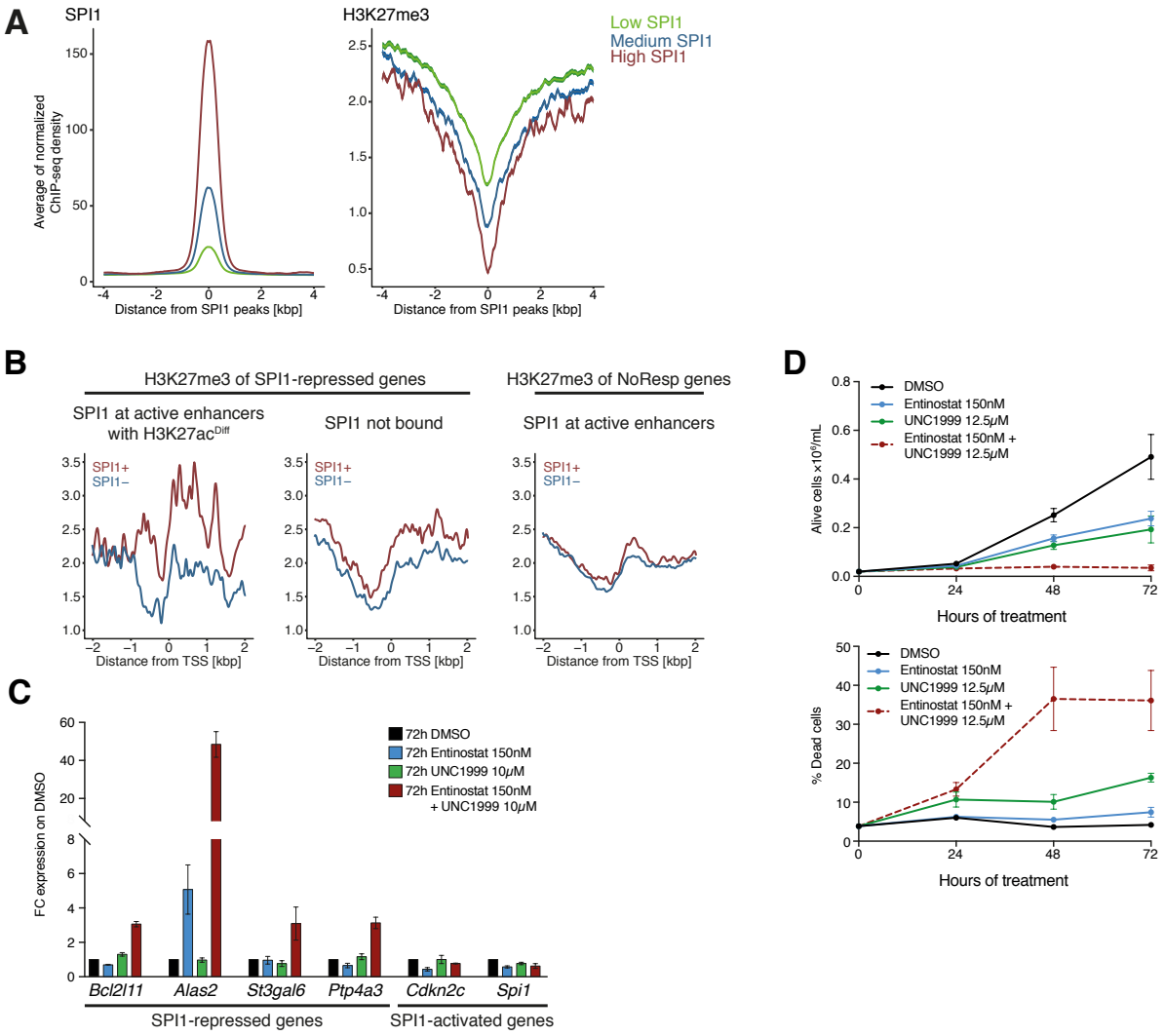


FIGURE 4.18 – **PRC2 et HDAC1 synergisent pour réprimer les gènes cibles de SPI1. A.** L'intensité du signal ChIP-seq de SPI1 et H3K27me3 a été divisée en trois groupes en fonction du signal ChIP-seq de SPI1 en utilisant l'algorithme des k-means (niveau de SPI1 faible, moyen, élevé SPI1) et centré sur le sommet du pic de SPI1. **B.** Profils de densité moyenne de H3K27me3 dans les conditions SPI1+ et SPI1- centrés sur les TSSs des gènes réprimés et fixés par SPI1 dans les enhancers actifs différentiels pour H3K27ac (gauche), des gènes réprimés par SPI1 mais non fixés par SPI1 (milieu) et des gènes NoResp fixés par SPI1 dans les enhancers actifs (droite). **C.** Les cellules ont été traitées avec l'Entinostat 150nM, l'UNC1999 12.5 μM ou la combinaison des deux inhibiteurs jusqu'à 72h. La moyenne +/-SEM du nombre de cellules vivantes (haut) et de la mortalité (bas) correspond au test au bleu trypan de 3 expériences. **D.** L'expression des gènes réprimés par SPI1 analysé dans des cellules traitées traitées avec l'Entinostat 150nM ou l'UNC1999 10μM ou la combinaison des deux inhibiteurs pendant 72h a été évaluée par RT-qPCR, FC +/-SEM par rapport aux cellules non traitées normalisée sur l'ARNm de *PolR2a*.

Chapitre 5

Caractérisation de la liaison de SPI1 à l'ARN et conséquences

Dans la section précédente je me suis intéressée à la première thématique de mon projet de thèse qui concerne le rôle de SPI1 dans la répression génique. Dans cette section je vais me concentrer sur la seconde thématique dont **le but est de caractériser la liaison de SPI1SPI1 à l'ARN et les conséquences de cette liaison sur l'expression génique**. Pour cela, je vais dans un premier temps présenter les stratégies et les méthodes adoptées pour l'analyse des données de CLIP-seq. En effet, comme je l'ai mentionné dans l'introduction (section 1.3.5), l'alignement des lectures et la recherche des sites de liaison d'un facteur de transcription sur l'ARN est une tâche complexe au cours de laquelle plusieurs paramètres sont à prendre en compte. Il m'a donc semblé important de dédier une section complète à ce travail. Dans un second temps, je présenterai les résultats que j'ai obtenus

5.1 Méthodes d'analyse des données de liaison de SPI1 à l'ARN

5.1.1 Alignement des lectures de CLIP-seq en deux étapes

L'alignement des données de CLIP-seq est une tâche complexe, les lectures peuvent provenir d'ARNm pré-mature (constitué à la fois d'exons et d'introns) et d'ARNm mature (ayant subi l'épissage, et incluant principalement les exons). Ainsi lors de l'alignement des lectures provenant d'ARNm mature il faut que le logiciel d'alignement soit capable d'insérer des gaps sur toute la longueur de l'intron pour produire un alignement correct sur les exons et de ne pas forcer l'alignement de la lecture sur l'intron. D'autre part, la reverse transcriptase, qui synthétise les molécules d'ADNc à partir de molécules d'ARN, peut rencontrer des problèmes de synthèse au niveau des régions de liaison de la protéine d'intêret. En effet si un morceau de

protéine reste sur l'ARN, alors la reverse transcriptase ne pourra pas synthétiser de façon correcte les nucléotides pour l'ADNc à ces positions puisque l'information de la base nucléique à synthétiser sera obstruée par la présence de la protéine. Ainsi, des erreurs peuvent être insérées dans l'alignement. Ceci implique que le logiciel d'alignement doit être capable de répérer ces régions où l'information de la base nucléique correcte est manquante afin d'y insérer des délétions, des insertions (communément appelés indels) ou des mutations dans l'alignement qui se révèleront importantes dans la suite de l'analyse. Elles seront, en effet, de potentiels sites de liaison de la protéine.

Une recherche bibliographique approfondie m'a amenée à opter pour une stratégie d'alignement en deux étapes [173] utilisant deux logiciels d'alignements différents :

— NovoAlign [174]

— BWA [122]

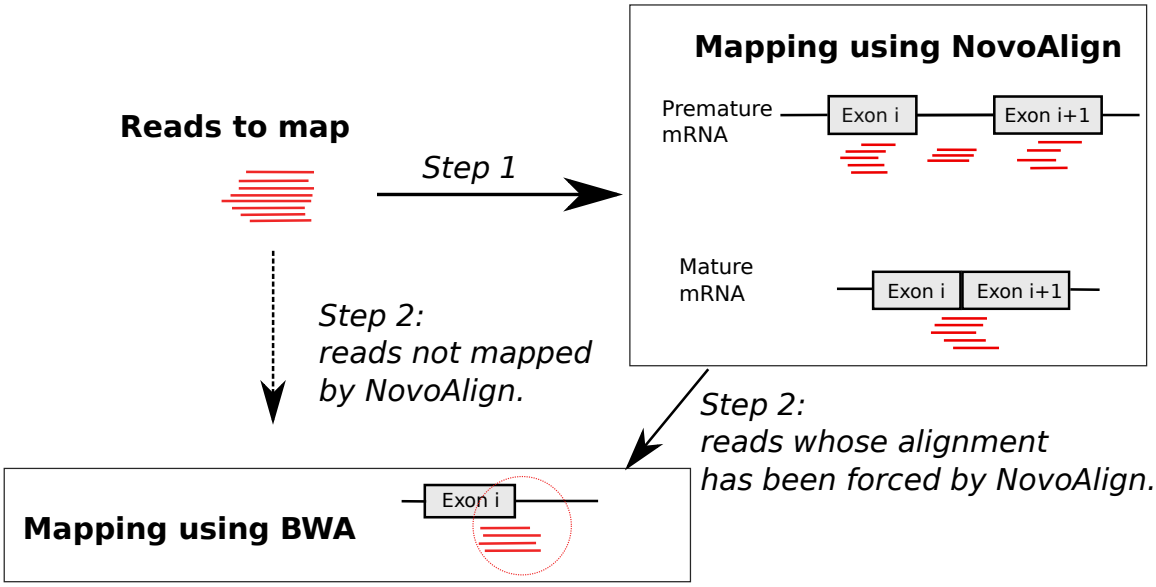


FIGURE 5.1 – **Alignement des lectures de CLIP-seq en deux étapes. Étape 1 :** Alignement séparément sur les introns et les exons en utlisant NovoAlign. **Étape 2 :** Alignement sur les jonctions exons-introns en utilisant BWA.

Alignement des lectures uniquement exoniques ou uniquement introniques avec NovoAlign

Le logiciel NovoAlign permet d'aligner **séparément** les lectures sur les exons et sur les introns (Fig. 5.1). Le transcriptome et les junctions exons-exons ont été définis en utilisant la suite USeq [175] et les programmes *MakeTranscriptome* et *MaskExonsInFastaFiles*. *MakeTranscriptome* prend en entrée :

— les fichiers fasta des chromosome du génome de référence (mm10).

— une table avec une annotation des gènes.

et construit le transcriptome avec les jonctions d'épissage, c'est à dire les jonctions exons-exons. *MakeExonsInFastaFiles* permet de masquer les séquences exoniques. Les commandes utilisées sont présentées ci-dessous :

```
java -jar MakeTranscriptome
-f Mouse/mm10/chromosomes
-u Mouse/mm10/genesAnnotation_mm10.txt -r 45 -n 60000
```

L'option -r concerne la longueur des séquences en paires de bases et l'option -n donne le nombre maximum d'épissage par transcrit.

```
java -Xmx4G -jar MaskExonsInFastaFiles
-f Mouse/mm10/chromosomes
-u Mouse/mm10/genesAnnotation_mm10.txt
-s Mouse/mm10/MaskedFastas
```

Le génome est indexé en utilisant l'outil *novoindex* :

```
novoindex Transcriptome.nix
Mouse/mm10/MaskedFastas2/geneMaskedGenome.fasta
Mouse/mm10/mm10_oldRad45Num60kMin10Splices.fasta
Mouse/mm10/mm10_oldRad45Num60kMin10Transcripts.fasta
```

Après ces étapes préliminaires, les adaptateurs retirés, les lectures sont alignées sur le génome par NovoAlign. L'alignement est basé sur une table de hachage construite en divisant les lectures en oligomères qui se chevauchent. La phase d'alignement utilise l'algorithme de Needleman-Wunsch [176]. Pour commencer, les lectures sont placées sur les séquences indexées du génome supposées être leurs origines. Ensuite, ces localisations sont évaluées à chaque itération grâce à l'algorithme de Needleman et Wunsch, les pénalités sont modifiables par l'utilisateur. Dans le but de favoriser la détection des insertions et délétions, dont nous avons besoin pour la suite de l'analyse, la pénalité d'ouverture d'un gap, initialement égale à 40, a été diminuée à 20 et la pénalité d'extension d'un gap, initialement égale à 6, a été rapportée à 5. L'algorithme de Needleman et Wunsch permet de trouver l'alignement global optimal. L'outil NovoAlign prend en entrée le transcriptome indexé avec les jonctions exons-exons (Transcriptome.nix) et le fichier fasta contenant les lectures (A203S*.trimmed.fasta). La commande utilisée est la suivante :

```
novoalign  -d Transcriptome.nix
-f A203S*.trimmed.fasta -F FA -g 20 -x 5
 -o SAM -r Random > A203S*.sam 2> A203S*_stats.txt
```

Une fois l'étape d'alignement terminée, il faut revenir aux coordonnées génomiques, pour cela le programme SamTranscriptomeParser de la suite USeq [175] est utilisé.

```
java -jar SamTranscriptomeParser
-f A203S*.sam -a 500000
-n 100 -u -s A203S*_converted.sam
```

La sortie est au format SAM que l'on peut facilement convertir au format BAM, le format binaire de SAM. Le logiciel NovoAlign permet d'aligner correctement les lectures provenant d'une jonction exon-exon (Fig. 5.1) ainsi que les lectures provenant de régions introniques (Fig. 5.1). Les lectures dont l'alignement a été forcé (lectures alignée sur une jonction **exon-intron**) sont répérables grâce au CIGAR string en colonne 6 des fichiers au format BAM. La lettre S, pour soft-clipped, en début ou en fin de CIGAR string illustre ce phénomène. En effet, les nucléotides correspondant à un S dans le CIGAR string ont été supprimés de l'alignement, cela veut dire que l'alignement de ces lectures a été forcé. De plus, certaines lectures n'ont pas pu être alignées par NovoAlign. Ainsi on récupère ces lectures qui n'ont pas pu être alignées ainsi que celles qui présentent au moins 4 fois la lettre S en début ou en fin de CIGAR string, preuve que leur alignement a été forcé. Leur alignement est réalisé lors d'une seconde étape, en utilisant le logiciel BWA. En effet, ces lectures qui n'ont pas pu être alignées ou dont l'alignement a été forcé proviennent très probablement de molécules d'ARNm pré-mature dont l'alignement n'était pas possible avec le logiciel NovoAlign.

Alignement des lectures à la fois exoniques et introniques avec BWA

BWA [122] est un logiciel d'alignement qui permet comme NovoAlign d'aligner de courtes séquences sur un génome de référence. Ceci est cohérent avec la technologie de CLIP-seq qui produit majoritairement des lectures assez courtes, au maximum 50bp dans notre cas (Fig. 5.5). BWA permet un bon compromis entre rapidité et performance. Comme mentionné ci-dessus, le but est maintenant d'aligner les lectures qui sont très courtes et qui n'ont pas pu être alignées par NovoAlign ainsi que les lectures provenant d'ARNm pré-mature, c'est à dire constituées à la fois de régions introniques et exoniques. L'algorithme utilisé par BWA requiert la création d'un index :

```
bwa index -a bwtsw mm10.fa
```

La commande suivante permet l'alignement des lectures présentes dans le fichier A203S*.trimmed.toRealign.fq :

```
bwa mem  -t 7
 -T 10 BWA/indexes/mm10.fa A203S*.trimmed.toRealign.fq
 | samtools view - -S -b > A203S*.trimmed.Realigned.bam
```

L'option -T 10 permet de n'avoir en sortie que les lectures qui ont été alignées avec un score au moins supérieur à 10. Ce score est calculé à partir des pénalités pour un match (+1), pour un mismatch (-4), pour l'ouverture d'un gap (-6) et pour l'extension d'un gap (un gap de longueur k coûte $(-1-k)$). La sortie est renvoyée vers le programme *view* de la suite Samtools [177] afin de convertir le fichier de sortie au format BAM. Le résultat de l'alignement par NovoAlign (duquel ont été retirées les lectures ré-alignées avec BWA) et le résultat de l'alignement par BWA sont alors rassemblés dans un seul fichier BAM. Cette stratégie d'alignement permet d'aligner le maximum de lectures en utilisant une stratégie spécifique à la technologie de CLIP-seq [173].

Un contrôle qualité de l'alignement des lectures est effectué, pour cela un seuil de 20 est appliqué sur le champ MAPQ des fichiers bam. Les duplicats (de PCR) sont retirés.

5.1.2 Recherche des pics de SPI1 sur l'ARN pour la détermination de ses sites de liaison sur l'ARN

La recherche des sites de liaison de SPI1 est appelée recherche de pics et permet d'identifier les régions du génome qui présentent un nombre de lectures plus élevé que les autres. Plusieurs paramètres sont à prendre en compte, en effet, la technologie de CLIP-seq induit des biais dont il faut s'extraire pour ne pas avoir un trop grand nombre de faux positifs qui pourraient amener à des conclusions biologiques fausses.

Impact de la quantité de transcrit dans le signal de CLIP-seq

Murigneux et al [173], Saulière et al [142] suggèrent que la quantité de transcrit présente lors de la liaison de la protéine d'intérêt à l'ARN *in vivo* induit un biais dans la quantité de liaison de la protéine sur les molécules d'ARN. Ainsi il faudrait prendre en compte la quantité de transcrit lors que de la recherche des zones de liaison de la protéine d'intérêt sur l'ARN. Je me suis donc intéressée à la corrélation qu'il peut exister entre le signal de CLIP-seq Spi1+ et le signal de RNA-seq Spi1+, pour cela j'ai utilisé la valeur de FPKM (Fragments Per Kilobase per Million reads mapped) de RNA-seq Spi1+ et le nombre de lectures par gène en CLIP-seq Spi1+. Ce nombre de lectures par gène a été calculé en utilisant l'outil "HTSeq-count" [178], la valeur est normalisée par mille fois la longueur du gène pour rester cohérent en terme d'échelle avec la valeur de FPKM utilisée pour les données de RNA-seq Spi1+. La Fig. 5.2 représentant en abscisse :

$$log2\left(\frac{(nbrReads_{CLIP-seqSpi1+})}{(longueurGeneKilobase)}\right) \tag{5.1}$$

et en ordonnée :

$$log2\,(FPKM_{RNAseqSpi1+}) \tag{5.2}$$

m'a orientée dans le choix de la méthode de recherche des sites de liaisons de SPI1 sur l'ARN. En effet, il existe une corrélation ($R = 0.45$, calculée sur l'ensemble des points) entre le signal

de CLIP-seq Spi1+ et le signal de RNA-seq Spi1+ (dont témoigne la zone 2 sur la Fig. 5.2) mais elle est relativement faible et il existe une partie des données qui est en faveur du fait que le signal de CLIP-seq Spi1+ n'est pas entièrement déterminé par la quantité de transcrit (zones 1 et 3 sur la Fig. 5.2). Ainsi, il me fallait tenir compte du biais induit par la quantité de transcrit lors de ma recherche des sites de liaison de SPI1 sur l'ARN.

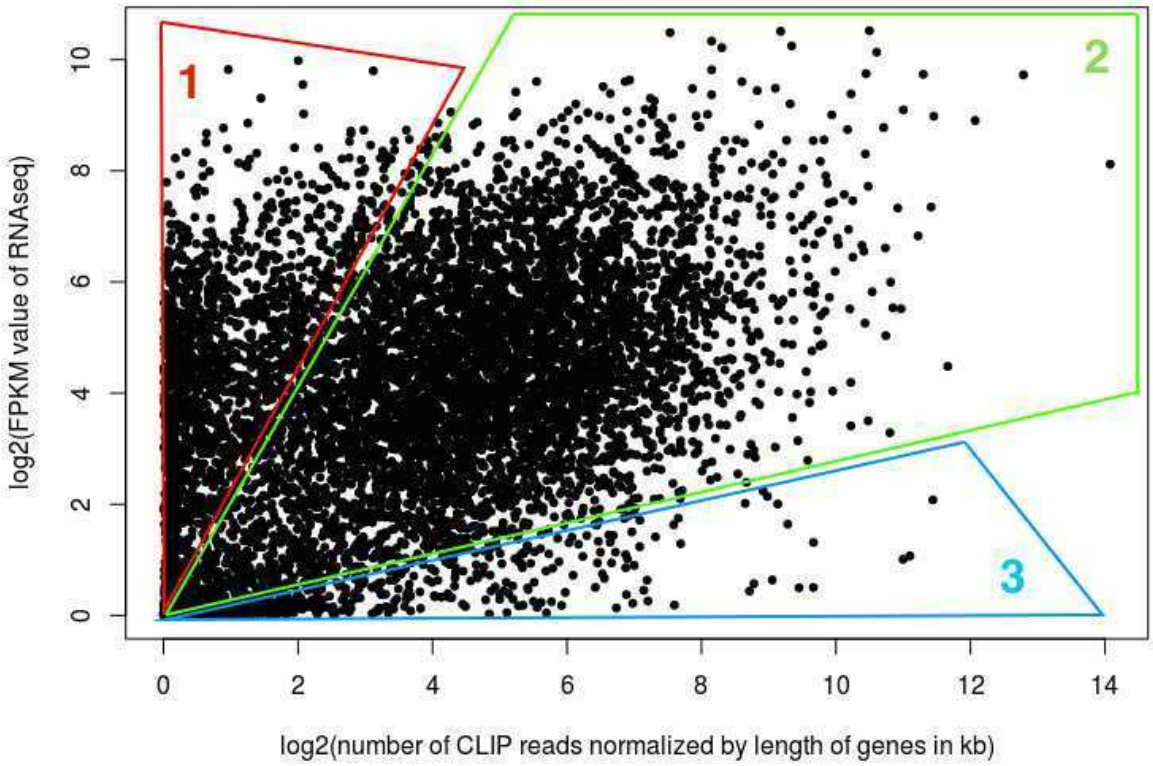


FIGURE 5.2 – **Corrélation des signaux de RNA-seq et de CLIP-seq.** La corrélation est égale à 0.45, toutefois les zones 1 et 3 sont en faveur du fait que le signal de CLIP-seq n'est pas entièrement déterminé par le signal de RNA-seq.

Importance des insertions, délétions insérées dans l'alignement des données de CLIP-seq

Le logiciel Piranha [180] est un logiciel de peak calling spécialement développé pour la recherche des sites de liaison de protéines sur l'ARN, l'utilisation de covariables permet de prendre en compte un ou plusieurs paramètres supplémentaires dans la recherche des sites de liaison. Toutefois un logiciel plus récent, omniCLIP [179], permet de prendre en compte **automatiquement** un autre biais induit par la technologie de CLIP-seq. En effet, dans la section 5.1.1 j'ai mentionné l'importance des délétions ou des insertions (indels) qui peuvent se trouver dans l'alignement. Ces indels sont les marqueurs de potentiels sites de liaison de la protéine d'intérêt. Le logiciel omniCLIP [179] se propose de repérer ces indels qui sont dénommés **évènements diagnostiques** dans la suite du texte. J'ai donc choisi ce logiciel car il permet de

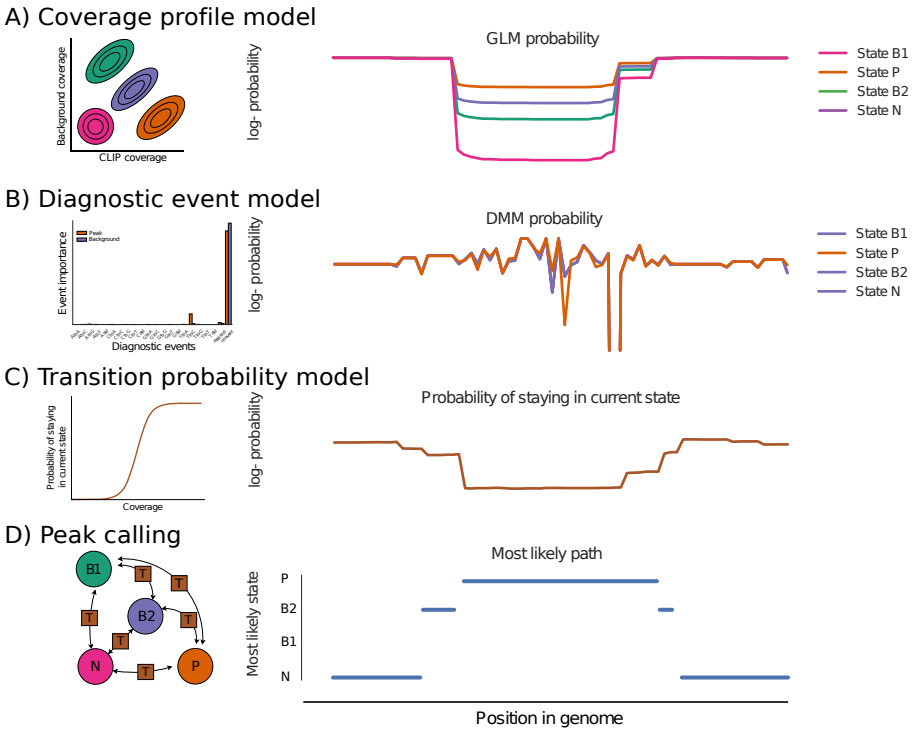


FIGURE 5.3 – **Illustration de la stratégie de omniCLIP (extrait de [179]). A.** La probabilité de chaque position et de chaque état est calculée à l'aide du profil de couverture, **B.** et des évènements diagnostiques. **C.** Les probabilités de transition sont calculées sur la base de la couverture à chaque position. **D.** Un modèle de Markov cache non homogène est appliqué pour séparer les régions de pics (P) et les régions non-pics (N, B1, B2).

prendre en compte automatiquement deux des biais introduits par la technologie de CLIP-seq. Une attention toute particulière doit être portée à la qualité du peak calling effectué sur les données de CLIP-seq afin d'éviter d'avoir trop de sites de liaison de SPI1 qui seraient de faux positifs menant à de fausses conclusions biologiques.

OmniCLIP [179] identifie les sites de liaison d'une protéine sur l'ARN en segmentant le génome de façon non supervisée en régions pics et non pics. Pour cela, un modèle de Markov caché non-homogène (NHMM) à quatre états est utilisé. Les lectures observées dans l'ensemble des librairies, CLIP-seq et RNA-seq (qui sera notre background), sont modélisées à l'aide du NHMM à quatre états :

- un état pour modéliser les régions qui sont des pics (**P**) : signal CLIP-seq > signal background
- deux états pour modéliser les régions où le signal du background est équivalent au signal de CLIP-seq (**B2**) ou plus grand que le signal de CLIP-seq (**B1**)
- un état pour modéliser les régions avec très peu de couverture dans l'ensemble des librairies (**N**)

Les **profils de couverture du CLIP-seq** et du background sont modélisés en utilisant l'hypo-

thèse que la couverture à chaque position du génome suit une distribution Binomiale Négative qui est déterminée par :

— la taille de la librairie
— l'expression des gènes
— l'état de la position (**P**, **B1**, **B2**, **N**)

La dépendance entre ces trois variables est modélisée en utilisant un modèle linéaire généralisé (GLM) (Fig. 5.3A).
Les **évènements diagnostiques** sont modélisés par un modèle mixte de Dirichlet multinomial. L'hypothèse est la suivante : les pics sont une mixture de plusieurs classes de positions qui ont un taux différent d'évènements diagnostiques. Dix classes sont construites et pour chaque classe le nombre d'évènements diagnostiques est modélisé par un modèle hiérarchique multinomial de Dirichlet (Fig. 5.3B).

La probabilité d'émission d'un état s du NHMM à une position i du génome est donnée par le produit de la probabilité de couverture $p(X_i|s)$ et la probabilité des évènements diagnostiques observée $p(Y_i|s)$.
Les **probabilités de transition** entre les états sont modélisés en utilisant une fonction logistique du profil de couverture X_i à chaque position i. Cela permet au modèle d'être plus ou moins rigide en fonction de la quantité de données disponibles à chaque position (Fig. 5.3C).
Les pics sont ensuite déterminés comme l'ensemble des régions consécutives pour lesquelles l'état le plus probable dans le NHMM est **P** en utilisant l'algorithme de Viterbi (Fig. 5.3D).

5.1.3 Recherche d'un motif de liaison de SPI1 sur l'ARN

Une fois que les sites de liaison de SPI1 sur l'ARN ont été déterminés, j'ai recherché un potentiel motif de SPI1 sur l'ARN.

Dans le but de ne réaliser la recherche de motifs que dans les vraies régions de liaison de SPI1, j'ai décidé d'être très stringeante dans leur sélection. Pour cela, j'ai utilisé l'un des avantages d'OmniCLIP [179], en effet, les évènements diagnostiques permettent de définir la position la plus probable pour la liaison de la protéine sur l'ARN (symbolisé par un trait vertical bleu sur la Fig. 5.4A). Ainsi, nous avons conservé uniquement les sites de liaison de SPI1 qui étaient présents dans les 2x2 réplicats biologiques et dont les distances des positions les plus probables étaient inférieures à 10bp (Fig. 5.4B). Ensuite pour chacun des sites de liaison conservés la position la plus probable pour les 2x2 réplicats biologiques est évaluée comme la moyenne des quatre positions les plus probables (Fig. 5.4C) et les dix premières paires de bases sont conservées de chaque côté.

Deux stratégies ont été adoptées pour réaliser la recherche de motifs, dans un premier temps l'outil AME de la suite MEME [132] a été utilisé dans le but de faire de la recherche de motif connus enrichis. En effet, rien n'excluait l'hypothèse selon laquelle le motif de liaison de SPI1

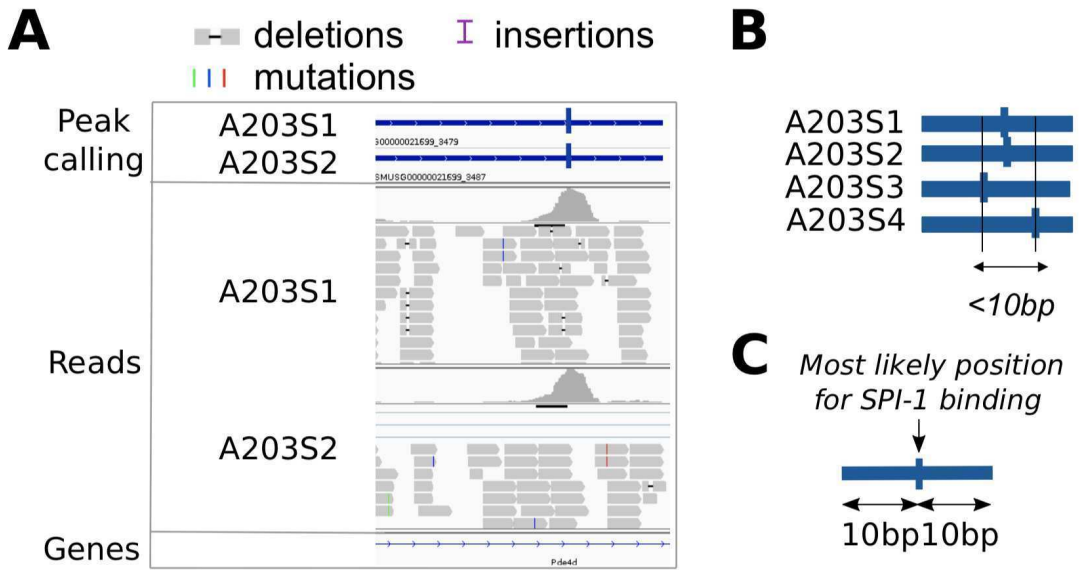


FIGURE 5.4 – **Illustration de la stratégie adoptée pour la recherche de motif de liaison de SPI1 sur l'ARN. A.** Capture d'écran IGV montrant un site de liaison de SPI1 sur deux réplicats techniques, le trait vertical bleu illustre la position la plus probable pour SPI1 déterminée par omniCLIP. **B.** Illustration d'un site de liaison de SPI1 dont la position la plus probable est à une distance inférieure à 10bp dans les 2x2 réplicats biologiques. **C.** La position conservée est la moyenne des quatre positions les plus probables dans les 2x2 réplicats biologiques et celle-ci est étendue de +/- 10bp pour la recherche de motifs.

à l'ARN était identique à celui sur l'ADN. Pour utiliser l'outil AME de la suite MEME [132] il nous a fallu rallonger les séquences de part et d'autre du site de liaison le plus probable de SPI1 à 30bp de chaque côté. Dans un second temps, nous avons recherché les motifs *de novo*, en utilisant le logiciel Zagros [181] spécifiquement implémenté pour la recherche de motifs à partir de données de CLIP-seq. Zagros possède quatre modes opératoires différents :

- en utilisant l'information de séquence uniquement.
- en utilisant l'information de séquence et de structure secondaire de l'ARN.
- en utilisant l'information de séquence et l'information d'évènements diagnostiques.
- en utilisant l'information de séquence, de structure secondaire de l'ARN et l'information de diagnostic events.

Dans la mesure où les séquences sur lesquelles la recherche de motif est réalisée ne font que 20bp en longueur l'information de structure secondaire de l'ARN n'est pas disponible. De plus, l'utilisation du logiciel OmniCLIP [179] a permis de réaliser une sélection des potentiels sites de liaison en incluant déjà l'information sur les évènements diagnostiques. La stratégie que j'ai adoptée pour sélectionner les régions sur lesquelles la recherche de motifs est réalisée a permis de privilégier la présence des évènements diagnostiques dans ces régions. Ainsi, nous avons décidé d'utiliser le premier mode opératoire de Zagros [181] qui consiste à ne donner en input

que l'information de séquence.

5.1.4 Gestion des réplicats biologiques

Nous disposons de deux réplicats biologiques pour la diminution de SPI1 au sein desquels deux réplicats techniques de CLIP-seq Spi1+ ont été réalisés. Le premier réplicat biologique pour la diminution de SPI1 s'appelle **7B3** et le second **722**. Dans la section de résultats, j'ai séparé les résultats sur les réplicats 7B3 (**A203S1** et **A203S2**) et sur les réplicats 722 (**A203S3** et **A203S4**). En particulier pour les résultats d'alignement je n'ai pas rassemblé les réplicats techniques (A203S1 avec A203S2 et A203S3 avec A203S4). En revanche pour caractériser la liaison de SPI1 sur l'ARN, je n'ai gardé que les pics qui étaient retrouvés dans les deux réplicats techniques. Ainsi pour le réplicat biologique 7B3, un peak est conservé s'il est retrouvé à la fois dans le réplicat technique A203S1 et dans le réplicat technique A203S2. Pour le réplicat biologique 722, un pic n'est conservé que s'il est retrouvé dans les réplicats techniques A203S3 et A203S4. J'ai considéré qu'un pic était retrouvé dans deux réplicats techniques si les deux pics avaient un recouvrement d'au moins 1 bp.
Pour l'étude du lien entre la liaison à l'ADN et à l'ARN j'ai uniquement utilisé le réplicat biologique 7B3 puisque les données de la liaison de SPI1 à l'ADN ne sont disponibles que pour ce réplicat.

5.2 Résultats

5.2.1 Qualité de l'alignement des données de CLIP-seq

Les résultats de l'alignement en deux étapes réalisé en utilisant les outils NovoAlign [174] et BWA [122] donnent des résultats similaires à ceux obtenus par d'autres équipes sur le même type de données [142].

- Tab. 5.1 : Résultats de l'alignement avec l'outil NovoAlign [174], la première colonne correspond au nombre de lectures initial, la seconde au nombre de lectures effectivement alignées après filtrage sur la qualité (Q20) et la dernière colonne le pourcentage correspondant.
- Tab. 5.2 : Résultats pour la seconde étape d'alignement réalisée avec BWA [122] le nombre de lectures initial est indiqué en première colonne, il est différent de celui du Tab. 5.1 car cette deuxième étape d'alignement a pour but d'aligner les lectures qui n'ont pas pu être alignées avec Novoalign ou qui n'ont pas été bien alignées.
- Tab. 5.3 : Nombre total de lectures alignées à l'issue de la stratégie d'alignement en deux étapes après filtrage sur la qualité (Q20).

Plusieurs équipes [180], [182] suggèrent que parmi le nombre total de lectures alignées et ayant passées le filtrage qualité, 8% à 20% d'entre elles doivent contenir des délétions, c'est ce

que l'on obtient avec nos données (Tab. 5.4).

	Nombre de lectures initial	Nombre de lectures alignées	Pourcentage de lectures alignées
A203S1	49 077 961	12 072 112	24.6 %
A203S2	49 645 891	5 082 114	10.2 %
A203S3	45 782 934	12 768 338	27.9 %
A203S4	49 068 892	9 717 743	19.8 %

TABLE 5.1 – **Tableau présentant le nombre et le pourcentage de lectures alignées par NovoAlign** [174]

	Nombre de lectures initial	Nombre de lectures alignées	Pourcentage de lectures alignées
A203S1	28 150 823	586 454	2.1 %
A203S2	40 656 046	173 691	0.4 %
A203S3	24 526 780	384 699	1.6 %
A203S4	31 975 428	344 480	1.1 %

TABLE 5.2 – **Tableau présentant le nombre et le pourcentage de lectures alignées par BWA** [122]

	Nombre de lectures initial	Nombre de lectures alignées	Pourcentage de lectures alignées
A203S1	49 077 961	12 072 112	25.7 %
A203S2	49 645 891	5 082 114	10.5 %
A203S3	45 782 934	12 768 338	28.3 %
A203S4	49 068 892	9 717 743	20.5 %

TABLE 5.3 – **Tableau présentant le nombre et le pourcentage final de lectures alignées** [122]

	Pourcentage de lectures alignées contenant des délétions
A203S1	8.2 %
A203S2	12.3 %
A203S3	11.6 %
A203S4	9.2 %

TABLE 5.4 – **Tableau présentant le pourcentage de lectures alignées contenant des délétions**

Les pourcentages présentés dans le Tab. 5.3 peuvent paraître faibles. Il est important de préciser qu'il est très difficile d'aligner les lectures de CLIP-seq en raison de la taille des lectures qui sont souvent très courtes (inférieures à 50bp et parfois moins encore, voir Fig. 5.5). En effet, ces lectures sont généralement éliminées à l'étape de filtrage qualité en raison d'un trop mauvais score. De plus, un autre alignement de ces données avait été réalisé avec le logiciel

TopHat [183] sur le génome mm9 et le pourcentage moyen de lectures alignées était égal à 8.6%. Par conséquent, l'alignement que j'ai réalisé, permettant d'obtenir un pourcentage moyen de lectures alignées égal à 21.25%, a amélioré les résultats.

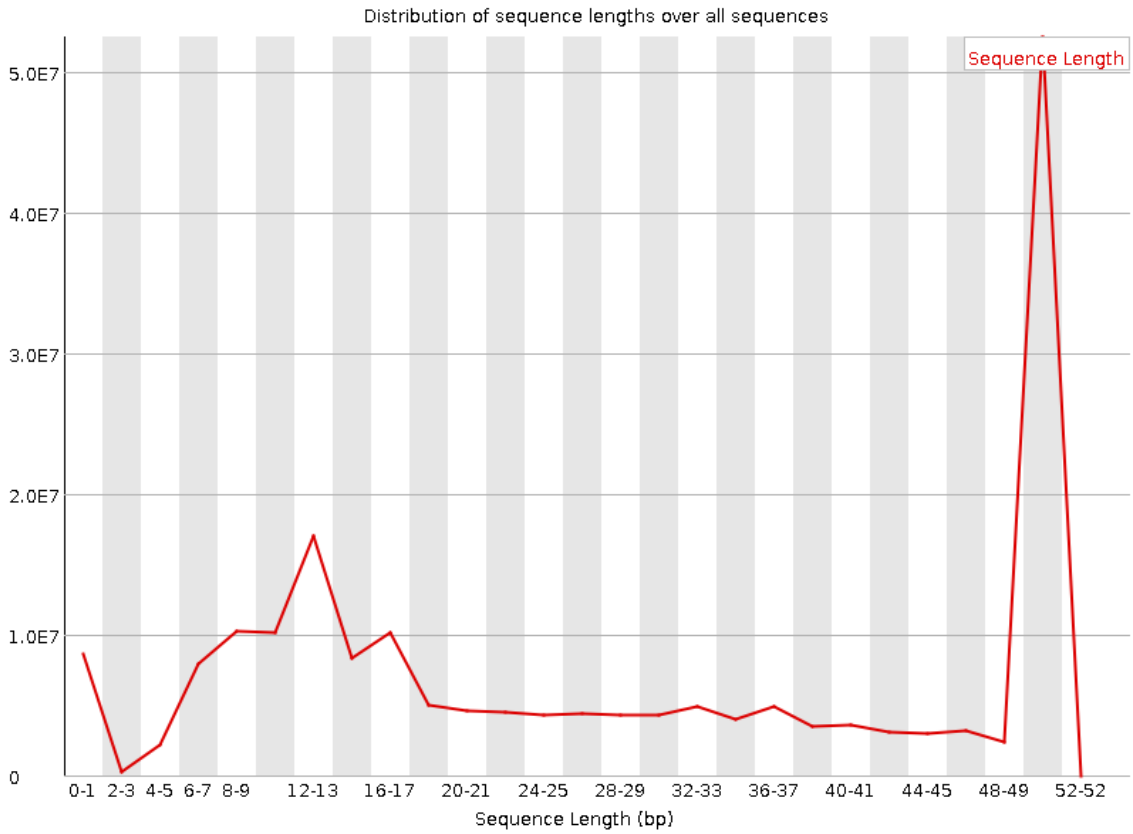


FIGURE 5.5 – **Histogramme présentant la taille moyenne des lectures de CLIP-seq.** La taille moyenne des lectures du CLIP-seq Spi1+ est de 50 paires de bases. Il existe toutefois une quantité non négligeable de lectures dont la taille est égale approximativement à 13 paires de bases, qu'il est impossible d'aligner.

5.2.2 Nombre de sites de liaison de SPI1 sur l'ARN

Le Tab. 5.5 présente le nombre de sites de liaison de SPI1 sur l'ARN pour chacun des réplicats techniques pour les deux réplicats biologiques (7B3, 722) ainsi que le nombre de gènes distincts présentant un pic de SPI1 sur l'ARN.

5.2.3 Caractérisation de la liaison de SPI1 à l'ARN

La première question qui se pose est :

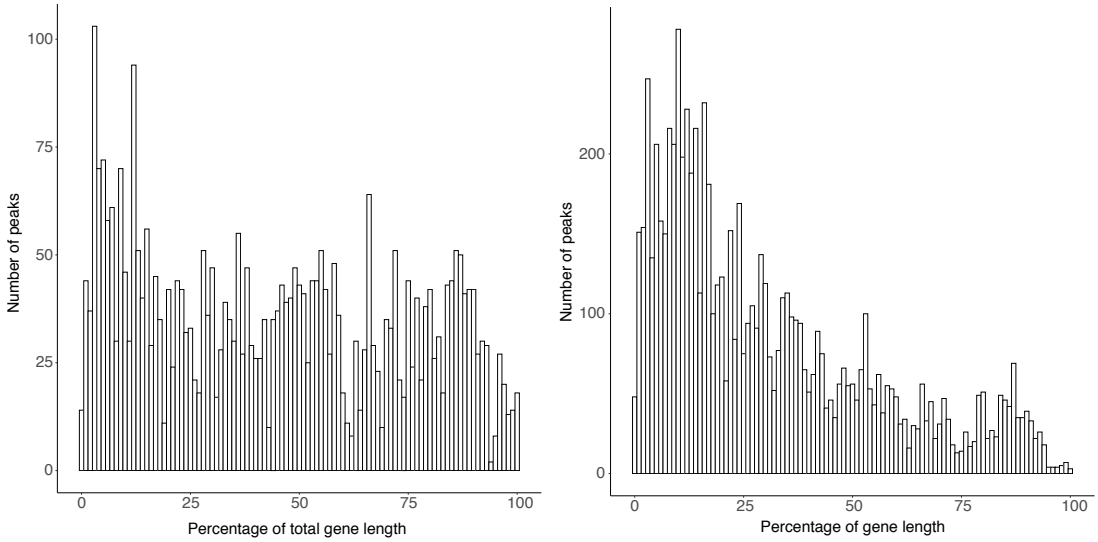
Où se lie le facteur de transcription SPI1 sur l'ARN ?

Pour répondre à cette question, je me suis intéressée à la distribution des pics de SPI1 sur toute la longueur des ARNs. Les positions des pics de SPI1 (a) pour le réplicat biologique 7B3

Réplicat technique	Réplicat biologique	Nombre de sites de liaison	Nombre de gènes distincts
A203S1	7B3	4652	1428
A203S2	7B3	4677	1153
A203S3	722	9803	1679
A203S4	722	8890	1537

TABLE 5.5 – **Nombre de sites de liaison de SPI1 sur l'ARN et de gènes différents fixés.**

et (b) pour le réplicat biologique 722 sont présentées en pourcentage de la longueur totale du gène sur lequel le pic a été annoté (Fig. 5.6).



(a) 7B3 (b) 722

FIGURE 5.6 – **Distribution des sites de liaison de SPI1 sur l'ARN des gènes fixés par SPI1 sur leur ARN.** La distribution est visualisée en nombre de pics et la longueur des gènes est rapportée en pourcentage pour le réplicat biologique 7B3 (a) et pour le réplicat biologique 722 (b).

Nous pouvons observer que la distribution des pics de SPI1 sur l'ARN se fait à proximité des TSSs sur la Fig. 5.6 (a) pour 7B3 et (b) pour 722. Cela est similaire à ce que l'on voit pour la liaison de SPI1 à l'ADN. Sur la Fig. 5.6 l'ensemble des gènes fixés par SPI1 sur leur ARN sont pris en compte, certains gènes peuvent être fixés par SPI1 à la fois sur leur ADN et sur leur ARN. Afin de ne voir que l'effet de la liaison à l'ARN on élimine tous les pics situés sur des gènes sur lesquels SPI1 fixe également l'ADN. Les résultats sont présentés dans la Fig. 5.7. Cette figure montre que la distribution des pics de SPI1 sur l'ARN, lorsque SPI1 ne fixe que la molécule d'ARN pour un gène donné ne présente pas de distribution particulière. En revanche, si l'on s'intéresse aux gènes dont SPI1 fixe à la fois l'ADN et l'ARN (Fig. 5.8), il existe un enrichissement dans la distribution au début de la molécule d'ARN, proche des TSSs, comme

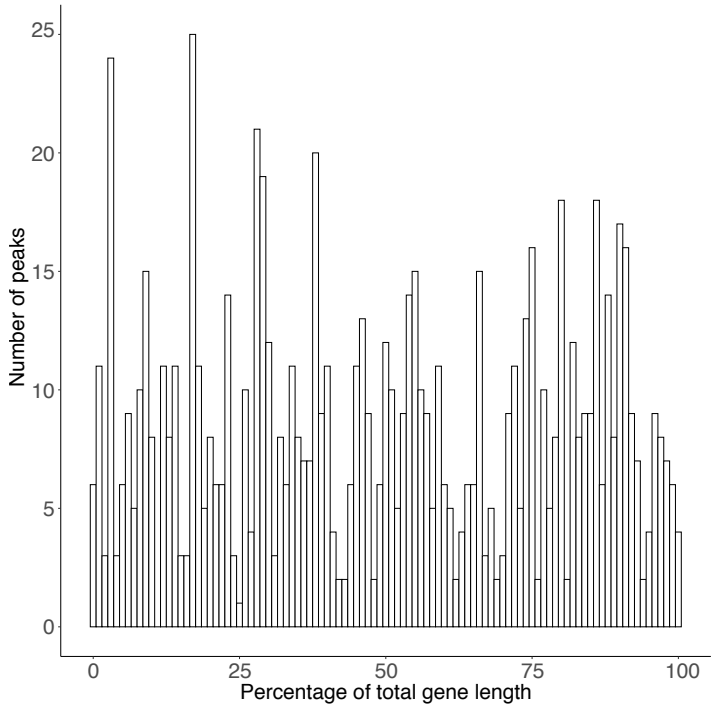


FIGURE 5.7 – **Distribution des sites de liaison de SPI1 sur l'ARN des gènes fixés par SPI1 uniquement sur leur ARN - réplicat biologique 7B3.** La distribution est visualisée en nombre de pics et la longueur des gènes est rapportée en pourcentage.

dans le cas de la liaison de SPI1 à l'ADN. Nous avons par la suite étudié la question suivante :

Est ce que la liaison de SPI1 sur l'ARN est influencée par sa liaison sur l'ADN ?

Pour répondre à cette question, j'ai réalisé un test de Fisher sur les gènes qui sont fixés à la fois sur leur ADN et sur leur ARN par SPI1. Le test de Fisher est un test exact utilisé pour l'analyse des tables de contingences, le calcul du odds-ratio permet de quantifier l'association de deux évènements A et B. Soit l'évènement A : "SPI1 se fixe sur l'ADN d'un gène donné", et l'évènement B : "SPI1 se fixe sur l'ARN d'un gène donné". Le test de Fisher va nous permettre de tester si il est très probable, ou non, que ces deux évènements se produisent simultanément, c'est à dire est-il très probable que SPI1 se fixe à la fois sur l'ADN et sur l'ARN d'un même gène ?

Le tableau de contingence utilisé pour le test est la Tab. 5.6. L'ensemble des gènes du RNA-seq exprimés (n=12871) est utilisé pour réaliser ce test. Les gènes définis comme exprimés sont les gènes dont la valeur de RPKM est strictement supérieure à 0. Parmi les gènes définis comme exprimés, 276 gènes sont fixés sur leur ARN par SPI1, au sein desquels **205** sont également fixés sur leur ADN par SPI1. Il y a donc seulement **71** gènes qui sont fixés uniquement sur leur ARN par SPI1. Parmi les 8874 gènes qui sont fixés par SPI1, 5611 gènes sont exprimés. Ainsi, il y a donc **5406** gènes exprimés qui sont fixés uniquement sur leur ADN par SPI1. Pour finir,

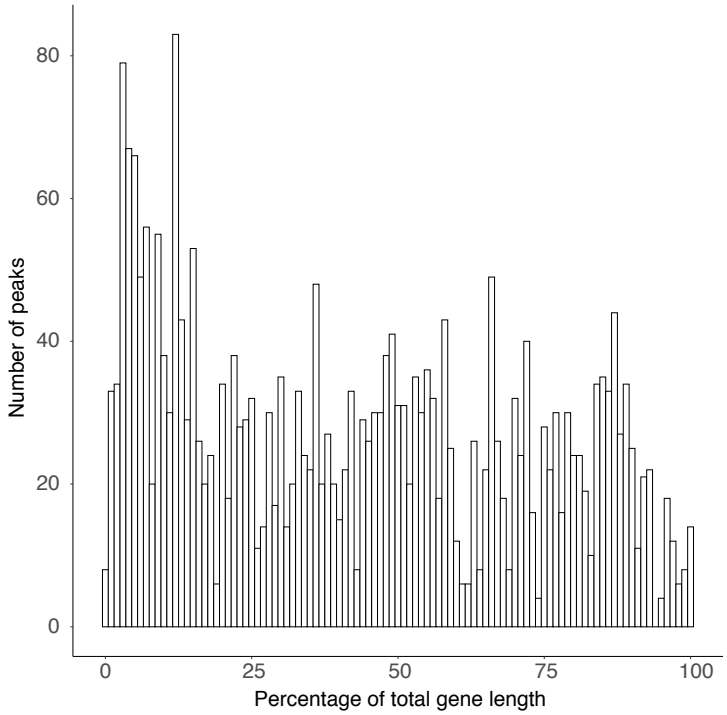


FIGURE 5.8 – **Distribution des sites de liaison de SPI1 sur l'ARN des gènes fixés par SPI1 à la fois sur leur ADN et sur leur ARN - réplicat biologique 7B3.** La distribution est visualisée en nombre de pics et la longueur des gènes est rapportée en pourcentage.

7189 parmi les 12871 gènes exprimés ne sont fixés ni sur leur ADN ni sur leur ARN par SPI1.

	SPI1 **se fixe** sur l'ADN du gène	SPI1 **ne se fixe pas** sur l'ADN du gène
SPI1 **se fixe** sur l'ARN du gène	205	71
SPI1 **ne se fixe pas** sur l'ARN du gène	5406	7189

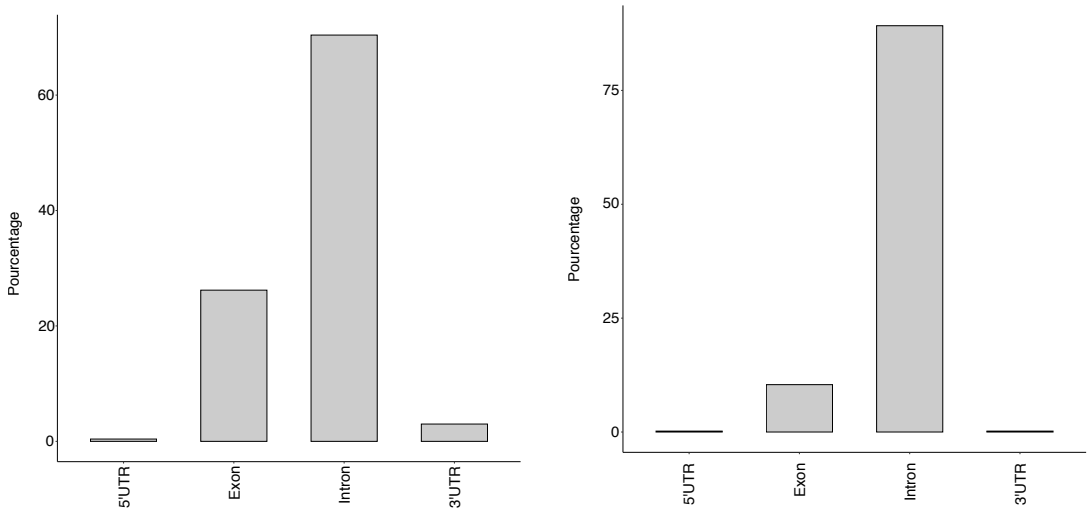
TABLE 5.6 – **Tableau de contingence pour le test de Fisher sur le réplicat biologique 7B3.** Les deux évènements considérés sont "Spi1 se fixe sur l'ADN d'un gène" et " Spi1 se fixe sur l'ARN du même gène".

Le test de Fisher sur ces données donne un odds-ratio égal à 3.8 avec une p-valeur inférieure à $2,2.10^{-16}$. Ainsi, il y a donc presque 4 fois plus de chance que le hasard que SPI1 se fixe sur l'ADN et sur l'ARN d'un gène donné. La liaison de SPI1 à l'ARN est donc influencée par sa liaison à l'ADN. Ce résultat ainsi que la distribution des pics de SPI1 sur l'ARN sur les gènes pour lesquels il fixe également l'ADN pose la question suivante : **Existe-t-il un crosstalk entre les molécules d'ADN et d'ARN d'un gène donné pour la liaison de SPI1 ?**

Quelles sont les régions génomiques sur lesquelles SPI1 se fixe principalement sur l'ARN ?

L'annotation des pics de SPI1 sur l'ARN permet de voir si une catégorie d'annotation est

enrichie par rapport aux autres catégories pour 7B3 (Fig. 5.9(a)) et pour 722 (Fig. 5.9(b)). L'ordonnée présente le pourcentage de pics dans chaque catégorie d'annotation par rapport au nombre de pics total. La catégorie "Intron" se distingue sur les deux figures : le pourcentage de pics de SPI1 sur l'ARN annoté en intron est égal à $70,4\%$ pour le réplicat biologique 7B3 et à $89,2\%$ pour le réplicat biologique 722. Les régions introniques sont uniquement présentes sur l'ARNm pré-mature, c'est-à-dire l'ARNm qui n'a pas encore subi les processus de maturation des ARNm que sont l'épissage, l'addition de la coiffe et de la queue polyA. Ainsi, SPI1 se fixe majoritairement sur des régions introniques sur l'ARN présentes sur les molécules d'ARNm pré-matures qui n'ont pas encore subi les trois étapes de maturation citées ci-dessus.



(a) 7B3 (b) 722

FIGURE 5.9 – **Régions génomiques fixées préférentiellement par SPI1 sur l'ARN** Barplot représentant le pourcentage de pics de SPI1 dans chaque catégorie d'annotation pour le réplicat biologique 7B3 (a) et pour le réplicat biologique 722 (b).

5.2.4 Rôle de la liaison de SPI1 à l'ARN

Pour étudier le rôle de la liaison de SPI1 à l'ARN, il me fallait m'extraire des influences de sa liaison à l'ADN. C'est pourquoi je me suis intéressée ici uniquement aux gènes fixés par SPI1 sur leur ARN. La liaison d'un facteur de transcription sur l'ARN d'un gène peut avoir plusieurs rôles tels que : l'épissage et la stabilité des ARNm.

Afin de voir si la liaison de SPI1 à l'ARN, et non à l'ADN, de certains gènes participait à l'activité transcriptionnelle, de ce dernier j'ai étudié l'indépendance des évènements : (i) "SPI1 se fixe à l'ARN d'un gène" ; (ii) "Ce gène est activé/réprimé par SPI1". Des tests de Fisher m'ont permis d'obtenir les valeurs de odds-ratio associant les régions génomiques de liaison de SPI1 au statut transcriptionnel des gènes fixés (Fig. 5.10). Une valeur égale à 1 pour un odds-ratio

indique qu'il n'y a pas plus de chance que le hasard que la conjugaison des deux évènements discutés se produise. La valeur de odds ratio supérieure à 20 pour l'enrichissement en liaison de SPI1 dans la région 5'UTR des gènes activés par SPI1 n'est pas du tout significative. En effet, lorsque SPI1 se fixe à l'ARN il ne se fixe que dans deux régions 5'UTR, c'est la raison pour laquelle le odds-ratio est si élevé. Les valeurs de odds-ratio présentées sont toutes autour de la valeur 1 en dehors de celle discutée ci-dessus (Fig. 5.10). Ainsi, la liaison de SPI1 à l'ARN et la régulation des gènes par SPI1 semblent être deux évènements indépendants.

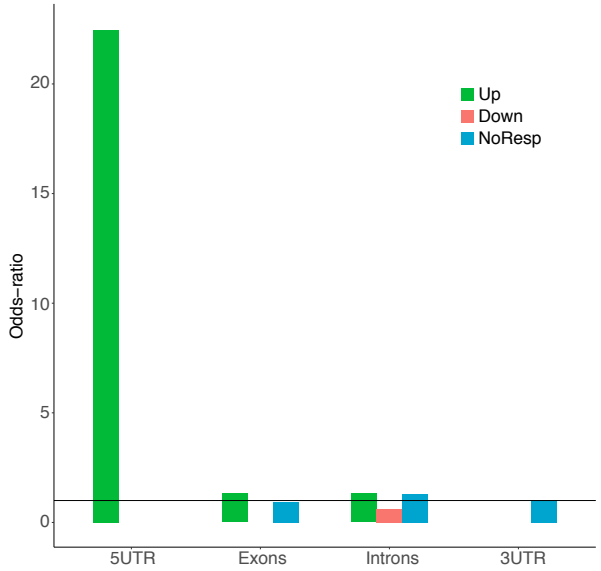


FIGURE 5.10 – **La liaison de SPI1 à l'ARN et la régulation des gènes par SPI1 semblent être deux évènements indépendants.** Enrichissement en molécules d'ARN fixées par SPI1 pour chaque catégorie de régulation transcriptionnelle dans les différentes régions génomiques des ARNs.

5.2.5 Recherche d'un motif de liaison à l'ARN

En utilisant l'outil AME de la suite MEME [132], nous avons regardé si le motif de liaison à l'ADN de SPI1 était retrouvé dans les potentiels sites de liaison de SPI1 à l'ARN. Deux motifs, celui de NR2C1 et de HSF1, ont été retrouvés dans les 97 séquences (Fig. 5.11). Toutefois leur e-values égales respectivement à $8.47e-1$ et $2.43e0$ ainsi que la valeur de leur TP en pourcentage (4.1% pour les deux) ne nous permettent pas d'interpréter ces résultats.

Les résultats obtenus par Zagros [179] sont disponibles en annexe D, un seul motif *de novo* a été déterminé (Fig. 5.12). J'ai utilisé l'outil Centrimo de la suite MEME [132] afin de voir si ce motif était bien enrichi dans les potentiels sites de liaison de SPI1 sur l'ARN. Le résultat obtenu (Fig. 5.13) est très intéressant puisque le motif est distribué de façon symétrique autour des potentiels sites de liaison de SPI1 dans les différentes régions. De plus, la p-valeur associée est égale à $2.4e-5$ ce qui est très significatif.

Logo	Database	ID	*p*-value	*E*-value	TP Thresh	TP (%)	FP (%)
	HOCOMOCOv11 core MOUSE mono meme format	NR2C1_MOUSE.H11MO.0.C	2.36e-3	8.47e-1	115.26	4 (4.1%)	0 (0.0%)
	HOCOMOCOv11 core MOUSE mono meme format	HSF1_MOUSE.H11MO.0.A	6.80e-3	2.43e0	50.69	4 (4.1%)	1 (0.1%)

FIGURE 5.11 – **Motifs déterminés par AME [132]**

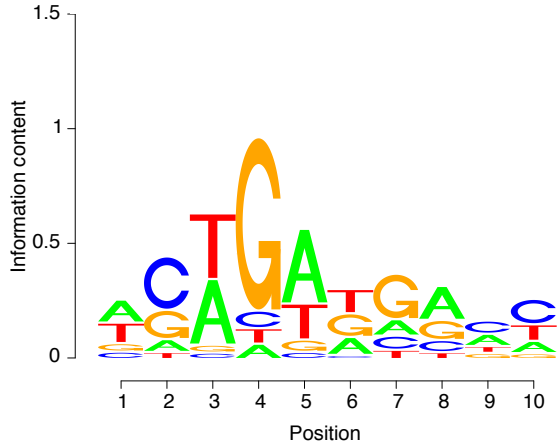


FIGURE 5.12 – **Logo du motif déterminé par Zagros [179] pour la liaison de SPI1 sur l'ARN**

L'ensemble de ces résultats montrent que :

- **La liaison de SPI1 à l'ARN se fait majoritairement dans des régions introniques.**
- **La liaison de SPI1 à l'ARN est influencée par la liaison de celui-ci à l'ADN.**
- **La liaison de SPI1 à l'ARN et la régulation des gènes par SPI1 semblent être deux évènements indépendants.**

Ces résultats seront discutés dans le chapitre suivant.

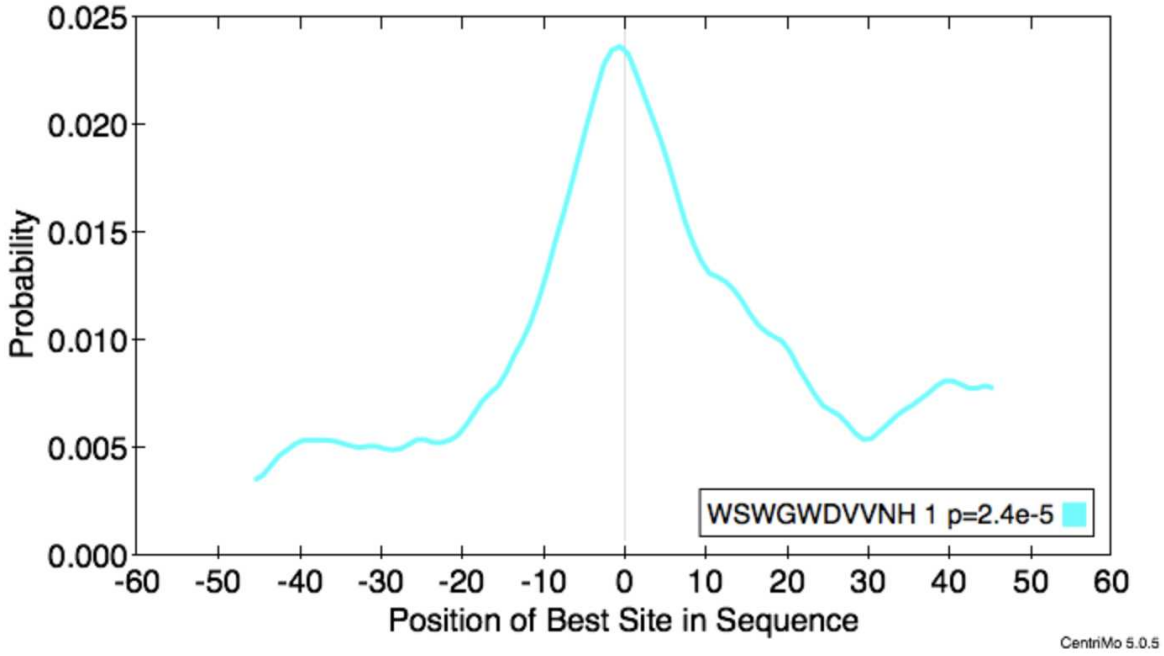


FIGURE 5.13 – **Distribution du motif déterminé par Zagros [179] autour de la position la plus probable pour SPI1 sur l'ARN**. La distribution est réalisée par l'outil CentriMo de la suite MEME [132]

Chapitre 6

Discussion et perspectives

Dans ce travail je me suis intéressée à la régulation de l'expression génique par le facteur de transcription SPI1 dans l'érythroleucémie, plus précisément j'ai mis en évidence que SPI1 réprime les gènes par fixation directe à la chromatine et par la coordination de deux mécanismes impliquant et contrôlés par HDAC1 et PRC2. Pour cela, j'ai analysé des données de séquençage à haut débit et les ai intégrées ensemble afin de permettre en plus de la caractérisation de la fixation à l'ADN et à l'ARN de SPI1, la compréhension des conséquences de la surexpression de celui-ci sur :

— la répression génique.
— les modifications des protéines histones et de la RNAPolII.
— les modifications de l'accessibilité à la chromatine.

L'une des contributions scientifiques de ce travail concerne le développement d'une méthode d'analyse de données de ChIP-seq. En effet, pour comprendre le rôle de SPI1 sur les modifications des protéines histones, de la structure chromatinienne et de la RNAPolII, j'ai dû analyser et comparer des jeux de données de ChIP-seq et d'ATAC-seq dans les conditions Spi1+ (Spi1 surexprimé) et Spi1- (Spi1 réprimé). Or, la comparaison des signaux de ChIP-seq entre différentes conditions a toujours été une question délicate. Il faut normaliser ces signaux et plusieurs biais expérimentaux sont à prendre à compte. Plusieurs méthodes ont été développées mais elles sont basées sur les pics communs entre les conditions ou utilisent l'input. Aucune ne permet d'intégrer des informations supplémentaires pour une détermination plus fine et exacte des paramètres de normalisation. Ainsi, une partie de mon travail de thèse a été consacré au développement d'un package R pour la normalisation de données de ChIP-seq comprenant plusieurs échantillons ou plusieurs conditions dans le but de les rendre comparables.

6.1 Une nouvelle méthode de normalisation des données de ChIP-seq inter-conditions basée sur l'invariance du signal pour des gènes spécifiques

L'étape de normalisation du signal de ChIP-seq est une étape obligatoire pour permettre la comparaison des intensités de liaison des facteurs de transcription ou des marques d'histone dans les différentes conditions biologiques. CHIPIN dispose de deux méthodes de normalisation : régression linéaire avec interception non nulle et normalisation quantile. Les deux méthodes développées dans CHIPIN sont plus efficaces que les autres approches de normalisations publiées [140], [139] que nous avons testées sur deux jeux de données (Fig. 3.6 Tab. 3.1, Fig. 3.7 Fig. 3.9 Tab. 3.2, Fig. 3.8 Fig. 3.10 Tab. 3.3). CHIPIN prend en entrée des fichiers .bigWig générés par la majorité des logiciels de recherche de site de fixation. Cependant, dans notre étude de validation nous avons utilisé préférentiellement le logiciel HMCan [124]. En effet, HMCan est un logiciel de recherche de site de fixation qui corrige pour le biais du contenu en GC, le biais en nombre de copies et retire le signal du bruit de fond (si un input est fourni). Son utilisation est fortement recommandée lorsque des échantillons normaux sont comparés à des échantillons de tumeurs présentant des différences dans les profils de nombre de copies.

Bien que nous ayons développé CHIPIN pour intégrer des données d'expression pour inférer les paramètres de normalisation, l'outil peut techniquement fonctionner sans les données d'expression. Lorsque l'utilisateur ne fournit pas de données d'expression, la procédure de normalisation est basée sur le signal calculé sur toute la longueur des gènes et dans les régions adjacentes. Néanmoins, cette façon de normaliser le profil n'est pas recommandée car elle peut entraîner une sur-correction du signal.

La méthode CHIPIN a été initialement développée pour des données de ChIP-seq, cependant elle est tout à fait générale et peut donc être appliquée sans aucune limitation à d'autres profils de densité tels que ceux construits pour des données d'ATAC-seq ou DNAse. Récemment, une étude [150] a démontré que le choix de la méthode de normalisation pour la comparaison de l'accessibilité à la chromatine dans différentes régions, à partir d'expériences d'ATAC-seq, était d'une importance cruciale pour éviter les fausses interprétations biologiques. Comme il est connu que l'état relâché de la chromatine au niveau des régions promotrices est en corrélation positive avec l'expression des gènes, l'utilisation de gènes dont l'expression est constante pour les différents échantillons/conditions, comme mis en oeuvre dans CHIPIN, représente une solution appropriée pour normaliser les données d'ATAC-seq.

CHIPIN ne doit pas être utilisé lorsque la protéine ciblée est surexprimée ou réprimée dans l'une des conditions. Il convient plutôt d'utiliser un protocole de spike-in suivi par une normalisation avec une méthode appropriée qui peut tenir compte de l'information spike-in, par exemple, HMCan-diff [184] pour déterminer les régions différentiellement enrichies.

La fonction *plot_expression* permet à l'utilisateur de vérifier que l'anticorps utilisé pour l'expérience est spécifique de la protéine ciblée Fig. 3.5. Pour cela, des connaissances biologiques sur les effets de la liaison de la protéine d'intérêt sur l'expression des gènes peuvent être utilisées.

Deux ensembles de données ont été utilisés pour comparer CHIPIN à d'autres méthodes de normalisation disponibles. Dans le second ensemble de données, deux réplicats techniques sont disponibles. Pour la validation de notre méthode nous avons choisi d'utiliser le réplicat 1 dans la condition 1 et le réplicat 2 dans la condition 2 afin d'augmenter l'effet de lot ("batch effect"). Dans ce cas CHIPIN a surpassé les autres méthodes (Fig. 3.7 Fig. 3.9, Tab. 3.2). Cependant, nous avons également comparé CHIPIN aux trois autres méthodes en utilisant le même réplicat pour les deux conditions (Fig. 6.1, Fig. 3.10, Tab. 3.3). Il est important de noter que dans ce cas également, CHIPIN surpasse les autres méthodes de normalisation. Toutefois, nous avons pu constater que la normalisation au même nombre de reads a également donné des résultats presque parfaits (Fig. 3.10, Tab. 3.3). Cette analyse démontre l'importance de la technique de normalisation utilisée lors de la comparaison de différents échantillons ou pour des expériences effectuées à des temps différents, tandis que la normalisation au même nombre de reads peut être appliquée à des expériences effectuées au même moment.

CHIPIN est un package R facile d'utilisation, gratuitement disponible sur GitHub à l'adresse `https://github.com/BoevaLab/CHIPIN/`. Des exemples sont présentés dans la vignette, celle-ci est disponible sur GitHub, des données tests sont également fournies.

6.2 Mise en évidence d'un mécanisme de répression génique en coopération avec HDAC1 et PRC2 par fixation de SPI1 à la chromatine

Nous avons mis en évidence un mécanisme de répression exercé par SPI1 qui implique HDAC1 et PRC2. En effet, j'ai montré que SPI1 réprime des gènes du réseau érythroïde principal en se fixant dans des régions enhancers que sont les corps de gène et les régions intergéniques. La coopération avec la protéine HDAC1 est illustrée par une dé-acétylation locale de l'histone H3 aux sites de fixation de SPI1 et aux TSSs associés suggèrant un mécanisme de régulation contrôlé entre ces deux régions. De plus, nous avons montré que les gènes associés à une dé-acétylation au niveau de ces deux régions présentent une diminution de l'accessibilité à la chromatine et de la charge de RNAPolII au niveau des TSS associés. Notre travail montre également que la répression de l'expression de *Spi1* n'entraîne pas l'ouverture de la chromatine aux sites de fixation de SPI1 sur les enhancers.

Ces résultats suggèrent un mécanisme impliquant la fixation de SPI1 dans les enhancers qui agit sur les régions promotrices résultant en la diminution de l'accessibilité à la chromatine pour

la machinerie de transcription, dont RNAPolII, sans moduler l'accessibilité des enhancers où il se fixe.

Nous avons montré qu'une partie des gènes réprimés par SPI1 sont des gènes régulés par le complexe érythroïde GATA1/KLF1/TAL1, ce qui est en accord avec la fonction de SPI1 dans la répression de la différenciation érythroïde (Fig. 4.8B) [151]. Ce complexe peut aussi inclure LDB1, ETO2 et MTGR1 et aura un rôle activateur ou répresseur de l'expression génique selon la présence des protéines dans le complexe. Ce complexe peut agir à distance, par looping, comme par exemple sur le gène de la β-globin [185]. Soler et al [151] ont démontré que la composition de ce complexe était variable au cours de la différenciation. En particulier le nombre de sites de fixation de ETO2 et MTGR1 (des régulateurs négatifs de la différenciation) diminue lorsque la différenciation est induite. Or, ETO2 est capable de bloquer le mécanisme de looping, dont ce complexe est responsable, avant l'engagement terminal de la différenciation érythroïde [186]. SPI1 augmente l'expression de ETO2, normalement diminuée au cours de l'engagement terminal de la différenciation érythroïde (post-CFUE-proE).

Plusieurs hypothèses concernant le scénario de la répression orchestrée par SPI1 dans les cellules pré-leucémiques sont alors possibles :

— Mécanisme additif indirect par augmentation de l'expression de ETO2, un régulateur négatif de la différenciation.

— ETO2 interagit avec des histones dé-acétylases [186, 187], SPI1 pourrait stabiliser l'interaction entre ETO2 et un complexe dé-acétylase. Comme ETO2 fait partie d'un complexe contenant GATA1, cela expliquerait la co-localisation des motifs de SPI1 et de GATA1 pour un certain nombre de gènes où la répression transcriptionnelle de SPI1 s'illustre par une dé-acétylation.

D'autre part, la protéine LDB1 a la capacité d'induire des interactions entre des régions éloignées, telles que des enhancers et des promoteurs [188]. La présence de LDB1 sur les régions de régulation des gènes érythroïdes est augmentée dans les cellules érythroïdes qui se différencient. Ainsi, une hypothèse est que SPI1 bloque l'activité de looping du complexe LDB1 en coopération ou non avec ETO2. Ce blocage entrainerait une diminution de l'ouverture de la chromatine et de la charge de la RNAPolII au niveau des promoteurs. Il serait également associé à une dé-acétylation des sites de fixation de SPI1, et probablement des sites occupés par le complexe LDB1 ainsi que des TSSs associés. Le rôle de l'acétylation dans la formation du looping n'est pas connu. De plus, ce travail ne nous permet pas de définir si c'est d'abord H3K27ac qui est diminué induisant ainsi la diminution de l'accessibilité de la chromatine et de charge de RNAPolII aux TSSs ou l'inverse.

Une question importante de ce travail concerne le rôle de HDAC1 dans la répression génique des gènes cibles de SPI1. Nous savons que HDAC1 et SPI1 interagissent à la chromatine et que la présence de SPI1 est associée à une dé-acétylation des enhancers et des TSSs associés. Nous avons aussi montré que HDAC1 est nécessaire à la répression des gènes par SPI1.

Une hypothèse est que SPI1 recrute HDAC1 aux enhancers pour dé-acétyler les histones et que HDAC1 agirait également sur le TSS du gène correspondants dans le cas où les deux régions interagissent. Ainsi, il faudrait définir les sites où se trouvent HDAC1 dans nos cellules pré-leucémiques en présence et en absence de SPI1 afin de voir si SPI1 et HDAC1 co-localisent à la chromatine. Sebastian Gregoricchio a essayé plusieurs protocoles et modèles pour répondre à cette question. Pour le moment, aucun n'a apporté de réponse en raison de la difficulté d'immunoprécipiter HDAC1 de façon spécifique dans des complexes où il n'est pas chimiquement fixé à la chromatine.

De plus, dans les régions où la présence de SPI1 diminue fortement l'acétylation des enhancers et des TSSs associés, des motifs de fixation et des sites de fixation du facteur de transcription GATA1 sont trouvés à proximité des motifs de SPI1. Nous avons prouvé que les sites de fixation des deux facteurs de transcription sont distincts, ainsi ils sont tous deux fixés à la chromatine contrairement à ce qui est publié [92, 166]. Déterminer les sites où HDAC1 se trouve nous permettrait de voir si SPI1, HDAC1 et GATA1 co-localisent à la chromatine dans les gènes où SPI1 et GATA1 sont tous deux fixés à la chromatine. Ainsi, nous pourrions étudier si HDAC1, en coopération ou non avec ETO2, conduit à la dé-acétylation de GATA1, donc à son inactivation transcriptionnelle et potentiellement au blocage de l'activité de looping du complexe LDB1.

Nous avons également montré que la répression des gènes par SPI1 est associée à une augmentation de H3K27me3 au niveau des promoteurs de ces gènes. Cette augmentation est observée plus fortement aux promoteurs des gènes réprimés fixés par SPI1 en comparaison avec les gènes réprimés par SPI1 mais non fixés par ce dernier. On sait maintenant que PRC2 est responsable du dépôt de la marque H3K27me3 à la chromatine sur des gènes faiblement transcrits. L'idée est que la marque H3K27me3 vient maintenir la répression transcriptionnelle.

PRC2 et SPI1 interagissent physiquement dans les cellules pré-leucémiques érythroïdes [66]. Notre résultat suggère que la fixation de SPI1 dans ses gènes cibles exerce un effet amplificateur de la répression en augmentant l'activité de PRC2 aux promoteurs associés. En effet, le groupe de Fernando Rodrigues-Lima à l'Institut Jacques Monod a mesuré *in vitro* que SPI1 augmente l'activité de PRC2 sur la lysine 27 de l'histone H3 (données non montrées).

Le traitement combiné des cellules pré-leucémiques par des inhibiteurs de EZH1/2, l'enzyme du complexe PRC2, (UNC1999) et de HDAC1/3 (Entinostat) a induit une mortalité de plus de 36% des cellules après 48H de traitement versus 5% pour les deux inhibiteurs séparés. L'inhibition combinée de EZH2 et de HDAC1 est donc synergique et en accord avec la coopération des deux mécanismes régulés par EZH2 et HDAC1. Ce traitement combiné a également induit une augmentation de l'expression de certains gènes réprimés par SPI1 dont :

— *Ptp4a3* et *Alas2* qui sont des gènes de la différenciation érythroïde, cibles du complexe formé par LDB1, GATA1, TAL1.

— *Bcl2l11* un gène pro-apoptotique dont la répression induit une résistance à l'apoptose

[66].

Ces deux inhibiteurs n'ont néanmoins pas ré-induit la différenciation érythroïde car les cellules meurent. Ils pourrait être utilisés *in vivo* pour tester leur efficacité dans l'élimination de l'expansion des cellules pré-leucémiques.

L'ensemble de ces résultats apporte de nouvelles perspectives sur le rôle du facteur de transcription hématopoïétique SPI1 pour les mécanismes de répression des gènes, dont très peu sont décrits [107, 108]. Ces travaux ont été réalisés dans des cellules pré-leucémiques indépendantes de l'Epo pour leur prolifération et leur survie et dont la différenciation est bloquée par SPI1. Aucun mécanisme global de répression génique par SPI1 n'avait été mis en évidence dans ces cellules auparavant.

Pendant ce travail de thèse, il a été montré que SPI1 restreint la défense des neutrophiles en inhibant l'accessibilité des enhancers par le recrutement de HDAC1 [108]. Ces modifications épigénétiques ont empêché le facteur de transcription immunostimulateur AP-1 JUNB de pénétrer dans la chromatine et d'activer ses cibles. Il semblerait que l'interaction de HDAC1 et de SPI1 soit impliquée dans la répression des gènes dans plusieurs lignages hématopoïétiques.

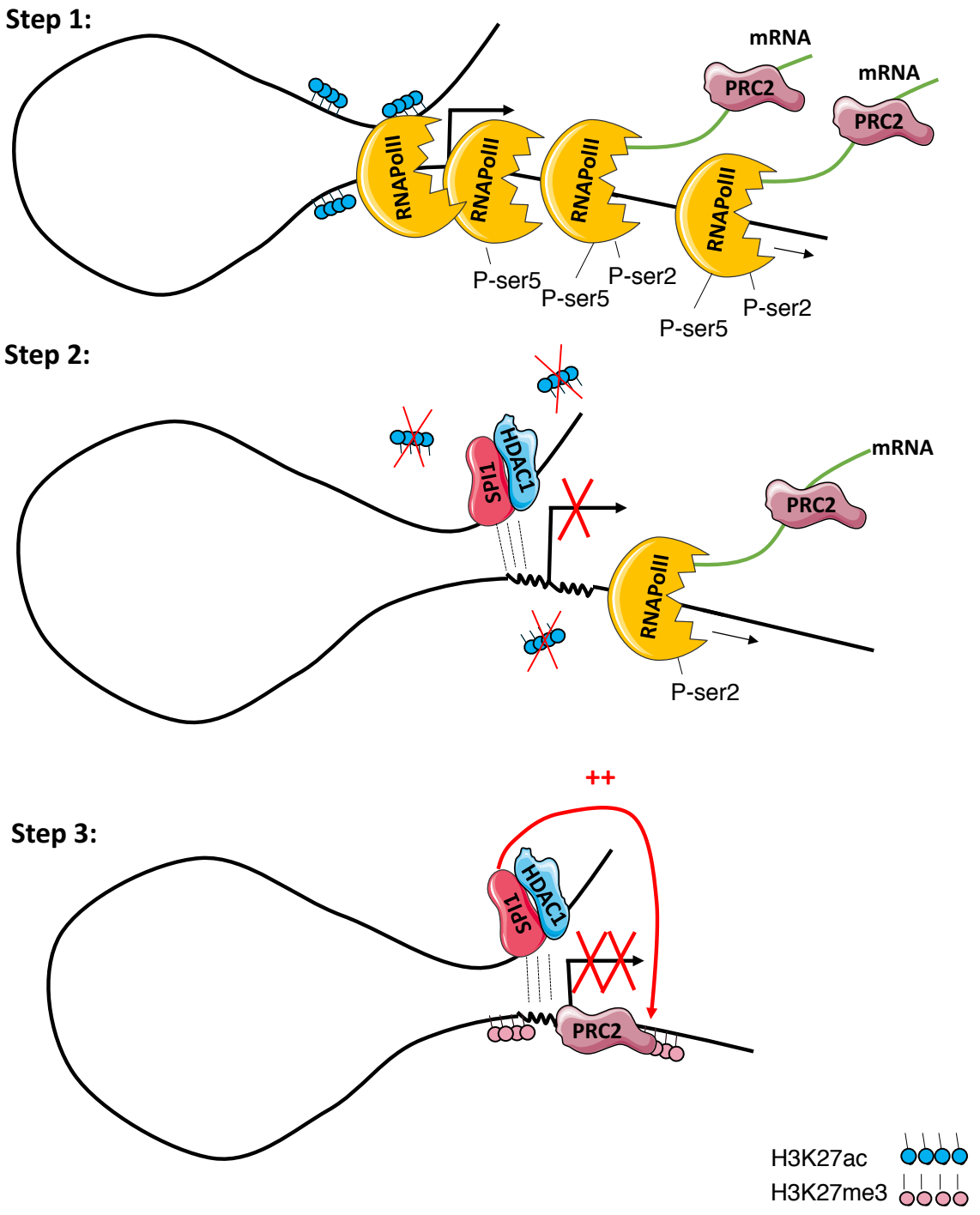


FIGURE 6.1 – **Modèle de travail sur la base des résultats obtenus. Étape 1 :** Gène activé en absence de SPI1, expression génique par formation d'une boucle enhancer-promoteur, PRC2 fixé à l'ARN nascent. **Étape 2 :** SPI1 se fixe aux enhancers actifs (enhancer en intergénique représenté ici), répression génique initiée par la coopération de SPI1 et de HDAC1 : diminution de l'acétylation au niveau du site de SPI1 et au TSS, condensation de la chromatine au TSS. Perte de la charge de RNAPolII au promoteur. **Étape 3 :** La répression génique et SPI1 induisent la venue de PRC2 au niveau du TSS et ainsi le dépôt de la marque H3K27me3 au TSS.

6.3 Caractérisation de la fixation de SPI1 à l'ARN

Nous avons démontré que le facteur de transcription SPI1 présente des modalités de fixation à l'ARN différentes suivant la nature de la molécule d'ARN fixée. En effet, si SPI1 se fixe uniquement sur l'ARN d'un gène donné, sa fixation ne présente pas de distribution particulière (Fig. 5.7). En revanche, lorsque SPI1 se fixe à l'ARN d'un gène dont il fixe aussi la molécule d'ADN alors la fixation de SPI1 se fera préférentiellement en 5' de la molécule d'ARN (Fig. 5.8). De plus, la fixation de SPI1 sur l'ARN est influencée par sa fixation à l'ADN, en effet, le test de Fisher permettant d'évaluer l'indépendance entre les évènements "SPI1 se fixe sur la molécule d'ADN d'un gène" et "SPI1 se fixe sur l'ARN du même gène" donne une valeur de odds-ratio égale à 4 associée à une $pval < 2.2e-16$. L'annotation des pics de SPI1 sur l'ARN nous a permis de montrer que SPI1 se fixe majoritairement dans des régions introniques (Fig. 5.9). L'analyse transcriptionnelle des gènes pour lesquels SPI1 se fixe uniquement sur l'ARN nous a permis de montrer qu'il n'y avait pas de lien entre la fixation de SPI1 à l'ARN et le statut transcriptionnel des gènes. En effet, aucun enrichissement en gènes réprimés ou activés par SPI1 n'a été observé dans les gènes fixés uniquement sur leur molécule d'ARN par SPI1 (Fig. 5.10). De plus, nous avons extrait un possible motif de fixation de SPI1 à l'ARN (Fig. 5.12) dont la distribution autour des potentiels sites de fixation de SPI1 est symétrique (Fig. 5.13) et associée à une p-valeur égale à $2.4e-5$. Tous ces résultats ont été obtenus en essayant au maximum de s'extraire des biais introduits par l'expérience de CLIP-seq. Toutefois des méthodes développées plus récemment telles que le iCLIP [143] ou le eCLIP [144] et sur lesquelles le recul est aujourd'hui suffisant pourraient permettre de valider les résultats que nous avons obtenus.

Il a été montré que SPI1 contrôle l'épissage de l'ARNm prémature [111], sa fonction s'exerce de deux façons différentes :

— Par sa capacité à fixer l'ADN et à transactiver.

— Sans fixation à l'ADN, le mécanisme est encore inconnu mais pourrait reposer sur une séquestration des facteurs d'épissage par SPI1 tel que TLS [110].

De plus, SPI1 interagit avec le facteur d'épissage potentiel TLS, il existe des complexes *in vivo* contenant SPI1 et TLS. Ainsi l'une des hypothèses sur le rôle de la fixation de SPI1 à l'ARN pourrait concerner sa capacité à interagir avec des facteurs d'épissage entrainant ainsi leur séquestration et un épissage différent de celui attendu en absence de SPI1.

La fixation des facteurs de transcription à l'ARN n'a été que très peu caractérisée. Lorsque j'ai commencé ma thèse aucune étude utilisant des techniques de séquençage à haut débit pour caractériser la fixation d'un facteur de transcription à l'ARN n'existait. Une étude a permis de montrer que le suppresseur de tumeur P53, lorsqu'il est surexprimé induit la suppression de la traduction de l'ARNm du gène *mdmx* par fixation à celui-ci [189]. La protéine MDMX est un régulateur négatif de la trans-activité de p53. Une étude plus récente montre que le facteur de

transcription SOX2, essentiel pour l'auto-renouvellement des cellules souches embryonnaires indifférenciées et des cellules souches neurales, se fixe à des molécules d'ARN [190]. Plus particulièrement cette fixation se fait avec une haute affinité et une faible spécificité (*in vitro* et *in vivo*) et présente une préférence de fixation pour les ARNs doubles brins. Dans notre étude, les ARNs doubles brins ne sont pas inclus. En effet, Liu et al [191, 192] suggèrent que le crosslinking utilisant les UVs, comme dans la méthode de HITS-CLIP que nous avons utilisée, ne permet pas de récupérer les molécules d'ARNs doubles brins. Beaucoup de mécanismes de régulation mettent en jeu ces structures formées par des miRNAs se fixant sur des ARNs simples brins ou par le repliement des ARNs simples brins pour former des structures double brin. Ces mécanismes n'ont pas pu être mis en évidence dans notre étude puisque les sites de fixation de SPI1 sur ces ARNs double brins n'ont pas été capturés lors de l'étape de réticulation du protocole de HITS-CLIP utilisant les UVs [191, 192].
Il a été montré que SPI1 fixe des molécules d'ARN par son domaine de fixation à l'ADN [109]. Le motif de SPI1 sur l'ADN n'a pas été retrouvé lorsque nous avons réalisé la recherche de motifs sur l'ARN en utilisant AME [132]. Or, les structures des ARNs simples brins sont très différentes de la structure de l'ADN double brin, ainsi si notre étude n'a pas permis de récupérer les ARNs double brin fixés par SPI1, on ne peut pas exclure l'hypothèse selon laquelle la fixation des ARNs doubles brins par SPI1 se fait avec son motif de fixation à l'ADN. Toutefois, pour le facteur SOX2 il a été montré que cette fixation à l'ARN se faisait avec le domaine de fixation à l'ADN de SOX2 mais en utilisant des résidus différents [190].

6.4 Conclusion globale

En conclusion, nous avons mis en évidence un mécanisme global de répression des gènes par SPI1 dans les cellules pré-leucémiques. Le mode d'action de SPI1 dans l'activation transcriptionnelle est bien caractérisé, très peu de mécanismes de répression ont été mis en évidence. Ainsi, ces résultats ouvrent la voie pour une meilleure compréhension du rôle de SPI1 dans l'érythroleucémie murine et à plus large échelle pour les nombreuses leucémies dans lesquelles SPI1 est impliqué.

Afin de mettre en évidence ce mécanisme de répression, j'ai analysé et intégré des données de ChIP-seq pour différentes marques d'histones, d'ATAC-seq et de RNA-seq dans différentes conditions. Pour cela, j'ai développé une nouvelle méthode pour l'analyse des données de ChIP-seq entre ces différentes conditions. Cette méthode est applicable à plus de deux conditions et peut également s'appliquer pour différents échantillons de tumeurs de patients. Elle donne de très bons résultats par rapport aux méthodes existantes et permet à l'utilisateur de visualiser l'efficacité de la normalisation et de récupérer les fichiers de densité normalisés qui sont directement utilisables et visualisables dans IGV. Aucun traitement post-normalisation n'est nécessaire.

Nous avons également caractérisé la fixation du facteur de transcription hématopoïétique SPI1 sur l'ARN, cela n'avait jamais été fait auparavant. La méthode expérimentale que nous avons utilisée ne nous a probablement pas permis de voir la fixation de SPI1 sur les ARNs double brins, et l'impact de la fixation de SPI1 sur l'ARN sur les mécanismes d'épissage n'a pas été étudié. Toutefois nous avons pu mettre en évidence une fixation localisée au début des ARNs simples brins et dans les régions introniques, ce qui ouvre la voie vers de nouvelles hypothèses quant au rôle de SPI1 par sa fixation à l'ARN. Il faudrait caractériser sa fixation sur les ARNs double brins afin de voir si cela peut avoir un lien avec la stabilité des ARNs.

Annexe A

Matériels et méthodes des expérimentations biologiques

A.1 Cell culture and chemicals

Cell suspensions from enlarged spleen of *Spi1* transgenic mice (TgSpi1) that developed an erythroleukemia (HS1) have been cultured in alpha Eagle's minimal essential medium (Gibco) supplemented with 5% Foetal Bovine Serum gold (FBS-gold, PAA, A15-151), L-Glutamine (2mM, Gibco), antibiotics penicillin/streptomycin (Gibco) and erythropoietin (EPO) at 1U/mL at 37 ° C in the presence of 5% CO_2. Cells are diluted every two days at 100 000 cells/mL [65]. The erythroleukemic cells were engineered to express anti-Spi1 shRNAs (shSpi1-A2B) in the presence of doxycycline (dox, 100ng/mL) as previously described [67].

HDACs inhibition was performed by incubation with 3μM entinostat (MS-275, Santa Cruz) for 6h and 0.5 μM or 2μM TMP195 (M6176, AbMole) for 12h after 48h dox treatment. Equal amount of DMSO has been used as control.

A.2 Co-immunoprecipitation and immunobloting

For immunoprecipitations 4010^6 cells, cells were lysed in 200µL IP-buffer (20mM HEPES, 2mM $MgCl_2$, 0.5% NP-40, 100mM NaCl, 0.1% Triton) and incubated on ice for 10 minutes. Then, 6.3μM 5M NaCl (final concentration of 420 mM) were added, incubated on ice for 5 minutes and centrifuged for 10 minutes at 20 000g. Supernatants were incubated 2h under rotation with the right amount of antibody (Fig. A.1) and then for 1h with 70µL of magnetic beads (Protein G coated Dynabeads, Invitrogen) for immunoprecipitations. Beads then have been washed 5 times with IP-buffer and then boiled for 10min at 95°C in 20µL Laemmli buffer 2X supplied with DTT 100.

For whole-cell protein extraction, cells are centrifuged at 250g for 7min at 4°C, washed

twice in DPBS-orthovanadate (1mM) and then lysed in Laemmli 2X buffer supplied with DTT 100 at concentration of 50 000 cells/µL, sonicated by Bioruptor-Standard sonicator (Diagenode) (intensity High, 3 cycles of 30s ON and 90s OFF) and then boiled 10min at 95°C.

Whole-cell extracts or immunoprecipitated/pulled-down proteins were resolved by SDS-PAGE and transferred to a nitrocellulose 0.22µm membrane (Biorad). The membrane was incubated with primary antibody diluted in antibody solution (DPBS 1X, Tween20 0.1%, milk powder 5%) as described in Fig. A.1 and with secondary antibody (see Fig. A.1) solution for 1h at room temperature. Proteins were detected with LAS4000 digital imager (GE Healthcare Life Sciences).

Target	Reference	Lot	Tech	Usage
Anti-Mouse-HRP	Jackson, 115-035-003	144935	WB	1:40 000, 1h RT
Anti-Rabbit-HRP	Jackson, 111-035-033	116154	WB	1:40 000, 1h RT
GATA1 [EPR17362]	Abcam, ab181544	GR208183-2	ChIP	10µg per 30µg chromatin, Protein G agarose beads
H3K27ac	Abcam, ab4729	GR225006-16	ChIP WB	2µg per 15µg chromatin, Protein G magnetic beads 1:5000, O/N +4°C
H3K27me3	Cell Signaling, cs9733	8	ChIP WB	10µg per 30µg chromatin, Protein G magnetic beads 1:1000, O/N +4°C
H3K36me3	Abcam, ab9050	ND	ChIP	5µg per 15µg chromatin, Protein G magnetic beads
H3K4me1	Abcam, ab8895	GR276935-1	ChIP	2µg per 15µg chromatin, Protein G magnetic beads
H3K4me3	ActiveMotif, 39159	01609004	ChIP	2µg per 15µg chromatin, Protein A magnetic beads
H4	Abcam, ab10158	GR3231419-1	WB	1:5000, 1h30 RT
HDAC1	Abcam, ab7028	GR3230011-7	ChIP IP WB	100µg per 30µg chromatin, Protein G magnetic beads 12µg per 40×10^6 cells
HSP90b	Enzo LifeScience, ADI-SPA-	ND	WB	1:1000, O/N +4°C
IgG (rabbit)	Abcam, ab171870	GR3228514-2	ChIP IP	same quantities and methods used for the target IP
SIN3A	Abcam, ab3479	GR3215134-7	IP WB	10µg per 40×10^6 cells 1:5000, O/N +4°C
Protein A-HRP	Pierce, 32400	OJ194200	WB	1:10 000, 1h RT
RNA-pol II	Santa Cruz, sc-899 X	ND	ChIP	5µg per 30µg chromatin, Protein A agarose beads
SPI1	Santa Cruz, sc-352 X	D1113 / D2910	ChIP	10µg per 30µg chromatin, Protein A agarose beads
SPI1*	Homemade (rabbit serum)	/	WB	1:2000, O/N +4°C

* Moreau-Gachelin, F. et al. Spi-1/PU.1 transgenic mice develop multistep erythroleukemias. Mol. Cell. Biol. 16, 2453–2463 (1996).

FIGURE A.1 – **Liste des anticorps utilisés et mode d'emploi**

A.3 ChIP assay and quantification by real time quantitative PCR (RT-qPCR)

For SPI1 and RNA-pol II ChIP, a total of 8.10^6 cells were used per ChIP assay using the ChIP assay kit (Millipore) following the manufacturer's protocol as previously described [37]. For the others ChIP assays, 3 to 10.10^6 cells per condition are cross-linked with 1% formaldehyde for 10min in mild agitation. Glycine is added (125mM) to the cells to block the cross-linking reaction and pelleted cells are washed twice with DPBS at 4°C. Cell lysis is performed for 10min in cell lysis buffer (HEPES-KOH pH7.5 50mM, NaCl 140mM, EDTA 1mM, NP-40

0.5%, TritonX-100 0.25%, Glycerol 10%, EDTA-free Protease Inhibitor Complex (PIC, Roche) 1X ; 500µL/10^7 cells). Nuclei are resuspended nuclei lysis buffer (Tris-HCl 50mM, EDTA 10mM, SDS 1%, PIC 1X ; 1mL per 25×10^6 cells). Nuclei are sonicated by Covaris S220 sonicator (Peak power 220, Duty factor 20 and Cycle Burst 200 – SonoLab7 software) for 17min. The solution is centrifuged at 20 000g for 10min at 4°C. A supernatant fraction is collected to check the sonication quality in an electrophoretic 1.8% agarose gel (after incubation for 1h at 37°C with 0.7µg/µL of RNase A). DNA is diluted 1 :10 in IP buffer (SDS 0.01%, Triton X-100 1.1%, EDTA 1.2mM, Tris-HCl pH8 16.7mM, NaCl 167mM, PIC 1X). Chromatin is pre-cleared by incubation with Dynabeads (Invitrogen) blocked with DPBS 1X and BSA 0.5% (see Fig. A.1 for details on beads type). A fraction corresponding to 3% of an IP is collected as Input sample. For each condition 50µL of Dynabeads are coupled for 6h with specific antibodies (listed in Fig. A.1) at 4°C under rotation and then added to pre-cleared chromatin over-night at 4°C. Beads are washed with low salt buffer (NaCl 150mM, Tris pH 8 20mM, EDTA 2mM, Triton X-100 1%, SDS 0.1%), high salt buffer (NaCl 500mM, Tris pH 8 20mM, EDTA 2mM, Triton X-100 1%, SDS 0.1%), LiCl buffer (NP-40 1%, Na-DOC 1%, LiCl 250mM, EDTA 1mM, Tris pH 8 10mM) and twice with TE buffer (EDTA 1mM, Tris pH 8 10mM). Chromatin is then eluted by $NaHCO_3$ 0.1M and SDS 1% at 65°C for 30min. Chromatin is de-cross-linked by adding NaCl 200mM and incubation for 4h at 65°C. Then 0.1µg/µL of proteinase K (invitrogen) are added and incubated 1h at 45°C. DNA is purified by phenol/chlorophorme-isoamylalcol (24 :1) and precipitated by adding 0.7 volumes isopropanol, washed with cold ethanol 70% and resuspended in DNA/RNase-free distilled H_2O.

ChIP assays were repeated at least three times. The DNA precipitated from the different experiments performed using antibodies with the same specificity in the same cells was combined after verification of the quality of each ChIP on positive targets by quantitative PCR. Quantitative PCR was performed on a same volume of immunoprecipitated DNA or input sample three times in duplicate on the 7500 RT-qPCR System (Applied Biosystems) using home-designed oligonucleotides (primers listed in Fig. A.2) and SYBR-green buffer (Applied Biosystems). Immunoprecipitated DNA is compared to input sample by the comparative C_t method. Immunoprecipitated DNA enrichment is expressed as percentage of input (% input).

Probes for mRNA expression (RT-qPCR)

Gene	Probe reference (Applied Biosystems)
Alas2	Mm00802083_m1
Bcl2l11	Mm00437796_m1
Hdac1	Mm02391771_g1
Nprl3	Mm01193449_m1
Polr2a	Mm00839493_m1
Ptp4a3	Mm00477233_m1
Spi1	Mm00488142_m1

Primers for ChIP quantification (qPCR)

Target region	Forward primer sequence (5'-3')	Reverse primer sequence (5'-3')
Ahi1_iso2+13kb	CTCTTCTGCAAGAGGAAGTGAG	GGCTGGCTGTGTAGTCATAAA
Alas2+2.4kb	CTAGGGAGTAGCCAGACTCTAAT	CCTGGCCATGAAGGCTAAA
Bcl2l11+5.6kb	ACACCAGTTGATCTTTGCACACGC	TCGGAGTCCCTGCTGACAATCAAT
Nprl3+21kb	ACACTGTCTCCAAGCTTATCC	GCACAGTCATGGAGTGTAGAT
Ptp4a3_iso1+20kb	GAGGTGAATGTCACTAGCTCTG	TCTCTGGGATAAGCTCCTCTT
Rraga-0.2kb	GATTGTCAAGCCTACAGCACTA	TGAACCTGGGAGGGAGAAT

Regions are indicated as distance from the TSS (transcription Starting Site) of the corresponding gene

FIGURE A.2 – **Listes des sondes pour l'ARN et des amorces pour les ChIP GATA1**

A.4 ChIP library preparation and sequencing check bp for reads and single/paired end

ChIP samples were purified using Agencourt AMPure XP beads (Beckman Coulter) and quantified with the Qubit (Invitrogen). ChIP-seq libraries were prepared from 3-10ng of double-stranded purified DNA using the MicroPlex Library Preparation kit v2 (C05010014, Diagenode s.a., Seraing, Belgium), according to manufacturer's instructions. In the first step, the DNA was repaired and yielded molecules with blunt ends. In the next step, stem-loop adaptors with blocked 5'-ends were ligated to the 5'-end of the genomic DNA, leaving a nick at the 3'-end. The adaptors cannot ligate to each other and do not have single-strand tails, avoiding non-specific background. In the final step, the 3'-ends of the genomic DNA were extended to complete library synthesis and Illumina compatible indexes were added through a PCR amplification (7 to 10 cycles). Amplified libraries were purified and size-selected using Agencourt AMPure XP beads (Beckman Coulter) to remove unincorporated primers and other reagents. The libraries were sequenced on Illumina Hiseq2500 (for SPI1 and RNApol II) or Hiseq4000 (for the others ChIP) sequencer as Paired-End 100bp (for H3K27ac upon Entinostat treatment, H3K4me1) or

50bp (for H3K27ac, H3K27me3, H3K4me3) or Single-End 100bp (for RNAPolII, H3K36me3, SPI1) reads following Illumina's instructions. Image analysis and base calling were performed using RTA v2.7.7 and bcl2fastq v2.17.1.14. Adapter dimer reads were removed using DimerRemover (https ://sourceforge.net/projects/dimerremover/).

A.5 RNA extraction and quantification by RT-qPCR

Total RNA was isolated from cells using RNeasy Plus Mini Kit (Qiagen) and 1µg of cDNA were prepared using Superscript IV Reverse Transcriptase (LifeTechnologies). Quantitative PCR was performed on 20ng of cDNA three times in duplicate on the 7500 RT-qPCR System (Applied Biosystems) using TaqMan Gene Expression Assays (Applied Biosystems) (probes are listed in Fig. A.2). Fold changes in mRNA expression levels were calculated relative to control condition and normalized to *Polr2a* mRNA level by the comparative C_t method using the formula :

$$\left(2^{-\Delta\Delta Ct}\right) \tag{A.1}$$

A.6 RNA library preparation and sequencing

RNA-seq libraries were generated from 1mg of total RNA using TruSeq Stranded mRNA LT Sample Preparation Kit (Illumina, San Diego, CA), according to manufacturer's instructions. The final cDNA libraries were checked for quality and quantified using Agilent Bioanalyzer 2100. For cells treated only with doxycycline, libraries were sequenced on the Illumina Hiseq 2500 as paired-end 100bp reads following Illumina's instructions, whereas for cells treated with Entinostat libraries were sequenced on the Illumina Hiseq 4000 as paired-end 100bp reads following Illumina's instructions. Image analysis and base calling were performed using RTA v2.7.3 and bcl2fastq v2.17.1.14. Sequencing was performed to obtain at least 15.10^7 reads for each sample. Sequence reads were mapped onto mm10/GRCm38 assembly of mouse genome using TopHat [183] v2.0.14 and bowtie [121] v2-2.1.0.

A.7 RNA-seq analyses

For RNA-seq, RNA quality was evaluated using an Agilent Fragment Analyzer apparatus as described in the manufacturer's procedure (Agilent Technologies, Basel, Switzerland). RNA-seq libraries were generated from 300 ng of total RNA Illumina TruSeq RNA Sample Preparation Kit v2 (Part Number RS-122-2001). One hundred base-pair paired-stranded reads were sequenced on an Illumina HiSeq 2500 system with multiplexing. Sequencing was performed to obtain at least 15x107 reads for each sample. Sequence reads were mapped onto

mm10/GRCm38 assembly of mouse genome using STAR v2.5.3a [193]. Gene expression was quantified using htseq-count v0.6.1p1 [178] with gene annotations from Ensembl release 98. Statistical analysis was performed using R (3.3.2). Read counts have been normalized across samples with the median-of-ratios method proposed by Anders and Huber [194] to make these counts comparable between samples. Differential gene expression analysis was done using the method proposed by Love et al. [188] and implemented in the Bioconductor package DESeq2 version 1.16.1. P-values were adjusted for multiple testing using the Benjamini and Hochberg method [195].

A.8 ATAC assay, library preparation and sequencing

Cells were harvested and frozen in culture media containing FBS and 10% DMSO. Cryopreserved cells were sent to Active Motif to perform the ATAC-seq assay. The cells were then thawed in a 37° C water bath, pelleted, washed with cold PBS, and tagmented as previously described [196], with some modifications based on [197]. Briefly, cell pellets were resuspended in lysis buffer, pelleted, and tagmented using the enzyme and buffer provided in the Nextera Library Prep Kit (Illumina). Tagmented DNA was then purified using the MinElute PCR purification kit (Qiagen), amplified with 10 cycles of PCR, and purified using Agencourt AMPure SPRI beads (Beckman Coulter). Resulting material was quantified using the KAPA Library Quantification Kit for Illumina platforms (KAPA Biosystems), and sequenced with PE42 sequencing on the NextSeq 500 sequencer (Illumina).

A.9 RIME Assay

A.9.1 Chromatin Immunoprecipitation

Cells were fixed with 1% methanol-free formaldehyde for 8min and quenched with 0.125M glycine. Chromatin was isolated by the addition of lysis buffer, followed by disruption with a Dounce homogenizer. Lysates were sonicated and the DNA sheared to an average length of 300-500bp. Genomic DNA (Input) was prepared by treating aliquots of chromatin with RNase, proteinase K and heat for de-crosslinking, followed by ethanol precipitation. Pellets were resuspended and the resulting DNA was quantified on a NanoDrop spectrophotometer. Extrapolation to the original chromatin volume allowed quantitation of the total chromatin yield. An aliquot of chromatin ($150\mu g$) was precleared with protein G agarose beads (Invitrogen). Proteins of interest were immunoprecipitated using 15µg of antibody against SPI1 and protein G magnetic beads. Protein complexes were washed, then trypsin was used to remove the immunoprecipitate from beads and digest the protein sample. Protein digests were separated from the beads and

purified using a C18 spin column (Harvard Apparatus). The peptides were vacuum dried using a speedvac.

A.9.2 Mass Spectrometry

Digested peptides were analyzed by LC-MS/MS on a Thermo Scientific Q Exactive Orbitrap Mass spectrometer in conjunction with a Proxeon Easy-nLC II HPLC (Thermo Scientific) and Proxeon nanospray source. The digested peptides were loaded on a 100µm x 25 mm Magic C18 100Å 5U reverse phase trap where they were desalted online before being separated using a $75\mu m$ x 150mm Magic C18 200 Å3U reverse phase column. Peptides were eluted using a 90-minute gradient with a flow rate of 300nL/min. An MS survey scan was obtained for the m/z range 300-1600, MS/MS spectra were acquired using a top 15 method, where the top 15 ions in the MS spectra were subjected to HCD (High Energy Collisional Dissociation). An isolation mass window of 1.6m/z was for the precursor ion selection, and normalized collision energy of 27% was used for fragmentation. A five second duration was used for the dynamic exclusion.

A.9.3 Database Searching

Tandem was set up to search the UP_mouse_CrapE2F1_rev database (106444 entries), the cRAP database of common laboratory contaminants (www.thegpm.org/crap ; 114 entries) plus an equal number of reverse protein sequences assuming the digestion enzyme trypsin. Tandem was searched with a fragment ion mass tolerance of 20 PPM and a parent ion tolerance of 20 PPM. Glu->pyro-Glu of the n-terminus, ammonia-loss of the n-terminus, gln->pyro-Glu of the n-terminus, deamidated of asparagine and glutamine, oxidation of methionine and tryptophan and dioxidation of methionine and tryptophan were specified in Tandem as variable modifications.

A.9.4 Criteria for Protein Identification

Scaffold (version Scaffold_4.8.7, Proteome Software Inc., Portland, OR) was used to validate MS/MS based peptide and protein identifications. Peptide identifications were accepted if they could be established at greater than 99.0% probability by the Scaffold Local FDR algorithm. Peptide identifications were also required to exceed specific database search engine thresholds. X ! Tandem identifications required at least. Protein identifications were accepted if they could be established at greater than 6.0% probability to achieve an FDR less than 5.0% and contained at least 2 identified peptides. Protein probabilities were assigned by the Protein Prophet algorithm [198]. Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony. Proteins sharing significant peptide evidence were grouped into clusters. List Filtering. Protein coverage

percentage is observed in the raw data file. Final list generation was done by taking all proteins with a spectral count of five and above from each replicate reaction and comparing them in a venn-diagram against IgG control replicates. Proteins unique to both experimental replicates were then applied to the PANTHER database for protein ontology results.

A.10 Inhibitors cytotoxicity/cytostaticity and synergy assays

Cells have been treated with the combination of different concentrations of Entinostat (MS-275, Santa Cruz) (50, 150, 300nM) and UNC1999 (M3107, AbMole) (5, 7.5, 10, 12.5, 15µM) for 48h and compared to treatment with equal quantity of DMSO to define the nature of the interaction between the two drugs. For the proliferation and mortality curves as function of time, cells were treated with 150nM Entinostat and 10µM or 12.5µM UNC1999 up to 72h. Viability and proliferation of cells, labeled with DAPI 0.2mM, have been performed by BD High Throughput Sampler (HTS) option for the BD LSRFortessa flow cytometer in a 96-wells plate or by BD FACSCanto II flow cytometer. FACS data have been then analyzed by FlowJo v10.6.2 software.

Annexe B

Tableau des pics de SPI1

Nous avons construit un tableur excel contenant toutes les informations sur les pics de SPI1 dans toutes les annotations différentes. Chaque ligne correspond à un pic de SPI1. Si un gène présente plusieurs pics de SPI1 alors une ligne sera dédiée à chaque pic. Les sept premières colonnes **A à G** sont communes à tous les sites de fixation de SPI1 :

- geneName
- RPKM Spi1+
- RPKM Spi1-
- FoldChangeExpression (Spi1+/Spi1-)
- Regulation
- $pvaleur_{adjusted}$
- chromosome

Ensuite les colonnes sont groupées par neuf pour chaque localisation génomique des pics de SPI1.

- Les colonnes **H à P** correspondent aux pics annotés dans la catégorie **Intergénique**.
- Les colonnes **H à AA** correspondent aux pics annotés dans la catégorie **Promoteur distal**.
- Les colonnes **AC à AK** aux pics annotés dans la catégorie **ImmediateUpStreamTSS**.
- Les colonnes **AM à AU** aux pics annotés dans la catégorie **ImmediateDownStreamTSS**.
- Les colonnes **AW à BE** aux pics annotés dans la catégorie **exon**.
- Les colonnes **BG à BO** aux pics annotés dans la catégorie **intron**.
- Les colonnes **BQ à BY** aux pics annotés dans la catégorie **3'UTR**.
- les colonnes **CA à CI** aux pics annotés dans la catégorie **5'UTR**.
- Les colonnes **CK à CS** aux pics annotés dans la catégorie **DownStreamTE**.

Chacune des colonnes dans les groupes de neuf colonnes est dédiée à une information :

- **Colonne 1 :**Présence d'un pic dans la région de l'annotation : TRUE ou FALSE.
- **Colonne 2 :**Position peakmax du pic.
- **Colonne 3 :**Hauteur du pic : valeur dans le fichier .bigWig à la position peakmax.

- **Colonne 4 :**Statut du pic si il se trouve dans une annotation autre que promoteur : E3, E2 ou E1
- **Colonne 5 :**Distance du pic au TSS du gène sur lequel il est annoté.
- **Colonne 6 :**Score moyen de H3K27ac en condition Spi1+ dans la région +/- 1000bp autour de la position peakmax ($H3K27ac_{Spi1+}$).
- **Colonne 7 :**Score moyen de H3K27ac en condition Spi1- dans la région +/- 1000bp autour de la position peakmax ($H3K27ac_{Spi1-}$).
- **Colonne 8 :**Valeur du rapport $\frac{H3K27ac_{Spi1+}}{H3K27ac_{Spi1-}}$.
- **Colonne 9 :**Valeur du rapport $\frac{H3K27ac_{Spi1-}}{H3K27ac_{Spi1+}}$.

Ce tableau étant très volumineux il est disponible sur le lien de téléchargement suivant : `https://drive.google.com/file/d/1hOo4O--KupuWq3dKOlC9zaNq6thFN_x7/view?usp=sharing`.

Annexe C

Formats de fichier

C.1 Format FASTQ

Ce format est consistant et contient les lectures issues du processus de séquençage non encore alignées. Chaque lecture est encodée sur quatre lignes. La première ligne correspond à l'identifiant de la lecture, la seconde à la séquence, la troisième contient un + et la dernière ligne correspond à la qualité du séquençage, celle-ci est encodée pour chaque base par des caractères spéciaux.

C.2 Format BAM

La plupart des outils d'alignement génère les alignements dans le format SAM (Sequence Alignement/Map). Ce format est volumineux, ainsi le format BAM permet de stocker ces mêmes informations sous forme binaire. On présente dans la Fig. C.1 les différents champs du format BAM.

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

FIGURE C.1 – Liste des différents champs du format BAM issu de [199]

Les champs 5 et 11 représentent la qualité de l'alignment. La qualité de chaque nucléotide

aligné est codée par une lettre. Chaque lettre est associée à un nombre, et chaque nombre est égal à $-10\log_{10}(p)$ où p représente la probabilité que la position où le nucléotide est aligné soit mauvaise. Cela est codé par le champ 11 et le champ 5 représente le hphred (score de qualité de l'alignement).

Le champ 6 appelé CIGAR représente les évènements d'alignements. Les différentes lettres qui codent ce champ sont détaillées dans la Fig. C.2 et illustrées par l'exemple de la Fig. C.3 pour lequel le CIGAR est 3M1I3M1D5M.

Op	BAM	Description	Consumes query	Consumes reference
M	0	alignment match (can be a sequence match or mismatch)	yes	yes
I	1	insertion to the reference	yes	no
D	2	deletion from the reference	no	yes
N	3	skipped region from the reference	no	yes
S	4	soft clipping (clipped sequences present in SEQ)	yes	no
H	5	hard clipping (clipped sequences NOT present in SEQ)	no	no
P	6	padding (silent deletion from padded reference)	no	no
=	7	sequence match	yes	yes
X	8	sequence mismatch	yes	yes

FIGURE C.2 – Codage du champ CIGAR du format BAM [1]

```
RefPos:     1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T  A  A  C
Read:                  A  C  T  A  G  A  A     T  G  G  C  T
```

FIGURE C.3 – Exemple de l'alignement d'une lecture avec un CIGAR 3M1I3M1D5M

On présente ci-dessous un extrait d'un fichier SAM constitué de deux lectures alignées provenant des données de CLIP-seq Spi1+.

```
HWI-ST1339:203:H8LPYADXX:2:1109:14719:46006 16 chr4_GL456216_random
13662 60 51M * 0 0 GGGTTTTATTTTTTAGCGTTTCATGGTTGTTTAGGTTTTTATGTTAAGCGT
JJJJJJJIJJJJJJJJJIJJ
JIJJJJJJJJJJJJJJJJHHHHHFFFFFCCC NM:i:1 MD:Z:27C23 AS:i:46 XS:i:19
HWI-ST1339:203:H8LPYADXX:1:1201:15651:45886 16 chr4_GL456216_random
15892 70 12M1D29M * 0 0 GGACGTCACCGCGCTGCCAATTCTCTTCTCCGCGTCCTCTT
DDFFFHHHJJJJJJJJJJJJ
JJJJJJJJHHHHHFFFFFCCC AS:i:55 UQ:i:55 NM:i:2 GN:Z:585 TN:Z:NM_001033326
NH:i:1 PG:Z:novoalign
```

1. http ://samtools.github.io/hts-specs/SAMv1.pdf

C.3 Format BED

Ce format sert à contenir des informations sur la localisation de régions génomiques d'intêret telles que des sites de fixation d'un facteur de transcription. Il est constitué de trois champs obligatoires :

— chromosome
— start
— stop

Les champs additionnels sont :

— nom
— score
— brin
— thickStart
— thickEnd
— itemRgb
— blockCount
— blockSizes
— blockStarts

Dans le cadre de ce travail de thèse nous utilisons majoritairement les six premiers champs.

C.4 Format bigWig

Ce format est un format binaire indexé du format Wig. Il contient des informations sur la densité d'une protéine ou d'une marque d'histone. La première ligne correspond à l'entête qui donne des informations sur la taille du pas (bin) utilisé pour encoder l'information de la densité. Il est visualisable sur le logiciel IGV.

Annexe D

Résultat Zagros

```
AC ZAGROS0
XX
TY Motif
XX
```

P0	A	C	G	T
01	40	10	12	35
02	11	53	28	5
03	45	3	4	45
04	6	7	78	6
05	58	4	8	27
06	25	3	34	35
07	18	15	55	9
08	48	15	27	7
09	33	36	13	15
010	19	44 7	27	

```
XX
BS tctgaggaca; chr1:72255073-72255093(+); 9; 10;  ;+ 1
BS ACAGATCTCA; chr1:143650245-143650265(+); 5; 10;  ;+ 1
BS TAAGTAGGCT; chr1:159264342-159264362(+); 7; 10;  ;+ 1
BS agtgatagac; chr1:179986528-179986548(+); 3; 10;  ;+ 1
BS AGATTTGGCC; chr1:180125216-180125236(+); 3; 10;  ;+ 1
BS ggagaggaac; chr1:192745623-192745643(+); 5; 10;  ;+ 1
BS AGAGAACAGC; chr3:37483512-37483532(+); 7; 10;  ;+ 1
BS ggtcttgaac; chr3:51237570-51237590(+); 3; 10;  ;+ 1
BS tcaggtctgc; chr3:60536227-60536247(+); 0; 10;  ;+ 1
BS cccgaggagc; chr3:84902118-84902138(+); 7; 10;  ;+ 1
```

```
BS GCTGTAGAAT; chr3:98033684-98033704(+); 10; 10;  ;+ 1
BS TGAAAGAACT; chr4:44764764-44764784(+); 1; 10;  ;+ 1
BS acaaaggtat; chr4:44789624-44789644(+); 5; 10;  ;+ 1
BS TCAAACGCCT; chr4:80002936-80002956(+); 7; 10;  ;+ 1
BS tctgagtctt; chr4:117190011-117190031(+); 1; 10;  ;+ 1
BS tctgagtctt; chr4:117210905-117210925(+); 9; 10;  ;+ 1
BS acgttggaAT; chr4:132352514-132352534(+); 5; 10;  ;+ 1
BS tcggaTAGTC; chr4:132352966-132352986(+); 9; 10;  ;+ 1
BS acagatagaa; chr4:147252206-147252226(+); 6; 10;  ;+ 1
BS gcacgagacc; chr5:146261170-146261190(+); 8; 10;  ;+ 1
BS tctccgctcc; chr6:47781686-47781706(+); 4; 10;  ;+ 1
BS GCTGAACAAC; chr6:134072424-134072444(+); 3; 10;  ;+ 1
BS ACAGATAAGT; chr6:134091145-134091165(+); 8; 10;  ;+ 1
BS agtgttgaac; chr6:134153697-134153717(+); 7; 10;  ;+ 1
BS ttagtaggac; chr7:97103018-97103038(+); 7; 10;  ;+ 1
BS actgaagaag; chr7:97138458-97138478(+); 5; 10;  ;+ 1
BS actggtcAGC; chr7:143267832-143267852(+); 5; 10;  ;+ 1
BS TCATGGGGTC; chr7:143286200-143286220(+); 10; 10;  ;+ 1
BS tgtgtagagc; chr10:40258518-40258538(+); 5; 10;  ;+ 1
BS aatgatgatt; chr11:69073466-69073486(+); 10; 10;  ;+ 1
BS TGTGTGAACT; chr11:95724725-95724745(+); 7; 10;  ;+ 1
BS TGAGTGGTTC; chr11:107430167-107430187(+); 5; 10;  ;+ 1
BS cgtgaggaca; chr12:59135344-59135364(+); 9; 10;  ;+ 1
BS acggagggag; chr12:59135644-59135664(+); 6; 10;  ;+ 1
BS cctgagaatg; chr12:77245650-77245670(+); 6; 10;  ;+ 1
BS agagatagac; chr12:77380705-77380725(+); 3; 10;  ;+ 1
BS AGTGTGTATT; chr12:77394222-77394242(+); 5; 10;  ;+ 1
BS tctgcagact; chr12:78255252-78255272(+); 0; 10;  ;+ 1
BS CTAGATCAAG; chr12:78265778-78265798(+); 4; 10;  ;+ 1
BS AGAGATACCC; chr13:35762870-35762890(+); 7; 10;  ;+ 1
BS AGAGATAGGA; chr13:35768156-35768176(+); 3; 10;  ;+ 1
BS TCTGAATAGA; chr13:109777102-109777122(+); 8; 10;  ;+ 1
BS AGGGAGGACA; chr13:109781948-109781968(+); 10; 10;  ;+ 1
BS ACAAAAGGTA; chr13:109786588-109786608(+); 5; 10;  ;+ 1
BS CGAGAAGACT; chr13:109789616-109789636(+); 7; 10;  ;+ 1
BS GCTGAGCACA; chr13:109798886-109798906(+); 5; 10;  ;+ 1
BS AGTGATGAAC; chr13:109806952-109806972(+); 1; 10;  ;+ 1
BS ATTGTGTCCC; chr13:113820528-113820548(+); 4; 10;  ;+ 1
```

```
BS AGATAGAAAC; chr14:21118134-21118154(+); 0; 10;  ;+ 1
BS tctgatgact; chr15:62040788-62040808(+); 3; 10;  ;+ 1
BS acagaaagaa; chr15:62191390-62191410(+); 9; 10;  ;+ 1
BS TGAGAAGAAT; chr15:83149705-83149725(+); 4; 10;  ;+ 1
BS acatttggtc; chrX:100027372-100027392(+); 3; 10;  ;+ 1
BS actcatgaag; chrX:100041012-100041032(+); 10; 10;  ;+ 1
BS acagagagac; chrY:90813736-90813756(+); 1; 10;  ;+ 1
BS tatgttggcc; chr1:33614570-33614590(-); 1; 10;  ;+ 1
BS tatgttcctc; chr1:192745623-192745643(-); 2; 10;  ;+ 1
BS tgagttggca; chr3:153374250-153374270(-); 5; 10;  ;+ 1
BS ACAGACCCAC; chr3:153401968-153401988(-); 2; 10;  ;+ 1
BS TGAGAGCCAT; chr3:153474642-153474662(-); 7; 10;  ;+ 1
BS ACTGTACCTT; chr3:153474664-153474684(-); 9; 10;  ;+ 1
BS TCAGATGACT; chr3:153503318-153503338(-); 2; 10;  ;+ 1
BS aatattgaac; chr3:153589030-153589050(-); 6; 10;  ;+ 1
BS TCTGAAGAAT; chr3:153611088-153611108(-); 5; 10;  ;+ 1
BS acagagggca; chr3:153624656-153624676(-); 8; 10;  ;+ 1
BS CATGAGGACC; chr3:153662956-153662976(-); 1; 10;  ;+ 1
BS ggtgaaaggt; chr3:153667796-153667816(-); 0; 10;  ;+ 1
BS GCTGTTGGCC; chr3:153675297-153675317(-); 6; 10;  ;+ 1
BS TCAGAGGCAC; chr3:153692664-153692684(-); 10; 10;  ;+ 1
BS TGTGAGGACT; chr3:153708814-153708834(-); 5; 10;  ;+ 1
BS GCTGCAGGGC; chr3:153714476-153714496(-); 0; 10;  ;+ 1
BS AGACATtcca; chr4:132352514-132352534(-); 1; 10;  ;+ 1
BS CCTGAAGAAC; chr4:133753304-133753324(-); 3; 10;  ;+ 1
BS TCCGATGATT; chr5:100426940-100426960(-); 6; 10;  ;+ 1
BS tcagatcaca; chr6:134170880-134170900(-); 4; 10;  ;+ 1
BS ccagttgaac; chr7:143267832-143267852(-); 10; 10;  ;+ 1
BS gcacattacc; chr8:31149866-31149886(-); 4; 10;  ;+ 1
BS TAAGAGCGCT; chr9:24745174-24745194(-); 5; 10;  ;+ 1
BS TTTGCGGACC; chr9:44818334-44818354(-); 0; 10;  ;+ 1
BS aaagttgtcc; chr11:34681900-34681920(-); 0; 10;  ;+ 1
BS ACTAAGGAGA; chr13:32201867-32201887(-); 3; 10;  ;+ 1
BS tttgaaggcc; chr13:97190602-97190622(-); 6; 10;  ;+ 1
BS TCTGTAGTCT; chr14:10331014-10331034(-); 4; 10;  ;+ 1
BS AGATAGAACC; chr14:61679714-61679734(-); 1; 10;  ;+ 1
BS GGAGATTAGA; chr15:53248576-53248596(-); 9; 10;  ;+ 1
BS ACCGTAGCAT; chr15:64265982-64266002(-); 4; 10;  ;+ 1
```

```
BS aatggacacc; chr15:64280898-64280918(-); 6; 10;  ;+ 1
BS ACAGGTGATC; chr15:95824651-95824671(-); 5; 10;  ;+ 1
BS CCTGTCGCAG; chr16:4088276-4088296(-); 10; 10;  ;+ 1
BS TCAGTTTCTC; chr16:35808850-35808870(-); 0; 10;  ;+ 1
BS agagagagac; chr16:92716556-92716576(-); 10; 10;  ;+ 1
BS GATGTAGAAA; chr16:92775680-92775700(-); 7; 10;  ;+ 1
BS CCTGAGGCCA; chr16:92777078-92777098(-); 5; 10;  ;+ 1
BS TCTGGAAGGA; chr16:92786722-92786742(-); 3; 10;  ;+ 1
BS actcgtggat; chr16:92798452-92798472(-); 5; 10;  ;+ 1
BS acagagagac; chr16:92801265-92801285(-); 6; 10;  ;+ 1
BS taagtgggcg; chr17:51141372-51141392(-); 9; 10;  ;+ 1
XX
//
```

Bibliographie

[1] S. Frietze and P. J. Farnham. Transcription factor effector domains. *Subcell Biochem*, 52(261–277), 2011.

[2] A. Mora, G. K. Sandve, O. Stokke Gabrielsen, and R. Eskeland. In the loop : promoter–enhancer interactions and bioinformatics. *Briefings in Bioinformatics*, page bbv097, Nov 2015.

[3] V. Haberle and A. Stark. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol*, 19(10) :621–637, 10 2018.

[4] O. Lancho and D. Herranz. The MYC enhancer-ome : Long-range transcriptional regulation of MYC in cancer. *Trends in Cancer*, 4(12) :810–822, 2018.

[5] Huret J.L., Ahmad M., Arsaban M., Bernheim A., Cigna J., Desangles F., Guignard J.C., Jacquemot-Perbal M.C., Labarussias M., Leberre V., Malo A., Morel-Pair C., Mossafa H., Potier J.C., Texier G., Viguié F., Yau Chun Wan-Senon S., Zasadzinski A., and Dessen P. Atlas of genetics and cytogenetics in oncology and haematology in 2013. *Nucleic Acids Res.*, 41, 2013.

[6] E. Wingender, T. Schoeps, M. Haubrock, M. Krull, and J. Dönitz. TFClass : expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Research*, 46(D1) :D343–D347, Oct 2017.

[7] A. J. Koleske and R. A. Young. An RNA polymerase II holoenzyme responsive to activators. *Nature*, 31, 1994.

[8] C. Esnault, Y. Ghavi-Helmn, S. Brun, J. Soutourina, N. Van Berkum, C. Boschiero, F. Holstege, and M. Werner. Mediator-dependent recruitment of tfiih modules in preinitiation complex. *Mol Cell*, 31(3) :337–346, 2008.

[9] A. Dvir. Promoter escape by RNA polymerase II. *Biochim Biophys Acta.*, 577(2) :208–223, 2002.

[10] Z. Ni, A. Saunders, N. J Fuda, J. Yao, J.-R. Suarez, W. W. Webb, and J. T. Lis. P-TEFb is critical for the maturation of RNA polymerase II into productive elongation in vivo. *Molecular and Cellular Biology*, 28(3) :1161–1170, 2008.

[11] Z. Yang, J. H. N. Yik, R. Chen, N. He, M. Kyoo Jang, K. Ozato, and Q. Zhou. Recruitment of p-TEFb for stimulation of transcriptional elongation by the bromodomain protein brd4. *Molecular Cell*, 19(4) :535–545, 2005.

[12] S R Eberhardy and P J Farnham. c-Myc mediates activation of the cad promoter via a post-RNA polymerase II recruitment mechanism. *Journal of Biological Chemistry*, 276(51) :48562 – 71, 2001.

[13] B R Cairns, Y J Kim, M H Sayre, B C Laurent, and R D Kornberg. A multisubunit complex containing the swi1/adr6, swi2/snf2, swi3, snf5, and snf6 gene products isolated from yeast. *Proceedings of the National Academy of Sciences*, 91(5) :1950–1954, 1994.

[14] T. Ito, M. Bulger, M. J. Pazin, R. Kobayashi, and J. T. Kadonaga. Acf, an iswi-containing and atp-utilizing chromatin assembly and remodeling factor. *Cell*, 90 :145–155, 1997.

[15] J. K. Tong, C. A. Hassig, G. R. Schnitzler, R. E. Kingston, and S. L. Schreiber. Chromatin deacetylation by an ATP-dependent nucleosome remodelling complex. *Nature*, 395(6705) :917–921, 1998.

[16] Health-innovations, {https ://health-innovations.org/2016/01/06}.

[17] O. I. Kulaeva, D. Gaykalova, and V. M. Studitsky. Transcription through chromatin by RNA polymerase II : Histone displacement and exchange. *Mutat Res.*, 618(1-2) :116–129, 2007.

[18] E. Calo and J. Wysocka. Modification of enhancer chromatin : What, how, and why ? *Molecular Cell*, 29 :825–837, 2013.

[19] A. Eberharter and P. B. Becker. Histone acetylation : a switch between repressive and permissive chromatin. *EMBO reports*, 3(3) :224–229, 2002.

[20] D. J. Gaffney, G. McVicker, A. A. Pai, Y. N. Fondufe-Mittendorf, N. Lewellen, K. Michelini, J. Widom, Y. Gilad, and J. K. Pritchard. Controls of nucleosome positioning in the human genome. *PLoS Genetics*, 8(11) :e1003036, Nov 2012.

[21] D. Tillo, N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, Y. Field, J. D. Lieb, J. Widom, E. Segal, and T. R. Hughes. High nucleosome occupancy is encoded at human regulatory sequences. *PLoS ONE*, 5(2) :e9129, Feb 2010.

[22] K. S. Zaret and J. S. Carroll. Pioneer transcription factors : establishing competence for gene expression. *Genes Dev*, 25(21) :2227–41, Nov 2011.

[23] E. Laurenti and B. Göttgens. From haematopoietic stem cells to complex differentiation landscapes. *Nature*, 553(7689) :418–426, Jan 2018.

[24] D. A. Jaitin, H. Keren-Shaul, N. Elefant, and I. Amit. Each cell counts : Hematopoiesis and immunity research in the era of single cell genomics. *Seminars in Immunology*, 27(1) :67 – 71, 2015.

[25] F. Paul, Y. Arkin, A. Giladi, D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, D. Winter, D. Lara-Astiaso, M. Gury, A. Weiner, E. David, N. Cohen, F. K. B. Lauridsen, S. Haas, A. Schlitzer, A. Mildner, F. Ginhoux, S. Jung, A. Trumpp, B. T. Porse, A. Tanay, and I. Amit. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 163 :1663–1677, 2015.

[26] D. A. Jaitin, E. Kenigsberg, N. Keren-Shaul, H.and Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, and I. Amit. Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science*, 343(6172) :776–779, 2014.

[27] D. Lara-Astiaso, A. Weiner, E. Lorenzo-Vivas, I. Zaretsky, D. A. Jaitin, E. David, H. Keren-Shaul, A. Mildner, D. Winter, S. Jung, N. Friedman, and I. Amit. Chromatin state dynamics during blood formation. *Science*, 345(6199) :943–949, 2014.

[28] J. Huang, X. Liu, D. Li, Z. Shao, H. Cao, Y. Zhang, E. Trompouki, T. V. Bowman, L. I. Zon, G. Yuan, S. H. Orkin, and J. Xu. Dynamic control of enhancer repertoires drives lineage and stage-specific transcription during hematopoiesis. *Developmental Cell*, 36 :9–23, 2016.

[29] G.-H. Wei, G. Badis, M. F. Berger, T. Kivioja, K. Palin, M. Enge, M. Bonke, A. Jolma, M. Varjosalo, A. R. Gehrke, J. Yan, S. Talukder, M. Turunen, M. Taipale, H. G. Stunnenberg, E. Ukkonen, T. R Hughes, M. L Bulyk, and J. Taipale. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J*, 29(13) :2147–60, Jul 2010.

[30] Christopher T Workman, Yutong Yin, David L Corcoran, Trey Ideker, Gary D Stormo, and Panayiotis V Benos. enologos : a versatile web tool for energy normalized sequence logos. *Nucleic Acids Res*, 33(Web Server issue) :W389–92, Jul 2005.

[31] C. Hagemeier, A. J. Bannister, A. Cook, and T. Kouzarides. The activation domain of transcription factor PU.1 binds the retinoblastoma (RB) protein and the transcription factor TFIID invitro : RB shows sequence similarity to TFIID and TFIIB. *Proc. Natl. Acad. Sci.*, 90 :1590–1584, 1992.

[32] J. M. R. Pongubala, l. Nagulapalli, M. J.Klemsz, S. R. McKercher, R. A. Maki, and M. L. Atchison. PU.1 recruits a second nuclear factor to a site important for immuno globulin K 3' enhancer activity. *Molecular and cellular biology*, 12(11) :368–378, 1992.

[33] T. A. Lodie, R. Jr. Savedra, D. T. Golenbock, C. P. Van Beveren, R. A. Maki, and M. J. Fenton. Stimulation of macrophages by lipopolysaccharide alters the phosphorylation

state, conformation, and function of PU.1 via activation of casein kinase II. *Journal of immunology (Baltimore, Md. : 1950)*, 158(4) :1848 – 1856, 1997.

[34] P. Rieske and J. M. Pongubala. AKT induces transcriptional activity of PU.1 through phosphorylation-mediated modifications within its transactivation domain. *The Journal of biological chemistry*, 276(11) :8460 – 8468, 2001.

[35] J.-M. Wang, M.-Z. Lai, and H.-F. Yang-Yen. Interleukin-3 stimulation of *mcl-1* gene transcription involves activation of the PU.1 transcription factor through a p38 mitogen-activated protein kinase-dependent pathway. *Molecular and cellular biology*, 23(6) :1896 – 1909, 2003.

[36] A. C. Azim, X. Wang, G. Y. Park, R. T. Sadikot, H. Cao, B. Mathew, M. Atchison, R. B. Van Breemen, M. Joo, and J. W. Christman. NF-kB-inducing kinase regulates cyclooxygenase 2 gene expression in macrophages by phosphorylation of PU.1. *The Journal of Immunology*, 179(11) :7868–7875, 2007.

[37] M. Ridinger-Saison, V. Boeva, P. Rimmelé, I. Kulakovskiy, I. Gallais, B. Levavasseur, C. Paccard, P. Legoix-Né, F. Morlé, A. Nicolas, P. Hupe, E. Barillot, F. Moreau-Gachelin, and Guillouf C. Spi-1/PU.1 activates transcription through clustered DNA occupancy in erythroleukemia. *Nucleic Acids Research*, 40(18) :8927–8941, Jul 2012.

[38] Y. Li, Y. Okuno, P. Zhang, H. S. Radomska, H. Chen, H. Iwasaki, K. Akashi, M. J. Klemsz, S. R. McKercher, R. A. Maki, and D. G. Tenen. Regulation of the PU.1 gene by distal elements. *Blood*, 98(10) :2958–2965, 2001.

[39] U. Steidl, C. Steidl, A. Ebralidze, B. Chapuy, H. Han, B. Will, F. Rosenbauer, A. Becker, K. Wagner, S. Koschmieder, S. Kobayashi, D. B. Costa, T. Schulz, K. B. O'Brien, R.G.W. Verhaak, R. Delwel, D. Haase, L. Trümper, J. Krauter, T. Kohwi-Shigematsu, F. Griesinger, and D. G. Tenen. A distal single nucleotide polymorphism alters long-range regulation of the PU.1 gene in acute myeloid leukemia. *The Journal of Clinical Investigation*, 117(9) :2611–2620, 9 2007.

[40] F. Rosenbauer, K. Wagner, J. L. Kutok, H. Iwasaki, M. M. Le Beau, Y. Okuno, K. Akashi, S. Fiering, and D. G. Tenen. Acute myeloid leukemia induced by graded reduction of a lineage-specific transcription factor, PU.1. *Nature Genetics*, (6) :624, 2004.

[41] F. Rosenbauer, B. M Owens, L. Yu, J. R Tumang, U. Steidl, J. L. Kutok, L. K. Clayton, K. Wagner, M. Scheller H. Iwasaki, C. Liu, B. Hackanson, K. Akashi, A. Leutz, T. L. Rothstein, C. Plass, and D. G. Tenen. Lymphoid cell growth and transformation are suppressed by a key regulatory element of the gene encoding PU.1. *Nature Genetics*, 38(1) :27–37, 2006.

[42] M. Leddin, C. Perrod, M. Hoogenkamp, S. Ghani, S. Assi, S. Heinz, N. K. Wilson, G. Follows, J. Schönheit, L. Vockentanz, A. M. Mosammam, W. Chen, D. G. Tenen, D. R. Westhead, B. Göttgens, C. Bonifer, and F. Rosenbauer. Two distinct auto-regulatory loops operate at the PU.1 locus in B cells and myeloid cells. *Blood*, 117(10) :2827–38, Mar 2011.

[43] A. K. Ebralidze, F. C. Guibal, U. Steidl, P. Zhang, S. Lee, B. Bartholdy, M. A. Jorda, V. Petkova, F. Rosenbauer, G. Huang, T. Dayaram, J. Klupp, K. B. O'Brien, B. Will, M. Hoogenkamp, K. L. B. Borden, C. Bonifer, and D. G. Tenen. PU.1 expression is modulated by the balance of functional sense and antisense RNAs regulated by a shared cis-regulatory element. *Genes Dev*, 22(15) :2085–92, Aug 2008.

[44] S. R. McKercher, B. E.Torbett, K. L. Anderson, G. W. Henkel, D. J. Vestal, H. Baribault, M. Klemsz, A. J. Feeney, G. E. Wu, C. J. P., and R. A. Maki. Targeted disruption of the PU.1 gene results in multiple hematopoietic abnormalities. *The EMBO Journal*, 15(20) :5647–5658, 1996.

[45] S. L. Nutt, D. Metcalf, A. D'Amico, M. Polli, and L. Wu. Dynamic regulation of PU.1 expression in multipotent hematopoietic progenitors. *J Exp Med*, 201(2) :221–231, 2005.

[46] J. Back, D. Allman, S. Chan, and P. Kastner. Visualizing PU.1 activity during hematopoiesis. *Experimental Hematology*, 33(4) :395–402, 2005.

[47] P. B Staber, P. Zhang, M. Ye, R. S. Welner, C. Nombela-Arrieta, C. Bach, M. Kerenyi, B. A. Bartholdy, H. Zhang, M. Alberich-Jordà, S. Lee, H. Yang, F. Ng, J. Zhang, M. Leddin, L. E. Silberstein, G. Hoefler, S. H. Orkin, B. Göttgens, F. Rosenbauer, G. Huang, and D. G. Tenen. Sustained PU.1 levels balance cell-cycle regulators to prevent exhaustion of adult hematopoietic stem cells. *Molecular Cell*, 49(5) :934–946, 2013.

[48] A. Dakic, D. Metcalf, L. Di Rago, S. Mifsud, L. Wu, and Stephen L. Nutt. PU.1 regulates the commitment of adult hematopoietic progenitors and restricts granulopoiesis. *J Exp Med*, 201(9) :1487–1502, 2005.

[49] H. Iwasaki, C. Somoza, H. Shigematsu, E. A. Duprez, J. Iwasaki-Arai, S. Mizuno, Y. Arinobu, K. Geary, P. Zhang, T. Dayaram, M. L. Fenyus, S. Elf, S. Chan, P. Kastner, C. S. Huettner, R. Murray, D. G. Tenen, and K. Akashi. Distinctive and indispensable roles of PU.1 in maintenance of hematopoietic stem cells and their differentiation. *Blood*, 106(5) :1590–1600, 2005.

[50] M.B. Kamath, I.B. Houston, A.J. Janovski, X. Zhu, S. Gowrisankar, A.G. Jegga, and R.P. DeKoter. Dose-dependent repression of T-cell and natural killer cell genes by PU.1 enforces myeloid and B-cell identity. *Leukemia*, 22 :1214–1225, 2008.

[51] V. Vegesna, S. Takeuchi, W.-K. Hofmann, T. Ikezoe, S. Tavor, U. Krug, A. C. Fermin, A. Heaney, C. W. Miller, and H. P. Koeffler. C/EBP-beta, C/EBP-delta, PU.1, AML1 genes : mutational analysis in 381 samples of hematopoietic and solid malignancies. *Leukemia research*, 26(5) :451 – 457, 2002.

[52] V.-P. Lavallée, I. Baccelli, J. Krosl, B. Wilhelm, F. Barabé, P. Gendron, G. Boucher, S. Lemieux, A. Marinier, S. Meloche, J. Hébert, and G. Sauvageau. The transcriptomic landscape and directed chemical interrogation of mll-rearranged acute myeloid leukemias. *Nature Genetics*, 47(9) :1030–1037, 2015.

[53] R. K. Vangala, M. S. Heiss-Neumann, J. S. Rangatia, S. M. Singh, C. Schoch, D. G. Tenen, W. Hiddemann, and G. Behre. The myeloid master regulator transcription factor PU.1 is inactivated by AML1-ETO in t(8 ;21) myeloid leukemia. *Blood*, 101(1) :270 – 277, 2003.

[54] R. Zheng, A. D. Friedman, M. Levis, L. Li, E. G. Weir, and D. Small. Internal tandem duplication mutation of FLT3 blocks myeloid differentiation through suppression of C/EBPalpha expression. *Blood*, 103(5) :1883 – 1890, 2004.

[55] B. U. Mueller, T. Pabst, J/ Fos, V. Petkovic, M. F Fey, N. Asou, U. Buergi, and D. G. Tenen. ATRA resolves the differentiation block in t(15 ;17) acute myeloid leukemia by restoring PU.1 expression. *Blood*, 107(8) :3330 – 3338, 2006.

[56] X. Gu, Q. Ebrahem, R. Z. Mahfouz, M. Hasipek, F. Enane, T. Radivoyevitch, N. Rapin, B. Przychodzen, Z. Hu, R. Balusu, C. V. Cotta, D. Wald, C. Argueta, Y. Landesman, M. P. Martelli, B. Falini, H. Carraway, B. T. Porse, J. Maciejewski, B. K. Jha, and Y. Saunthararajah. Leukemogenic nucleophosmin mutation disrupts the transcription factor hub that regulates granulomonocytic fates. *The Journal of Clinical Investigation*, 128(10) :4260–4279, 10 2018.

[57] I. Antony-Debré, A. Paul, J. Leite, K. Mitchell, H. M. Kim, L. A. Carvajal, T. I. Todorova, K. Huang, A. Kumar, A. A. Farahat, B. Bartholdy, S. Narayanagari, J. Chen, A. Ambesi-Impiombato, A. A. Ferrando, I. Mantzaris, E. Gavathiotis, A. Verma, B. Will, D. W. Boykin, W. David Wilson, G. M.K. Poon, and U. Steidl. Pharmacological inhibition of the transcription factor PU.1 in leukemia. *The Journal of Clinical Investigation*, 127(12) :4297–4313, 12 2017.

[58] D. Roos-Weil, B. Giacopelli, M. Armand, V. Della-Valle, H. Ghamlouch, C. Decaudin, M. Metzner, J. Lu, M. LeGarff-Tavernier, V. Leblond, P. Vyas, T. Zenz, F. Nguyen-Khac, O. A. Bernard, and C. C. Oakes. Identification of 2 DNA methylation subtypes of Waldenström macroglobulinemia with plasma and memory B-cell features. *Blood*, 136(5), 2020.

[59] M. Seki and J. Takita. Recurrent SPI1 fusions in pediatric T-cell acute lymphoblastic leukemia : novel mutations with poor prognosis. *[Rinsho ketsueki] The Japanese journal of clinical hematology*, 59(4) :439 – 447, 2018.

[60] D. L. Linemeyer, S. K. Ruscetti, E. M. Scolnick, L. H. Evans, and P. H. Duesberg. Biological activity of the spleen focus-forming virus is encoded by a molecularly cloned subgenomic fragment of spleen focus-forming virus DNA. *Proc Natl Acad Sci USA*, 78(3) :1401–1405, 1981.

[61] F Moreau-Gachelin, J Robert-Lezenes, F Wendling, A Tavitian, and P Tambourin. Integration of spleen focus-forming virus proviruses in friend tumor cells. *Journal of virology*, 53(1) :292 – 295, 1985.

[62] F. Moreau-Gachelin, A. Tavitian, and P. Tambourin. Spi-1 is a putative oncogene in virally induced murine erythroleukaemias. *Nature*, 331(6153) :277 – 280, 1988.

[63] F Moreau-Gachelin, D Ray, M G Mattei, P Tambourin, and A Tavitian. The putative oncogene Spi-1 : murine chromosomal localization and transcriptional activation in murine acute erythroleukemias. *Oncogene*, 4(12) :1449 – 1456, 1989.

[64] C. Friend, W. Scher, J. G. Holland, and T. Sato. Hemoglobin synthesis in murine virus-induced leukemic cells in vitro : Stimulation of erythroid differentiation by dimethyl sulfoxide. *Proceedings of the National Academy of Sciences of the United States of America*, 68(2) :378–382, 1971.

[65] F. Moreau-Gachelin, F. Wendling, T. Molina, N. Denis, M. Titeux, G. Grimber, P. Briand, W. Vainchenker, and A. Tavitan. Spi-1/PU.1 Transgenic Mice Develop Multistep Erythroleukemias. *Molecular and Cellular Biology*, 16(5) :2453–2463, 1996.

[66] M. Ridinger-Saison, E. Evanno, I. Gallais, P. Rimmelé, D. Selimoglu-Buet, E. Sapharikas, F. Moreau-Gachelin, and C. Guillouf. Epigenetic silencing of *bim* transcription by Spi-1/PU.1 promotes apoptosis resistance in leukaemia. *Cell Death Differ*, 20(9) :1268–78, Sep 2013.

[67] P. Rimmelé, O. Kosmider, P. Mayeux, F. Moreau-Gachelin, and C. Guillouf. Spi-1/PU.1 participates in erythroleukemogenesis by inhibiting apoptosis in cooperation with Epo signaling and by blocking erythroid differentiation. *Neoplasia*, 109(7), 2007.

[68] O. Kosmider, N. Denis, C. Lacout, W/ Vainchenker, P. Dubreuil, and F. Moreau-Gachelin. Kit-activating mutations cooperate with Spi-1/PU.1 overexpression to promote tumorigenic progression during erythroleukemia in mice. *Cancer Cell*, 8(6) :467–478, Dec 2005.

[69] R. Monni, L. Haddaoui, A. Naba, I. Gallais, M. Arpin, P. Mayeux, and F. Moreau-Gachelin. Ezrin is a target for oncogenic Kit mutants in murine erythroleukemia. *Blood*, 111(6), 2008.

[70] J. J. Hayes and A. P. Wolffe. The interaction of transcription factors with nucleosomal DNA. *Bioessays*, 14(9) :597–603, 1992.

[71] S. Ghisletti, I. Barozzi, F. Mietton, S. Polletti, F. De Santa, E. Venturini, L. Gregory, L. Lonie, A. Chew, and C.-L. Wei. Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity*, 32(3) :317–328, Mar 2010.

[72] S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, 38(4) :576–589, May 2010.

[73] I. Barozzi, M. Simonatto, S. Bonifacio, L. Yang, R. Rohs, S. Ghisletti, and G. Natoli. Coregulation of Transcription Factor Binding and Nucleosome Occupancy through DNA sample Features of Mammalian Enhancers. *Cell*, 54(5) :844–857, 2014.

[74] M. F. Garcia, C. D. Moore, K. N. Schulz, O. Alberto, G. Donague, M. M. Harrison, H. Zhu, and K. S. Zaret. Structural features of transcription factors associating with nucleosome binding. *Molecular Cell*, 75 :921–932, 2019.

[75] J. Minderjahn, A. Schmidt, A. Fuchs, R. Schill, J. Raithel, M. Babina, C. Schmidl, C. Gebhard, S. Schmidhofer, K. Mendes, A. Ratermann, D. Glatz, M. Nützel, M. Edinger, P. Hoffmann, R. Spang, G. Längst, A. Imhof, and M. Rehli. Mechanisms governing the pioneering and redistribution capabilities of the non-classical pioneer PU.1. *Nature Communications*, 11(402), 2020.

[76] X. Gu, Z. Hu, Q. Ebrahem, J. S. Crabb, R. Z. Mahfouz, T. Radivoyevitch, J. W. Crabb, and Y. Saunthararajah. RUNX1 regulation of PU.1 corepressor/coactivator exchange identifies specific molecular targets for leukemia differentiation therapy. *The Journal of Biological Chemistry*, 289(21) :14881–14895, 2014.

[77] O.T. Avery, C.M. Macleod, and M. McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type Iii. *J Exp Med.*, 79 :137–158, 1944.

[78] Jones P. A. Functions of DNA methylation : islands, start sites, gene bodies and beyond. *Nature Review Genetics*, 13(7) :484–492, 2012.

[79] L. D Moore, T. Le, and G. Fan. DNA methylation and its basic function. *Neuropsychopharmacology*, 38(1) :23–38, 2013.

[80] M. Curradi, A. Izzo, G. Badaracco, and N. Landsberger. Molecular mechanisms of gene silencing mediated by DNA methylation. *Molecular and Cellular Biology*, 22(9) :3157–3173, 2002.

[81] L. de la Rica, J. Rodríguez-Ubreva, M García, A. BMMK Islam, J. M. Urquiza, H. Hernando, J. Christensen, K. Helin, C. Gómez-Vaquero, and E. Ballestar. PU.1 target genes undergo TET2-coupled demethylation and DNMT3b-mediated methylation in monocyte-to-osteoclast differentiation. *Genome Biology*, 14, 2014.

[82] M. Suzuki, T. Yamada, F. Kihara-Negishi, T. Sakurai, E. Hara, D. G. Tenen, N. Hozumi, and T. Oikawa. Site-specific DNA methylation by a complex of PU.1 and Dnmt3a/b. *Oncogene*, 25(17) :2477–88, Apr 2006.

[83] C. Lio, J. Zhang, E. González-Avalos, P. G. Hogan, X. Chang, and A. Rao. TET2 and TET3 cooperate with B-lineage transcription factors to regulate DNA modification and chromatin accessibility. *eLife*, 5 :e18290, nov 2016.

[84] M. D. Young, T. A. Willson, M. J. Wakefield, E. Trounson, D. J. Hilton, M. E. Blewitt, A. Oshlack, and I. J. Majewski. ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Research*, 39(17) :7415–7427, Jun 2011.

[85] G. Deb, V. S. Thakur, and S. Gupta. Multifaceted role of EZH2 in breast and prostate tumorigenesis. *Epigenetics*, 8(5) :464–476, 2013.

[86] F. Kihara-Negishi, H. Yamamoto, M. Suzuki, T. Yamada, T. Sakurai, T. Tamura, and T. Oikawa. In vivo complex formation of PU.1 with HDAC1 associated with PU.1-mediated transcriptional repression. *Oncogene*, 20 :6039–6047, 2001.

[87] M. Suzuki, T. Yamada, F. Kihara-Negishi, T. Sakurai, and T. Oikawa. Direct association between PU.1 and MeCP2 that recruits mSin3a-HDAC complex for PU.1-mediated transcriptional repression. *Oncogene*, 22(54) :8688–8698, Nov 2003.

[88] H. M. Chan and N. B. La Thangue. p300/CBP proteins : HATs for transcriptional bridges and scaffolds. *Journal of Cell Science*, 114(13) :2363–2373, 2001.

[89] H. Yamamoto, F. Kihara-Negishi, T. Yamada, Y. Hashimoto, and T. Oikawa. Physical and functional interactions between the transcription factor PU.1 and the coactivator CBP. *Oncogene*, 18 :1495–1501, 1999.

[90] W. Hong, A. Y. Kim, S. Ky, C. Rakowski, S.-B. Seo, D. Chakravarti, M. Atchison, and G. A. Blobel. Inhibition of CBP-mediated protein acetylation by the ETS family oncoprotein PU.1. *Mol Cell Biol*, 22(11) :3729–43, Jun 2002.

[91] H.-L. Hung, J. Lau, A. Y. Kim, M. J. Weiss, and G. A. Blobel. CREB-binding protein acetylates hematopoietic transcription factor GATA-1 at functionally important sites. *Molecular and Cellular Biology*, 19(5) :3496–3505, May 1999.

[92] T. Stopka, D. F. Amanatullah, M. Papetti, and A. I. Skoultchi. PU.1 inhibits the erythroid program by binding to GATA-1 on DNA and creating a repressive chromatin structure. *EMBO J*, 24(21) :3712–23, Nov 2005.

[93] N. Rekhtman, K. S. Choe, I. Matushansky, S. Murray, T. Stopka, and A. I. Skoultchi. PU.1 and pRB interact and cooperate to repress GATA-1 and block erythroid differentiation. *Mol Cell Biol*, 23(21) :7460–74, Nov 2003.

[94] A. Heydemann, G. Juang, K. Hennessy, M.S. Parmacek, and M. C. Simon. The myeloid-cell-specific *c-fes* promoter is regulated by Sp1, PU.1, and a novel transcription factor. *Mol Cell Biol*, 16 :1676–1686, 1996.

[95] H.M. Chen, H.L. Pahl, R.J. Scheibe, D.E. Zhang, and D. G. Tenen. The Sp1 transcription factor binds the *cd11b* promoter specifically in myeloid cells in vivo and is essential for myeloid-specific promoter activity. *J. Biol. Chem.*, 268 :8230–8239, 1993.

[96] L. T. Smith, S. Hohaus, D. A. Gonzalez, S. E. Dziennis, and D. G. Tenen. C/EBP alpha regulate the granulocyte colony-stimulating factor receptor promoter in myeloid cells. *Blood*, 88 :1234–1247, 1996.

[97] S. Hohaus, M. S. Petrovick, M. T. Voso, Z. Sun, D. E. Zhang, and D. G. Tenen. PU.1 (spi-1) and C/EBP alpha regulate expression of the granulocyte-macrophage colony-stimulating factor receptor alpha gene. *Cell. Biol.*, 15 :5830–5845, 1995.

[98] J. M. Perkel and M. L. Atchison. A two-step mechanism for recruitment of pip by PU.1. *The Journal of Immunology*, 160 :241–252, 1998.

[99] A. Himmelmann, A. Riva, G. L. Wilson, B. P. Lucas, C. Thevenin, and J. H. Kehrl. PU.1/Pip and basic helix loop helix zipper transcription factors interact with binding sites in the CD20 promoter to help confer lineage- and stage-specific expression of CD20 in B lymphocytes. *Blood*, 90 :3984–3995, 1997.

[100] G. Behre, A. J. Whitmarsh, M. P. Coghlan, T. Hoang, C. L. Carpenter, D. E. Zhang, R. J. Davis, and D. G. Tenen. c-Jun is a JNK-independent coactivator of the PU.1 transcription factor. *J. Biol. Chem.*, (4939-4946), 1999.

[101] T.-H. Pham, J. Minderjahn, C. Schmidl, H. Hoffmeister, S. Schmidhofer, W. Chen, G. Längst, C. Benner, and M. Rehli. Mechanisms of in vivo binding site selection of the hematopoietic master transcription factor PU.1. *Nucleic Acids Research*, 41(13) :6391–6402, May 2013.

[102] D. G. Tenen, R. Hromas, J. D. Licht, and D.-E. Zhang. Transcription factors, normal myeloid development, and leukemia. *Transcription and myeloid development*, 1997.

[103] G. Natoli. Maintaining cell identity through global control of genomic organization. *Immunity*, 33(1) :12–24, Jul 2010.

[104] M. Yaneva, S. Kippenberger, N. Wang, Q. Su, M. McGarvey, A. Nazarian, L. Lacomis, H. Erdjument-Bromage, and P. Tempst. PU.1 and a TTTAAA element in the myeloid defensin-1 promoter create an operational TATA box that can impose cell specificity onto TFIID function. *The Journal of Immunology*, 176(11) :6906–6917, May 2006.

[105] S. N. Wontakal, X. Guo, B. Will, M. Shi, D. Raha, M. C. Mahajan, S. Weissman, M. Snyder, U. Steidl, and D. Zheng. A large gene network in immature erythroid cells is controlled by the myeloid and B cell transcriptional regulator PU.1. *PLoS Genetics*, 7(6) :e1001392, Jun 2011.

[106] J. Ungerbäck, H. Hosokawa, X. Wang, T. Strid, B. A. Williams, M. Sigvardsson, and E. V. Rothenberg. Pioneering, chromatin remodeling, and epigenetic constraint in early T-cell gene regulation by SPI1 (PU.1). *Genome Res*, 28(10) :1508–1519, 10 2018.

[107] H. Hosokawa, J. Ungerbäck, X. Wang, M. Matsumoto, K. I. Nakayama, S. M. Cohen, T. Tanaka, and E. V. Rothenberg. Transcription factor PU.1 represses and activates gene expression in early T cells by redirecting partner transcription factor binding. *Immunity*, 48 :1119–1134, 2018.

[108] J. Fischer, C. Walter, A. Tönges, H. Aleth, M. J. Costa Jordão, M. Leddin, V. Gröning, T. Erdmann, G. Lenz, J. Roth, T. Vogl, M. Prinz, M. Dugas, I. D. Jacobsen, and F. Rosenbauer. Safeguard function of PU.1 shapes the inflammatory epigenome of neutrophils. *Nature Immunology*, 20 :546–558, 2019.

[109] M. Hallier, A. Tavitian, and F. Moreau-Gachelin. The Transcription Factor Spi-1/PU.1 Binds RNA and Interferes with the RNA-binding Protein p54nrb. *The Journal of Biological Chemistry*, 271(19) :11177–11181, 1996.

[110] M. Hallier, A. Lerga, S. Barnache, A. Tavitian, and F. Moreau-Gachelin. The Transcription Factor Spi-1/PU.1 interacts with the potential splicing factor TLS. *The Journal of Biological Chemistry*, 273(9) :4838–4842, 1998.

[111] C. Guillouf, I. Gallais, and F. Moreau-Gachelin. Spi-1/PU.1 oncoprotein affects splicing decisions in a promoter binding-dependent manner. *Journal of Biological Chemistry*, 281(28) :19145–19155, 2006.

[112] J. X. Cheng, L. Chen, Y. Li, A. Cloe, M. Yue, J. Wei, K. A. Watanabe, J. M. Shammo, J. Anastasi, and Q. J. Shen. RNA cytosine methylation and methyltransferases mediate

chromatin organization and 5-azacytidine response and resistance in leukaemia. *Nature Communications*, 9(1), Mar 2018.

[113] J. D. Watson and F. H. Crick. The structure of DNA. *Cold Spring Harbor symposia on quantitative biology*, 18 :123 – 131, 1953.

[114] F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3) :441 – 448, 1975.

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

[116] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1) :195 – 197, 1981.

[117] Łukasz Ligowski, Witold R. Rudnicki, Yongchao Liu, and Bertil Schmidt. Chapter 11 - accurate scanning of sequence databases with the smith-waterman algorithm. In Wen mei W. Hwu, editor, *GPU Computing Gems Emerald Edition*, Applications of GPU Computing Series, pages 155 – 171. Morgan Kaufmann, Boston, 2011.

[118] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3) :403 – 410, 1990.

[119] Human genome project, {https ://www.genome.gov/human-genome-project}.

[120] R. A. Bonnin, T. Naas, and L. Dortet. Impact du séquençage d'adn à haut débit sur la surveillance des épidémies de bactéries multi-résistantes aux antibiotiques. *Feuillets de biologie*, 334, 2017.

[121] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, (4) :357, 2012.

[122] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14) :1754 – 1760, 2009.

[123] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based analysis of Chip-Seq (macs). *Genome Biol*, 9(9) :R137, 2008.

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

[125] C. Herrmann, B. Van de Sande, D. Potier, and S. Aerts. i-cisTarget : an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. *Nucleic Acids Research*, 40(15) :e114–e114, Jun 2012.

[126] I. V. Kulakovskiy, V. A. Boeva, A. V. Favorov, and V. J. Makeev. Deep and wide digging for binding motifs in ChIP-seq data. *Bioinformatics*, 26(20) :2622–2623, Oct 2010.

[127] N. T. T. Nguyen, B. Contreras-Moreira, J. A. Castro-Mondragon, W. Santana-Garcia, R. Ossio, C. D. Robles-Espinoza, M. Bahin, S. Collombet, P. Vincens, and D. Thieffry. RSAT 2018 : regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Research*, 46(W1) :W209–W214, May 2018.

[128] A. Medina-Rivera, M. Defrance, O. Sand, C. Herrmann, J. A. Castro-Mondragon, J. Delerce, S. Jaeger, C. Blanchet, P. Vincens, and C. Caron. RSAT 2015 : Regulatory Sequence Analysis Tools. *Nucleic Acids Research*, 43(W1) :W50–W56, Apr 2015.

[129] M. Thomas-Chollier, M. Defrance, A. Medina-Rivera, O. Sand, C. Herrmann, D. Thieffry, and J. van Helden. RSAT 2011 : regulatory sequence analysis tools. *Nucleic Acids Research*, 39(suppl) :W86–W91, Jun 2011.

[130] M. Thomas-Chollier, O. Sand, J.-V. Turatsinze, R. Janky, M. Defrance, E. Vervisch, S. Brohee, and J. van Helden. Rsat : regulatory sequence analysis tools. *Nucleic Acids Research*, 36(Web Server) :W119–W127, May 2008.

[131] J. Van Helden. Regulatory Sequence Analysis Tools. *Nucleic Acids Research*, 31(13) :3593–3596, Jul 2003.

[132] T. L. Bailey, M. Bodén, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. MEME SUITE : tools for motif discovery and searching. *Nucleic Acids Research*, 37 :202–208, 2009.

[133] K. Chen, Z. Hu, Z. Xia, D. Zhao, W. Li, and J. K. Tyler. The Overlooked Fact : Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses. *Mol Cell Biol*, 36(5) :662–7, Dec 2016.

[134] F. Ramírez, D. P. Ryan, B. Grüning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dündar, and T. Manke. Deeptools2 : A next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 8 :160–165, 2016.

[135] K. Liang and S. Keles. Normalization of ChIP-seq data with control. *BMC Bioinformatics*, 13(199), 2012.

[136] H. Ji, H. Jiang, W. Ma, D. S. Johnson, R. M. Myers, and W. H. Wong. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nature Biotechnology*, 26 :1293–1300, 2008.

[137] C. Taslim, J. Wu, P. Yan, G. Singer, J. Parvin, T. Huang, S. Lin, and K. Huang. Comparative study on ChIP-seq data : normalization and binding pattern characterization. *Bioinformatics*, 25(18) :2334–2340, Jun 2009.

[138] N. U. Nair, A. D. Sahu, P. Bucher, and B. M. E. Moret. ChIPnorm : A Statistical Method for Normalizing and Identifying Differential Regions in Histone Modification ChIP-seq libraries. *PLoS ONE*, 7(8) :e39573, Aug 2012.

[139] H. Jin, L. H. Kasper, J. D. Larson, G. Wu, S. J. Baker, J. Zhang, and Y. Fan. ChIPseqSpikeInFree : a ChIP-seq normalization approach to reveal global changes in histone modifications without spike-in. *Bioinformatics*, 36(4) :1270–1272, 09 2019.

[140] V. Boeva, C. Louis-Brennetot, A. Peltier, S. Durand, C. Pierre-Eugène, V. Raynal, H. C. Etchevers, S. Thomas, A. Lermine, and E. Daudigeos-Dubus. Heterogeneity of neuroblastoma cell identity defined by transcriptional circuitries. *Nature Genetics*, 49(9) :1408–1413, Jul 2017.

[141] A. D. Ellington and J. W. Szostak. In vitro selection of RNA molecules that bind specific ligands. *Nature*, (6287) :818, 1990.

[142] J. Sauliere, V. Murigneux, Z. Wang, E. Marquenet, I. Barbosa, O. Le Tonqueze, Y. Audic, L. Paillard, H. R. Crollius, and H. Le Hir. Clip-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. *Nature Structural and Molecular Biology*, (11) :1124, 2012.

[143] I. Huppertz, J. Attig, A. D'Ambrogio, L. E. Easton, C. R. Sibley, Y. Sugimoto, M. Tajnik, J. König, and J. Ulea. iCLIP : Protein–RNA interactions at nucleotide resolution. *Methods*, 65(3) :274–287, 2014.

[144] E. .L Van Nostrand, G. A Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, and G. W. Yeo. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, 13(6) :508–514, 2016.

[145] J. Ule, K.B. Jensen, M. Ruggiu, A. Mele, A. Ule, and R.B. Darnell. CLIP identifies Nova-regulated RNA networks in the brain. *Science*, 14(302) :1212–1215, 2003.

[146] P. J. Uren, E. Bahrami-Samania, P. Rosa de Araujob, C. Vogeld, M. Qiaoc, S. C. Burnsc, A. D. Smith, and L. O. F. Penalva. High-throughput analyses of hnRNP H1 dissects its multi-functional aspect. *RNA Biology*, 13(4) :400–411, 2016.

[147] P. M. Clark, P. Loher, K. Quann, J. Brody, E. R. Londin, and I. Rigoutsos. Argonaute CLIP-Seq reveals miRNA targetome diversity across tissue types. *Scientific Reports*, 4(1), Aug 2014.

[148] J. Lin, Y. Zhang, W. N. Frankel, and Z. Ouyang. PRAS : Predicting functional targets of RNA binding proteins based on CLIP-seq peaks. *Plos computational biology*, 15(8), 2019.

[149] Eci-ishr, {https ://eciofishr.wordpress.com/2019/04/22/technical-section-atac-seq}.

[150] J. J. Reske, M. R. Wilson, and R. L. Chandler. ATAC-seq normalization method can significantly affect differential accessibility analysis and interpretation. *Epigenetics & Chromatin*, 13(22), 2020.

[151] E. Soler, C. Andrieu-Soler, E. de Boer, J. C. Bryne, S. Thongjuea, R. Stadhouders, R.-J. Palstra, M. Stevens, C. Kockx, W. van IJcken, J. Hou, C. Steinhoff, E. Rijkers, B. Lenhard, and F. Grosveld. The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. *Genes & Development*, 24 :277–289, 2009.

[152] S. B. Rothbart, B. M. Dickson, J. R. Raab, A. T. Grzybowski, K. Krajewski, A. H. Guo, E. K. Shanle, S. Z. Josefowicz, S. M. Fuchs, C. D. Allis, T. R. Magnuson, A. J. Ruthenburg, and B. D. Strahl. An interactive database for the assessment of histone antibody specificity. *Molecular Cell*, 59 :502–511, 2015.

[153] A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, J. Hillman-Jackson, R. M. Kuhn, J. S. Pedersen, A. Pohl, B. J. Raney, K. R. Rosenbloom, A. Siepel, K. E. Smith, C. W. Sugnet, A. Sultan-Qurraie, D. J. Thomas, H. Trumbower, R. J. Weber, M. Weirauch, A. S. Zweig, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database : update 2006. *Nucleic Acids Research*, 1(34), 2006.

[154] G. Yu, L.G. Wang, and Q.Y. He. ChIPseeker : an r/bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, 31(14) :2382–2383, 2015.

[155] M. Lawrence, W. Huber, H. Pagès, P. Aboyoun, M. Carlson, R. Gentleman, M. Morgan, and V. Carey. Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9, 2013.

[156] Gencode VM16, {http ://genome.ucsc.edu}.

[157] J. Ernst and M. Kellis. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc*, 12(12) :2478–2492, Dec 2017.

[158] X. Jiao, B. T. Sherman, D. W. Huang, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki. David-ws : a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, 28(13) :1805–1806, Apr 2012.

[159] S. N. Wontakal, X. Guo, C. Smith, T. MacCarthy, E. H. Bresnick, A. Bergman, M. P. Snyder, S. M. Weissman, D. Zheng, and A. I. Skoultchi. A core erythroid transcriptional

network is repressed by a master regulator of myelo-lymphoid differentiation. *Proceedings of the National Academy of Sciences*, 109(10) :3832–3837, 2012.

[160] M.R Tallack, T. Whitington, W. S. Yuen, E. N. Wainwright, J. R. Keys, B. B. Gardiner, E. Nourbakhsh, N. Cloonan, S. M. Grimmond, T. L. Bailey, and A. C. Perkins. A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Research*, 20(8) :1052–1063, 2010.

[161] M. T Kassouf, J. R Hughes, S. Taylor, S. J. McGowan, S. Soneji, A. L. Green, P. Vyas, and C. Porcher. Genome-wide identification of TAL1's functional targets : insights into its mechanisms of action in primary erythroid cells. *Genome Research*, 20(8) :1064–1083, 2010.

[162] Rapid immunoprecipitation mass spectrometry of endogenous proteins, {https ://www.activemotif.com/catalog/1077/rime}.

[163] P. Tessarz and T. Kouzarides. Histone core modifications regulating nucleosome structure and dynamics. *Nature Reviews Molecular Cell Biology*, 15 :703–708, 2014.

[164] M. Tagore, M. J. McAndrew, A. Gjidoda, and M. Floer. The lineage-specific transcription factor PU.1 prevents Polycomb-mediated heterochromatin formation at macrophage-specific genes. *Molecular and Cellular Biology*, 35(15) :2610–2625, 2015.

[165] Jaspar database, {http ://jaspar.genereg.net}.

[166] N. Rekhtman, F. Radparvar, T. Evans, and A. I. Skoultchi. Direct interaction of hematopoietic transcription factors PU.1 and GATA-1 : functional antagonism in erythroid cells. *Genes & development*, 13 :1398–1411, 1999.

[167] P. Zhang, X. Zhang, A. Iwama, C. Yu, K. A. Smith, B. U. Mueller, S. Narravula, B. E. Torbett, S. H. Orkin, and D. G. Tenen. PU.1 inhibits GATA-1 function and erythroid differentiation by blocking GATA-1 DNA binding. *Blood*, 96(8) :2641–2648, 10 2000.

[168] J. Boyes, P. Byfield, Y. Nakatani, and V. Ogryzko. Regulation of activity of the transcription factor GATA-1 by acetylation-1 by acetylation. *Nature*, 396(6711) :594–598, 1998.

[169] D. Holoch and R. Margueron. Mechanisms regulating PRC2 recruitment and enzymatic activity. *Trends Biochem Science*, 42(7) :531–542, 2017.

[170] M. Beltran, C. M. Yates, L. Skalska, M. Dawson, F. P. Reis, K. Viiri, C. L. Fisher, C. R. Sibley, B. M. Foster, T. Bartke, J. Ule, and R. G. Jenner. The interaction of PRC2 with RNA or chromatin is mutually antagonistic. *Genome Research*, 26(7) :896–907, 2016.

[171] X. Wang and C. Davidovich. Targeting PRC2 : RNA offers new opportunities. *Oncotarget*, 8(64) :107346–107347, 2017.

[172] B. Yadav, K. Wennerberg, T. Aittokallio, and J. Tang. Searching for drug synergy in complex dose-response landscapes using an interaction potency model. *Computational and Structural Biotechnology Journal*, 25(13) :504–513, 2015.

[173] V. Murigneux, J. Saulière, H. Roest Crollius, and H. Le Hir. Transcriptome-wide identification of RNA binding sites by CLIP-seq. *Methods*, 63(1) :32 – 40, 2013. Diversity of the non-coding transcriptomes revealed by RNA-seq technologies.

[174] Novocraft, {www.novocraft.com/products/novoalign}.

[175] Useq, {www.useq.sourceforge.net}.

[176] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3) :443 – 453, 1970.

[177] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16) :2078–2079, 06 2009.

[178] S. Anders, P. T. Pyl, and W. Huber. Htseq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2) :166–169, Sep 2014.

[179] P. Drewe-Boss, H.-H. Wessels, and U. Ohler. omniCLIP : probabilistic identification of protein-RNA interactions from CLIP-seq data. *Genome Biology*, 19(183), 2018.

[180] P. J. Uren, E. Bahrami-Samani, S. C. Burns, M. Qiao, F. V. Karginov, E. Hodges, G. J. Hannon, J. R. Sanford, L. O. F. Penalva, and A. D. Smith. Site identification in high-throughput RNA–protein interaction data. *Bioinformatics*, 28(23) :3013–3020, Sep 2012.

[181] E. Bahrami-Samani, L. O.F. Penalva, A. D. Smith, and P. J. Uren. Leveraging cross-link modification events in CLIP-seq for motif discovery. *Nucleic Acids Research*, 43(1) :95–103, Dec 2014.

[182] C. Zhang and R. B Darnell. Mapping in vivo protein-rna interactions at single-nucleotide resolution from HITS-CLIP data. *Nat Biotechnol*, 29(7) :607–14, Jun 2011.

[183] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat : discovering splice junctions with rna-seq. *Bioinformatics*, 25(9) :1105–1111, 2009.

[184] H. Ashoor, C. Louis-Brennetot, I. Janoueix-Lerosey, V. B. Bajic, and V. Boeva. HMCan-diff : a method to detect changes in histone modifications in cells with different genetic characteristics. *Nucleic Acids Research*, page gkw1319, Jan 2017.

[185] I. Krivega, R. K. Dale, and A. Dean. Role of LDB1 in the transition from chromatin looping to transcription activation. *Genes & Development*, 28 :1278–1290, 2014.

[186] X. Guo, J. Plank-Bazinet, I. Krivega, R. K. Dale, and A. Dean. Embryonic erythropoiesis and hemoglobin switching require transcriptional repressor ETO2 to modulate chromatin organization. *Nucleic Acids Res*, 48(18) :10226–10240, Oct 2020.

[187] J. M. Amann, J. Nip, D. K. Strom, B. Lutterbach, H. Harada, N. Lenny, J. R. Downing, S. Meyers, and S. W. Hiebert. ETO, a Target of t(8 ;21) in Acute Leukemia, Makes Distinct Contacts with Multiple Histone Deacetylases and Binds mSin3A through Its Oligomerization Domain. *Molecular and Cellular Biology*, 21(19) :6470–6483, 2001.

[188] P. E Love, C. Warzecha, and L. Li. Ldb1 complexes : the new master regulators of erythroid gene transcription. *Trends Genet*, 30(1) :1–9, 2014.

[189] A-S. Tournillon, I. López, L. Malbert-Colas, S. Findakly, N. Naski, V. Olivares-Illana, K. Karakostis, B. Vojtesek, K. Nylander, and R. Fåhraeus. p53 binds the mdmx mRNA and controls its translation. *Nature*, 36 :723–730, 2016.

[190] Z. E. Holmes, D. J. Hamilton, T. Hwang, N. V. Parsonnet, J. L. Rinn, D. S. Wuttke, and R. T. Batey. The Sox2 transcription factor binds RNA. *Nature Communications*, 11(1805), 2020.

[191] Z. R. Liu, A. M. Wilkie, M. J. Clemens, and C. W. Smith. Detection of double-stranded RNA-protein interactions by methylene blue-mediated photo-crosslinking. *RNA*, 2(6) :611–621, 1996.

[192] Z.-R. Liu, B. Sargueil, and C. W.J. Smith. Methylene blue-mediated cross-linking of proteins to double-stranded RNA. *Methods in enzymology*, 318(2000) :22–33, 2000.

[193] A. Dobin, C. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. Gingeras. STAR : ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1) :15–21, 2013.

[194] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10) :R106, 2010.

[195] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate : A practical and powerful approachto multiple testing. *J R Statist Soc B*, 57(1) :289–300, 1995.

[196] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, (12) :1213, 2013.

[197] M R. Corces, A. E. Trevino, E. G. Hamilton, P. G. Greenside, N. A. Sinnott-Armstrong, S. Vesuna, A. T. Satpathy, A. J. Rubin, K. S. Montine, B. Wu, A. Kathiria, S. W. Cho, M. R. Mumbach, A. C. Carter, M. Kasowski, L. A. Orloff, V. I. Risca, A. Kundaje,

P. A. Khavari, T. J. Montine, W. J. Greenleaf, and H. Y. Chang. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods : Techniques for life scientists and chemists*, 14(10) :959, 2017.

[198] A. I. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem.*, 75(17) :4646–4658, 2003.

[199] Format sam ; {http ://samtools.github.io/hts-specs/SAMv1.pdf}.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre: Régulation de l'expression génique par le facteur de transcription SPI1/PU.1 dans l'érythroleucémie : mécanismes de répression des gènes par sa liaison à l'ADN, conséquences de sa liaison à l'ARN.

Mots clés: Transcription, Épigénétique, Régulation génique, ChIP-seq, CLIP-seq

Résumé: Dans le lignage érythroïde, une expression anormale et non-contrôlée du facteur de transcription SPI1 entraîne une leucémie aiguë, en partie en inhibant la différenciation érythroïde et l'apoptose des progéniteurs engagés dans la différenciation érythroïde. Mon travail de thèse concerne la caractérisation des mécanismes par lesquels SPI1 réprime l'expression génique dans un modèle d'érythroleucémie murine. En utilisant des cellules pré-leucémiques issues de souris transgéniques pour spi1, dans lesquelles l'expression de spi1 peut être contrôlée, j'ai comparé des données de séquençage à haut débit pour caractériser l'accessibilité de la chromatine, les modifications épigénétiques des protéines histones et l'expression des gènes en fonction de la présence ou de l'absence de SPI1. Pour comparer les signaux de ChIP-seq entre différentes conditions, nous avons développé un package R : ChIP-seq Intersample Normalization (CHIPIN). Nous avons ainsi démontré que la répression des gènes par SPI1, dont certains sont liés à la différenciation érythroïde et à l'apoptose, est basée sur la coordination de deux mécanismes qui impliquent et sont contrôlés par l'histone dé-acétylase 1 (HDAC1) et le complexe répressif Polycomb (PRC2). La caractérisation de la fixation de SPI1 à l'ARN en utilisant des données de CLIP-seq nous a permis de montrer que cette fixation n'était pas liée à la régulation de l'expression génique. Nous proposons un nouveau mécanisme pour la répression de l'expression génique par SPI1 dans l'érythroleucémie en coopération avec deux facteurs épigénétiques : PRC2 et HDAC1, et un nouveau package R pour la normalisation des données de ChIP-seq.

Title: Regulation of gene expression by the transcription factor SPI1/PU.1 in erythroleukemia: mechanisms of repression of genes by its binding to DNA, consequences of its binding to RNA.

Keywords: Transcription, Epigenetics, Gene regulation, ChIP-seq, CLIP-seq

Abstract: In the erythroid lineage, abnormal and uncontrolled expression of the transcription factor SPI1 leads to acute leukaemia, in part by inhibiting erythroid differentiation and apoptosis of progenitors involved in erythroid differentiation. My PhD thesis is dedicated to the characterization of the mechanisms by which SPI1 represses gene expression in a mouse erythroleukemia model. Using pre-leukemic cells from spi1 transgenic mice, in which spi1 expression can be controlled, I compared high-throughput sequencing data to characterise chromatin accessibility, epigenetic modifications of histone proteins and gene expression as a function of the presence or absence of SPI1. To compare ChIP-seq signals between different conditions, we developed an R package: ChIP-seq Intersample Normalization (CHIPIN). We demonstrated that gene repression by SPI1, including genes coding for apoptosis and erythroid differentiation, is based on the coordination of two mechanisms which involve and are controlled by histone deacetylase 1 (HDAC1) and the Polycomb repressive complex (PRC2). Characterisation of the binding of SPI1 to RNA using CLIP-seq data allowed us to show that this binding was not linked to regulation of gene expression. We propose a new mechanism for the repression of gene expression by SPI1 in erythroleukemia in cooperation with two epigenetic factors: PRC2 and HDAC1, and a new R package for the normalization of ChIP-seq data.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France