Université de Nantes

UFR Sciences et Techniques

Analyse de la fixation des sous-unités de la voie NF-κB dans une lignée cellulaire de myélome multiple

INSERM U892, Equipe 11 - LINA

Encadrants :
Florence MAGRANGEAS, UMR892 Equipe 11 - UMGC
Stéphane MINVIELLE, UMR892 Equipe 11 - UMGC
Jérémie BOURDON, LINA Université de Nantes
Etudiant : Victor Gaborit

Master 2 Bioinformatique

July 20, 2016

Résumé

La voie NF-κB a été décrite comme étant importante dans la défense de l'organisme suite à une infection. Elle régule également le développement, la différentiation, la prolifération et la survie des cellules. Cette voie est activée de façon constitutive chez des patients atteints de myélome multiple [1]. Le but de cette étude est de caractériser l'activité des cinq sous-unités NF-κB (RelA, RelB, cRel, p50 et p52) dans la cellule de myélome afin de comprendre et de cibler les différents gènes et fonctions biologiques impactées par l'activation canonique et non-canonique de cette voie. Une autre partie de l'étude sera de rechercher les différents co-facteurs qui peuvent intervenir dans l'activation ou la répression de la transcription par cette voie.

Pour permettre la compréhension de ce mécanisme, nous avons réalisé un séquençage ChIP-seq sur chacune des sous-unités de la NF-κB ainsi que sur le facteur oncogénique forkhead box FOXM1 qui a été décrit comme étant associé avec de nombreux sites κB dans le lymphome [3].

Ce rapport présente la mise en place d'un pipeline d'analyse pour l'étude de données de ChIP-seq pour des facteurs de transcriptions ainsi que les résultats obtenus pour les données de la voie NF-κB. Certaines de ces étapes ont été automatisée afin de permettre l'analyse générale de n'importe quel facteur de transcription: alignement, recherche de pics, annotation fonctionnelle des pics. D'autres étapes ne peuvent être automatisée mais sont essentielles pour la réalisation de cette étude (annotation fonctionnelle du génome, associations des sites de fixation à des processus biologiques). Enfin, des analyses de co-binding sont utiles dans le cas de l'étude de l'association de plusieurs facteurs de transcription comme ceux de la voie NF-κB.

L'ensemble du code utilisé pour mettre en place le pipeline ainsi que le manuel d'utilisation est disponnible sur le github: `https://github.com/victor2410/ChIPpipe`

Mots-clés: Voie NF-κB, ChIP-seq, Cistrome, Découverte de motifs, Annotation Fonctionnelle, Co-Binding.

Abstract

NF-κB pathway have been described as a major part of body's defense against infection and also mediate development, differentiation, proliferation and survival. Some studies have shown that NF-κB pathway is constitutivelly activated in multiple myeloma cells.
The aim of this study is characterizing activity of NF-κB subunits (RelA, RelB, cRel, p50 and p52) in myeloma cell to understand and identify genes and biological pathways that are impacted by canonical and non-canonical activation of NF-κB signaling pathway. In other hand, we will try to study potential co-factors which can be associated with NF-κB subunits for activate or repress transcription mediated by this pathway.
To allow understanding of that mechanism, we performed chromatin immunoprecipitation followed by high throughput DNA-sequencing (ChIP-seq) on each NF-κB subunits and on oncogenic forkhead box FOXM1 which have been as co-occupying κB sites in B lymphoma. ChIP-seq have been performed for cRel subunits but no signal was obtained for this factor during sequencing, inducing the hypothesis that this factor is not present in nucleus in myeloma cell. This hypothesis will be further explore by performing Western Blot analysis on cRel subunit in nucleus.
This report present the establishing of analysis pipeline to study ChIP-seq datas for transcription factor and the results we get for NF-κ ChIP-seq analysis.

Keywords : NF-κB pathway, ChIP-seq, Cistrome, Motif Discovery, Functionnal Annotation, Co-Binding
A compléter

Sommaire

1 Contexte Biologique

1.1 Le myélome multiple

Le myélome multiple est une hémopathie maligne qui se caractérise par la prolifération de cellules plasmocytaires à l'intérieur de la moelle osseuse. Malgré des nouveaux traitements qui fournissent une meilleure réponse et un allongement de la survie du patient, cette maladie reste aujourd'hui incurable. De plus, la transformation des plasmocytes en clones malins, ainsi que les différents mécanismes responsables de la rechute du patient vis à vis de la maladie ne sont pas encore totalement connus.
Certaines anomalies génétiques ont été décrites dans le cadre de cette pathologie [2]: des translocations chromosomiques peuvent générer des recombinaisons aberrantes avec le locus des immunoglobulines qui peuvent placer des oncogènes sous le contrôle d'enhancers actifs ce qui conduit à une augmentation de leur expression (exemple: gènes CCND1, FGFR3, MMSET, c-maf, c-myc...). Certaines mutations génétiques dans différentes voies contribuent à déréguler la biologie intrinsèque de la cellule plasmocytaire et peut lui donner de nouvelles capacités comme la résistance au traitement ou la faculté de se développer dans le sang périphérique.
Différents stades ont été décrits dans le cadre de cette pathologie, la cellule plasmocytaire normale, le stade MGUS (Gammapathie monoclonale de signification indéterminée) qui représente une augmentation du nombre de plasmocytes dans la moelle osseuse, le stade SMM (Myélome multiple indolent) qui correspond à une plus forte augmentation des plasmocytes sans apparition de signes cliniques. Il y a ensuite le stade myélome avéré (MM) correspond à une apparition des signes cliniques du myélome.
Dans la moelle osseuse, la cellule plasmocytaire normale à besoin des signaux de son micro-environnement pour survivre. Au cours de son évolution, la cellule de myélome va passer par ces différents stades pour aboutir à une forme extra-médullaire appelée PCL (Plasma Cell Leukemia) où la cellule deviendra indépendante de son environnement par l'acquisition d'altérations génétiques, en particulier sur la voie NF-κB (Figure 1).

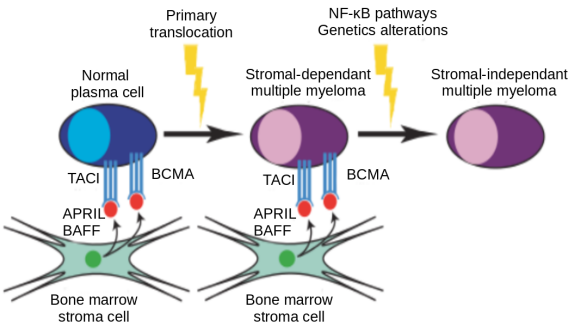


Figure 1: Évolution du plasmocyte en clone malin pour devenir indépendant du micro-environnement de la moelle osseuse (Figure adaptée d'après Staudt et al. [1])

1.2 La voie NF-κB

Cette voie est activée de façon normale au cours de la réponse immunitaire et dans la réponse au stress cellulaire. Cinq sous-unités de ce facteur ont été décrites dans cette voie chez les mammifères : RelA (p65),

RelB, cRel, p50 (p105) et p52 (p100). Chacune de ces sous-unités possède un domaine Rel-homologue en N-terminal qui va permettre la fixation spécifique sur les sites κB de l'ADN. De plus, RelA, RelB et cRel ont également un domaine d'activation de la transcription du coté C-terminal. Cela indique que seules les sous-unités RelA, RelB et cRel peuvent jouer un rôle en tant qu'activateur de transcription. Pour pouvoir jouer ce rôle, les sous-unités p50 et p52 doivent donc interagir (hétéro-dimérisation) avec d'autres facteurs activateurs/répresseurs de la transcription NF-κB, ou avec d'autres co-facteurs possédant un domaine activateur. Un hétéro-dimère p50:p52 sera donc généralement associé à une répression de la transcription s'il se fixe sans co-facteur.

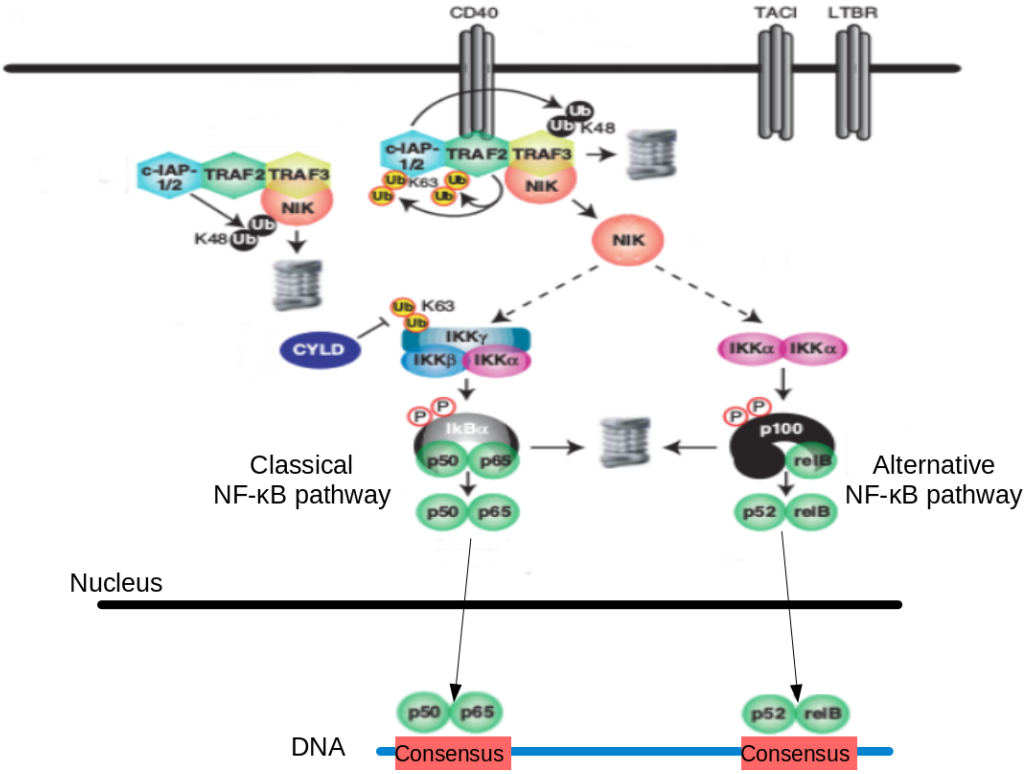


Figure 2: Activation de la voie NF-κB (Figure adaptée d'après Staudt et al. [1])

L'activation de cette voie va conduire à l'export d'homo et d'hétéro-dimères vers le noyau afin de pouvoir se fixer sur l'ADN et ainsi réguler la transcription de gènes (Figure 2). Il existe deux mécanismes activateurs de la voie NF-κB. Le premier mécanisme, dit voie canonique, est une réponse à des signaux pro-inflammatoires qui va induire la dégradation d'un inhibiteur de NF-κB (IκB). C'est un mécanisme essentiel pour la rapidité de la réponse immune et qui va permettre la translocation dans le noyau des dimères suivants : RelA:p50, cRel:p50, RelA:RelA, cRel:cRel. Pour la seconde voie, dite non canonique, qui permet l'organogénèse des organes lymphoïdes secondaires, le développement des cellules B et leur survie, ce sont les dimères RelB:p52, p52:p52 et p50:p52 qui seront exportés vers le noyau. Enfin une activation de ces deux mécanismes simultanément peut également conduire à la formation de dimères hybrides résultants de cette activation simultanée comme RelA:p52. En tout, 14 dimères NF-κB différents ont été observés lorsque les deux mécanismes sont actifs dans le lymphome [3].
Dans le cadre du myélome, des anomalies sur les régulateurs négatifs de la protéine NIK (Figure 2):

TRAF3,TRAF2 et ciap ou un haut niveau d'expression du récepteur membranaire CD40 auront pour conséquence une activation constitutive de ces deux voies canoniques et non-canoniques de NF-κB. Ces mutations vont également permettre à la cellule de s'affranchir de son micro-environnement et de rejoindre le sang périphérique.
De plus, une sur-expression du facteur oncogénique FOXM1 est généralement observée chez les patients atteints de myélome multiple de mauvais pronostic. Ce facteur semble jouer un rôle important au niveau de NF-κB. En effet FOXM1 est présent sur la moitié des sites NF-κB dans la lignée cellulaire GM12878 (EBV transformed lymphoblastoid B cell line) et est recruté sur des complexes NF-κB fixés sur l'ADN [3].
Ainsi, le but de cette étude est d'analyser le cistrome régulé par la voie NF-κB dans une cellule de myélome multiple. Il s'agira également d'étudier l'association des sous-unités de la voie NF-κB avec le facteur oncogénique FOXM1.
De nombreuses études ont été réalisées dans le myélome multiple afin de mieux comprendre les voies cytosoliques qui activent les facteurs de la voie NF-κB afin de trouver de nouvelles cibles thérapeutiques. Paradoxalement, peu de choses sont connues sur les mécanismes qui gouvernent la fonction nucléaire de ces facteurs. Dans ce stage, je me suis intéressé à l'analyse de la fixation des facteurs de la voies NF-κB dans le noyau de la cellule de myélome et aux différentes cibles impactées par ces facteurs (cistrome).

2 Données disponibles

J'ai utilisé des données issues de la technique de Chromatine ImmunoPrecipitation Sequencing (Chip-Seq) pour les sous-unités RelA, RelB, p50, p52 et FOXM1. Seules les données de la sous unité cRel n'ont pas pu être analysées car le séquençage n'a pas fourni un signal suffisament fort pour ce facteur. Un Western-blot sera réalisé plus tard afin de vérifier la présence de cRel dans le noyau pour la lignée cellulaire MM1S.

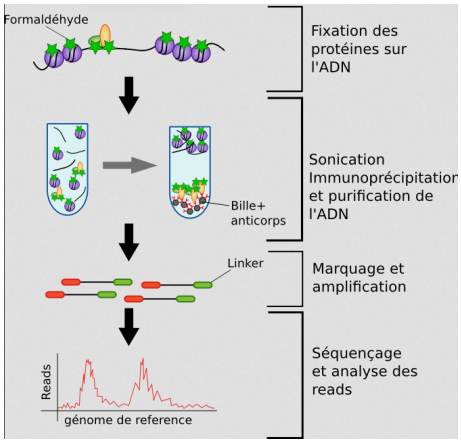


Figure 3: Protocole de séquençage par la technique du ChIP-Seq pour étudier la fixation des facteurs de transcription sur l'ADN (Figure adaptée d'après le site http://bioinfo-fr.net)

Ces données ont été obtenues dans la lignée cellulaire MM1S qui est issue d'une cellule de patient atteint de myélome multiple au stade PCL et qui est mutée sur le gène *TRAF3*. Le séquençage a été réalisé sans stimulation des cellules pour se rapprocher des conditions in vivo de ces cellules qui sont indépendantes de l'environnement, ce qui explique que l'on obtienne moins de signal que pour les données de disponibles

dans la littérature pour d'autres lignées cellulaires [3], le but étant de comprendre le fonctionnement de la cellule dans son état "normal". Nous disposons également d'un fichier fastq qui nous servira de contrôle correspondant à l'Input (Séquençage sans ImmunoPrécipitation).
Pour réaliser cette technique (Figure 3), il faut tout d'abord sélectionner une cible au Chip-Seq (généralement, un facteur de transcription). Les protéines liées à l'ADN sont fixées (généralement avec du formaldéhyde). Une fois toutes les liaisons bloquées, l'ADN va être découpé par une enzyme ou par sonication.
Un anticorps spécifique du facteur de transcription est ensuite ajouté pour permettre l'immuno-précipitation (IP), cela correspond à ne retenir que les fragments d'ADN liés à la protéine d'intérêt. Puis, l'ADN va être séparé du facteur de transcription et il va y avoir amplification de ces fragments d'ADN par PCR en vue de les séquencer.
Pour le séquençage, n'importe quelle technique issue du NGS peut être utilisée mais la plus courante reste Illumina. Lors de la réalisation de cette technique, il est important de faire des contrôles en même temps, deux sont généralement effectués: un premier que l'on nomme Input. Ce contrôle correspond à la même manipulation que celle réalisée pour le ChIP-seq mais sans immuno-précipitation. Cette technique permet de vérifier la sélectivité de l'anticorps utilisé en s'assurant que les fragments d'ADN séquencés ne sont pas sélectionnés aléatoirement.Ce contrôle est utilisé le plus fréquemment pour l'analyse de ce type de données. Le second contrôle est un ChIP-seq mais avec pour lequel on utilise un anticorps différents de celui du facteur de transcription: l'anticorps anti-IgG. Ce contrôle permet de s'assurer de la spécificité de l'anticorps utilisé.
La technique de ChIP-Seq permet de récupérer les protéines spécifiques de l'anticorps utilisé qui sont liées à l'ADN. En revanche, cette technique ne permet pas de donner plus d'information sur le type de la liaison, c'est à dire si le facteur est présent sur l'ADN directement ou bien si il est fixé via un co-facteur. L'analyse des sites de fixation et notamment la recherche de motifs consensus est donc une étape clé de l'analyse afin de comprendre comment le facteur étudié se fixe à l'ADN dans une condition ou une région particulière. Toutes les données ont été obtenues par ChIP-Seq sur la MM1S et sont des reads single-end de 75 nucléotides. Les données étant faites par un organisme extérieur qui est spécialisé dans le séquençage ChIP-Seq, il n'y a pas de réplicats biologiques pour les différents facteurs de transcription.

3 Pipeline général d'analyse des données

Au cours de mon travail, j'ai mis en place un pipeline d'analyse de données de ChIP-seq qui se détaille en sept étapes (Figure 4). Les étapes 1 à 4 ont pu être automatisées sous forme de sous programmes (implémentés en python) et réunis comme des fonctionnalités d'un seul programme : *ChIPpipe*. Ce pipeline permet de filtrer et d'effectuer les contrôles qualité des données à la sortie du séquenceur (les résultats de cette étape ne sont pas présentés dans ce rapport car ils ont été fournis par l'organisme produisant les données), puis d'aligner chacun des reads contre un génome de référence choisi.
Le principal aspect de ce pipeline est de permettre la recherche de pics basée sur le protocole mis en place par le consortium ENCODE [4] lorsque l'on dispose de réplicats biologiques ou bien d'effectuer une recherche de pics sans réplicats biologiques en étant le plus sélectif possible. Enfin il permet également d'effectuer une annotation fonctionnelle des pics à partir d'un fichier d'annotation fonctionnelle spécifique à la lignée cellulaire pour laquelle le séquençage a été réalisé.
Les étapes 5 à 7 ne peuvent pas être automatisées car elles dépendent d'applications et d'interfaces web mais sont présentées ici car elles font partie de la procédure mise en place pour analyser les données. Il s'agit de la recherche de profils de fixation. Cette étape est réalisée lorsque l'on étudie plusieurs facteurs de transcription comme c'est le cas ici. De la recherche de motifs dans les régions étudiées et enfin de

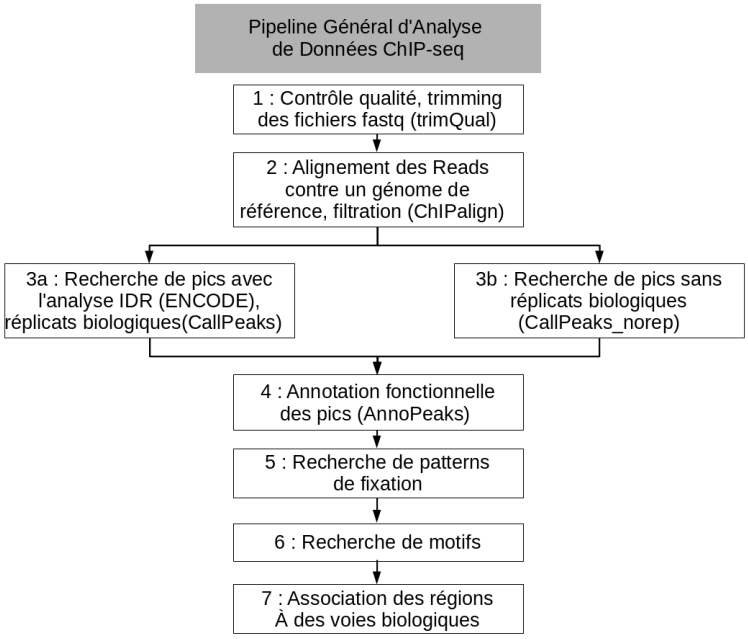


Figure 4: Présentation du pipeline général mis en place pour analyser les données issues de ChIP-seq pour des facteurs de transcription

l'étude de l'association de ces régions avec des processus biologiques connus.

4 Alignement des reads contre le génome de référence

Pour réaliser l'alignement des fichiers fastq, j'ai utilisé un pipeline d'automatisation mis en place durant mon stage : *ChIPalign* (Figure 5) qui va me permettre de réaliser l'alignement des reads avec le logiciel *Bowtie2* (version 2.1.0) et de créer les fichiers d'alignement (fichiers Bam). En plus de réaliser l'alignement des reads contre le génome de référence, ce pipeline permet également d'indexer le génome de référence (nécessaire pour l'alignement) mais surtout de filtrer le fichier d'alignement en sortie (ces étapes de filtrations sont optionnelles dans le pipeline et peuvent être paramétrées dans la commande):

- Filtrer des reads non mappés avec *Samtools* (Version 1.3.1)
- Filtrer des reads ayant une qualité de mapping inférieure à un seuil donné avec *Samtools* (Ici, le seuil utilisé est de 30).
- Filtrer les reads situés dans les coordonnées spécifiées dans un fichier texte (Pour cette analyse, le fichier contient les régions "blacklist" d'ENCODE incluant les régions répétées télomériques et centromériques, les satellites et les îlots à haute mappabilité) avec *Samtools*.
- Enlever les reads correspondant à des duplicats de PCR ou des reads mappés à plusieurs endroits dans le génome avec *Picard MarkDuplicates* (Broad Institute).
- Trier le fichier bam selon les coordonnées des reads avec *Samtools*.

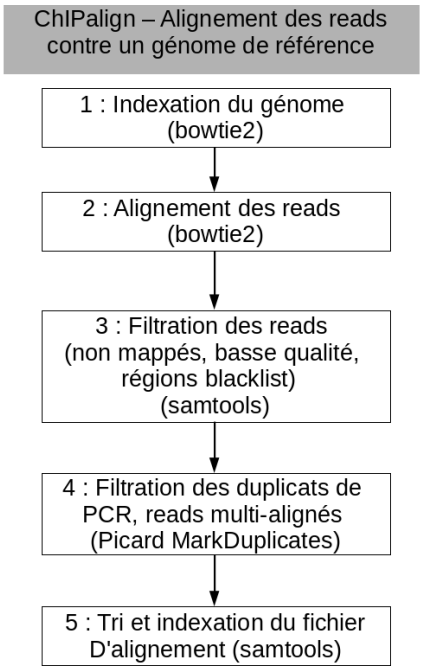


Figure 5: Pipeline *ChIPalign* créé pour aligner les données de ChIP-seq contre un génome de référence et filtrer les reads alignés selon leurs qualité et position

- Indexer le fichier bam final avec *Samtools*.

Table 1: Résultats fournis par le pipeline d'analyse *ChIPalign* pour les quatre sous-unité de la voie NF-κB et l'Input

	RelA	RelB	p50	p52	FOXM1	Input
Nombre de reads	39751988	39597090	39150624	41981223	37481240	31295367
Nombre de reads mappés	38485654	37864800	37829745	40505961	30280568	30354568
Nombre de reads (mapQ >30)	33438199	32913456	32712256	35415978	26193535	26181273
Nombre de reads (-dup)	19564675	24278832	27405603	11780864	21555587	23955866
Nombre de reads (-blacklist régions)	19504167	24206086	27325957	11743326	21498519	23869608
Pourcentage de reads conservés	49,06%	61,13%	69,80%	27,97%	57,36%	76,27%

Les résultats obtenus après utilisation de ce pipeline (Table 1 et Figure 6) divergent en fonction des données utilisées, on remarque que la majorité des reads ont été filtrés pour les données de p52 (seulement 27,97%

des reads sont conservés), on perd également beaucoup de données pour la sous-unité RelA(49,06% des reads sont conservés). Pour ce qui est des autres sous-unités et de l'Input, le taux de reads conservés après filtration est meilleur(respectivement 61,13%, 69,80% et 76,27%). En parallèle, l'alignement à également été réalisé de la même manière sur le fichier de données pour FOXM1. L'alignement du fichier fastq fourni puis la filtration conserve 57% des reads initiaux.

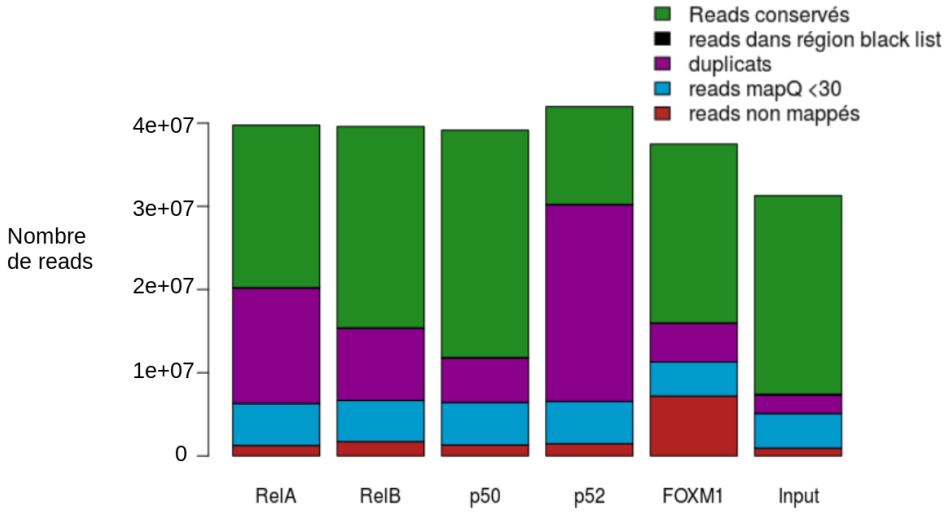


Figure 6: Résultats de la filtration des fichiers d'alignement pour les quatre facteurs de la voie NF-κB, l'input ainsi que pour le facteur oncogénique FOXM1

5 Recherche de pics par sous-unité

5.1 Protocole utilisé pour rechercher les pics sans utiliser de réplicats

Le peak calling est une tâche très importante dont une des difficultés majeure réside dans la sélection des pics d'intérêts tout en éliminant le bruit de fond. Du fait de l'absence de réplicat, cette tâche est rendu beaucoup plus complexe et le protocole standard pour sélectionner les pics, IDR, mis au point par ENCODE [4] ne peut pas être utilisé. Pour réaliser une recherche de pics la plus fiable possible, j'ai mis au point un autre pipeline qui n'est pas basé sur l'analyse IDR : CallPeaks_norep (Figure 7). Pour cela, j'ai tout d'abord transformé les fichiers bam issus de la précédente partie en fichier tagAlign (format BED modifié) avec les outils *Samtools* et *Bedtools bamToBed* (version 2.17.0). Le format tagAlign est un format Bed dans lequel le nom du read a été remplacé par la lettre N. Ce nouveau pipeline est en majeure partie relié à *Macs2* (version 2.1.1.20160309) qui est le logiciel qui va effectuer la recherche de pics.
J'ai choisi *Macs2* qui va associer en premier temps une p-value à chaque région génomique avant de sélectionner les pics. Globalement, la sélection des pics se fait en plusieurs étapes :

- Les fichiers de traitement et contrôles sont ajustés en fonction du ratio entre le nombre total de reads dans le traitement et le nombre total de reads dans le contrôle.

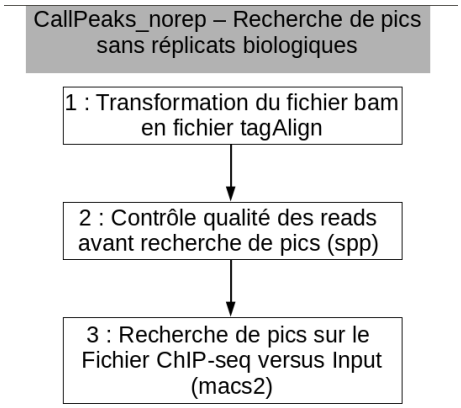
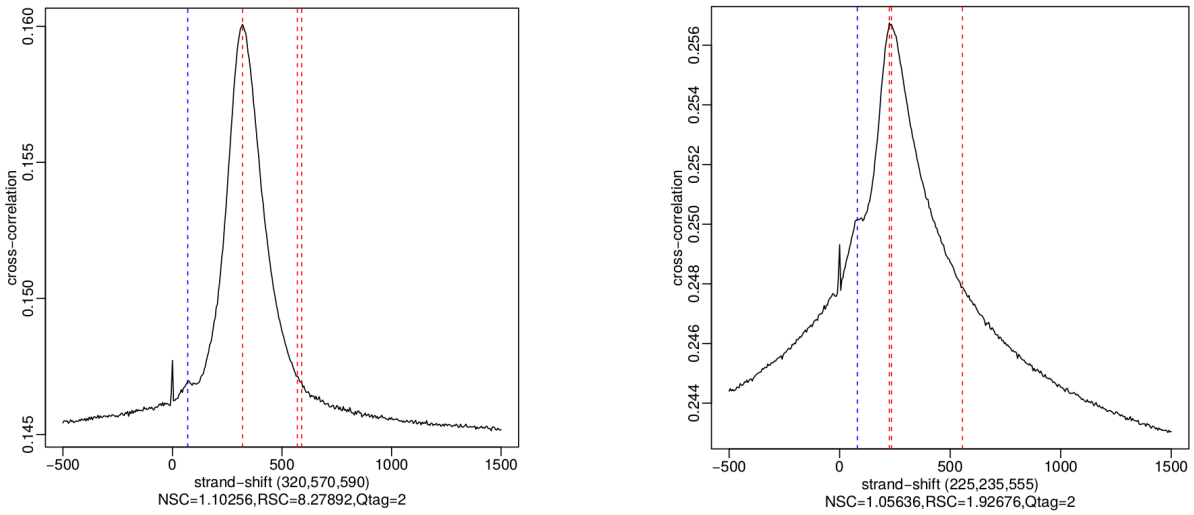


Figure 7: Pipeline CallPeaks_norep créé pour effectuer le peak calling des données de ChIP-Seq en étant le plus sélectif possible

- Les fichiers de traitements et de contrôle seront ensuite multipliés par la taille du plus petit pic.
- Enfin, une distribution de Poisson est utilisée pour calculer les p-values pour l'enrichissement.



(a) (b)

Figure 8: Exemple de graphe produits par le calcul des métriques de qualité pour p52 (a) et FOXM1 (b), la ligne rouge la plus à gauche représente la taille du fragment estimée la plus probable et coïncide avec le "vrai" pic. La ligne bleu représente la taille du read et coïncide avec le pic fantôme (artefact)

Le pipeline propose, avant d'effectuer la recherche de pics, d'effectuer un dernier contrôle qualité sur les fichiers via *SPP* (version 2.0). On utilisera un script spécifique de *SPP* (*run_spp_nodups.R* au lieu de *run_spp.R*) car les duplicats auront été enlevés lors de la filtration des fichiers bam. Dans mon analyse, j'utiliserai d'abord *SPP* pour contrôler la qualité des fichiers sur lesquels effectuer la recherche de pics avant d'utiliser *Macs2*. *SPP* va créer un graphe qui va montrer deux pics : un correspondant à la

taille du read (pic fantôme) et un autre qui va correspondre à l'enrichissement de la taille du fragment prédominante (Figure 8).
Le logiciel va également calculer deux valeurs qui seront utilisées pour évaluer la qualité : le NSC (Normalized Strand Coefficient) qui est le ratio entre le pic de cross-corrélation de la taille du fragment et la cross-correlation du fond, cette métrique doit être supérieur à 1.05 (bonne qualité quand >1.1), et l'autre est le RSC (Relative Strand Coefficient) qui est le ratio entre le pic de la taille du fragment et le pic de la taille du read et qui doit être supérieur à 0.8 (bonne qualité quand >1).
Enfin le Qtag montre la qualité basée sur le RSC (-2:veryLow, -1:Low, 0:Medium, 1:High, 2:veryHigh).
Une autre option de ce pipeline est de fixer le seuil de p-value pour la sélection des pics. Du fait de l'absence de réplicat, c'est ce seuil qui va réellement définir la sélection des pics, il faut donc être suffisamment sélectif pour s'abstraire du bruit de fond. La valeur par défaut de cette option est de 1e-3, dans cette analyse, j'ai utilisé un seuil plus sélectif de 1e-7.

5.2 Identification des pics NF-κB et FOXM1

Ainsi, j'ai utilisé le programme *Macs2* pour effectuer la sélection des pics avec un seuil de p-value de 1e-7 (Table 2). L'ensemble des données est validé par le contrôle qualité effectué par phantomPeakQualTools (seul le NSC de RelB est inférieur à 1.05 [1.04], mais reste relativement proche et possède une bonne valeur de RSC et Qtag).
Pour ce qui est des pics obtenus pour RelA, RelB, p50 et p52, j'ai identifié respectivement 1933, 2243, 12391 et 5817 pics avec ce protocole. Ainsi le nombre de pics identifiés varie grandement d'une sous-unité à une autre. Le grand nombre de pics pour la sous-unité p50 pourrait coller à l'hypothèse que ce facteur serait présent dans le noyau de façon constitutive. On a en revanche un faible nombre de pics pour les sous-unités RelA et RelB et donc un faible nombre de régions potentiellement activatrice de la transcription de gène. D'autre part, la sous-unité p52 a un nombre de pics moyen par rapport à la sous-unité p50.

Table 2: Résultats fournis par le pipeline d'analyse *CallPeaks_norep* qui fournit les métriques déterminées par *phantomPeakQualTools* et le nombre de pics trouvés par *Macs2* au seuil de p-value de 1e-7

	RelA	RelB	p50	p52	FOXM1
Nombre de reads	19504167	24206086	27325957	11743326	21498519
NSC	1,06	1,04	1,05	1,10	1,06
RSC	4,64	3,47	1,97	8,28	1,93
Qtag	2	2	2	2	2
Taille estimée du fragment	240	240	235	320	225
Nombre de pics(p-value 10-7)	1933	2243	12391	5817	10815

Les résultats obtenus nous indiquent des observations divergentes de celles présentées dans d'autres lignées

telle que la GM12878 (lymphome) [3], dans cette lignée le nombre de pics identifiés pour les cinq sous-unités NF-κB (RelA, RelB, cRel, p50 et p52) était respectivement de 20067, 16617, 6765, 4298 et 10814. Cela sous-entend que les mécanismes qui régissent l'activation de cette voie sont différents, car dans la lignée MM1S, p50 est majoritaire, RelA et RelB ont peu de sites de fixation et cRel est probablement absent dans le noyau (hypothèse à vérifier par Western Blot).
Enfin, on note également que l'on identifie un nombre de pics pour le facteur oncogénique FOXM1 du même ordre de grandeur (10815) que le nombre de pics obtenus pour p50.

5.3 Visualisation des pics NF-κB

Il est important par la suite de regarder l'allure des pics obtenus dans un génome browser. Lors de l'utilisation de *Macs2*, une option à été spécifiée afin de produire en plus de résultats pour les pics, un fichier au format bedGraph qui va pouvoir être directement utilisé pour visualiser les pics.
Pour cette visualisation, plusieurs outils existent, le plus utilisé est l'outil en ligne de l'université de Californie UCSC [5]. Cependant, c'est une application en ligne qui est donc limitée par des problèmes de transfert de fichiers (dans le cas de fichiers volumineux) mais elle dispose de grand nombre données qui sont utiles pour des consultations (comme certaines données d'ENCODE par exemple). Pour visualiser nos pics, nous utiliserons donc un autre outil qui est utilisable via un terminal shell : *IGV* (Figure 9).

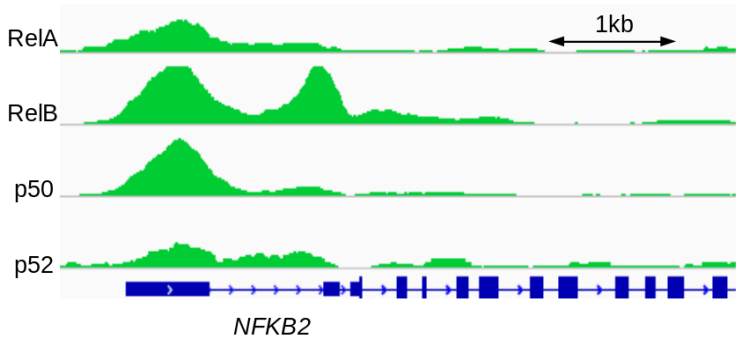


Figure 9: Visualisation des pics obtenus par *Macs2* sous *IGV* au niveau du gène *NFKB2*

6 Découverte du motif consensus

J'ai utilisé ensuite *Chipmunk* [6] pour effectuer une découverte de motif *de novo* dans l'ensemble des pics de chaque sous unité NF-κB que nous avons obtenu. Le but est d'identifier pour chaque sous unité la séquence consensus la plus représentée dans l'ensemble des régions. J'ai recherché des motifs d'une taille de 8 à 12 nucléotides. Dans un second temps, j'ai utilisé un outil issu de la suite *Meme* [7] : *Tomtom* pour comparer les motifs trouvés avec ceux de bases de données existantes. J'ai ressorti 4 motifs consensus (Figure 10): Seule la sous-unité RelB présente le motif consensus du site κB. Pour les motifs trouvés dans les sous unités RelA, p50 et p52, ce sont des motifs riches en GC qui correspondent à des motifs de la famille SP (SP2, EGR1,SP3...) et qui sont décris comme des facteurs de transcription qui se fixent majoritairement dans des séquences riches en G/C de promoteurs de gènes impliqués dans la régulation du cycle cellulaire mais qui jouent également un rôle dans la différenciation des cellules B. J'ai ensuite

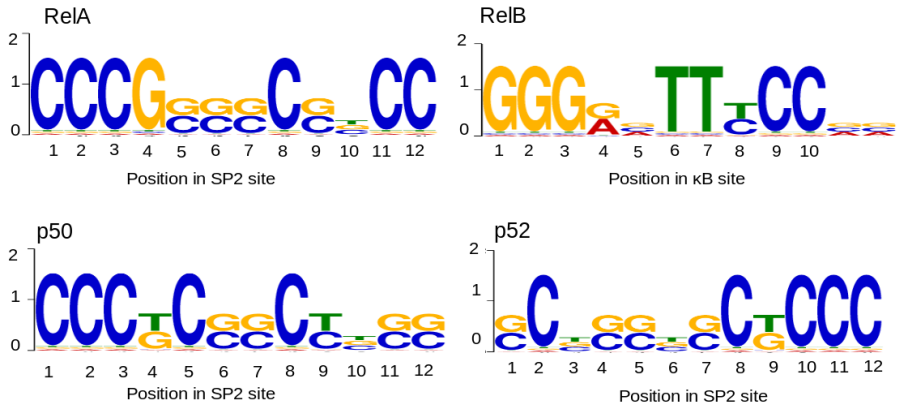


Figure 10: Motif consensus trouvé par *ChipMunk* le plus représenté pour chaque sous-unité de la voie NF-κB

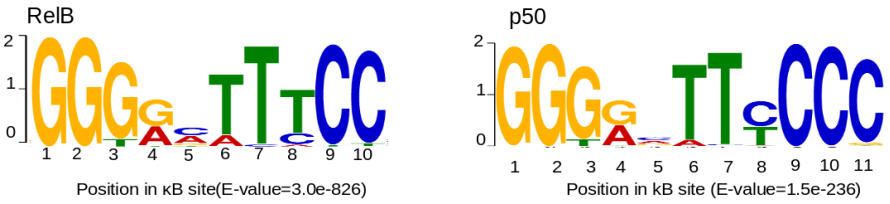


Figure 11: Motif consensus trouvé par *Meme* pour les sous-unités RelB et p50

utilisé la suite *Meme* pour effectuer une découverte de motif sur le top 1000 des pics de chaque sous-unité (Figure 11). On ne retrouve toujours pas le motif consensus κB pour RelA et p52, en revanche on obtient toujours le motif κB de 10 bases pour la sous-unité RelB (E-value = 3.0e-826). On remarque cependant que pour la sous-unité p50, on obtient le motif consensus κB de 11pb avec la cytosine supplémentaire en position 11 du coté 3' comme décrit précédemment dans la littérature [3] (E-value = 1.5e-236).

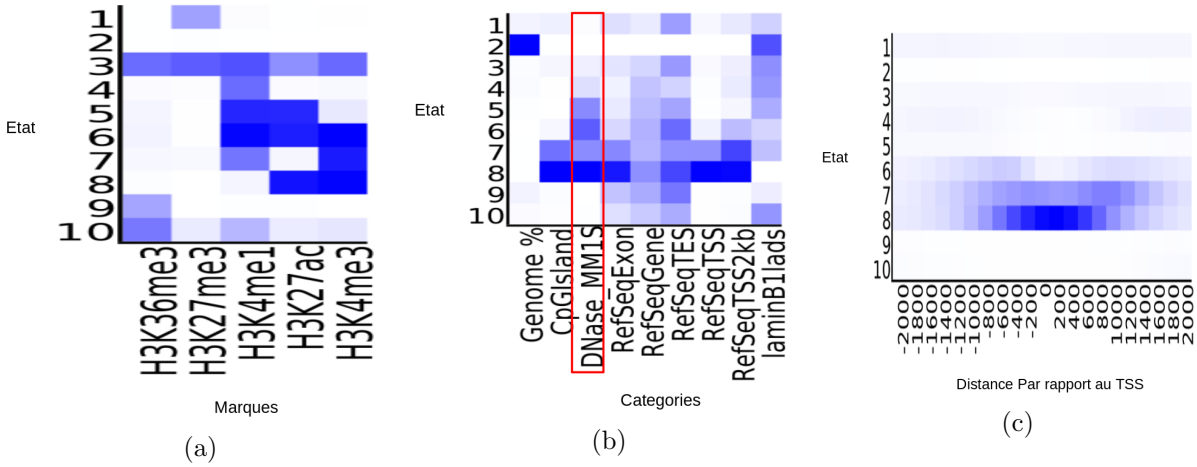
7 Annotation fonctionnelle

7.1 Annotation fonctionnelle des régions du génome de la lignée MM1S

La première étape pour annoter fonctionnellement les régions NF-κB obtenus lors de la recherche de pics est d'identifier sur le génome de la lignée MM1S, quelles sont les régions régulatrices potentielles (promoteurs et enhancers). Pour cela, un logiciel spécialisé dans l'annotation du génome va être utilisé : *ChromHMM* [8]. Cet outil est basé sur l'utilisation d'un modèle de Markov caché qui va permettre la modélisation de la présence ou de l'absence de certaines marques d'histone. Les différentes marques d'histone sont issues de séquençage ChIP-Seq sur des histones qui ont des modifications sur leurs résidus (types méthylation ou acétylation). Ces modifications vont être indicatives d'une région régulatrice ou répressive, d'une région transcrite ou bien encore d'une région active. Cinq marques sont disponibles actuellement pour la lignée MM1S (ENCODE):

- H3K4me1 : méthylation sur le résidu Lysine 4 de la sous-unité H3. Marque souvent associée à une région régulatrice (type enhancer).
- H3K4me3 : tri-méthylation sur le résidu Lysine 4 de la sous-unité H3. Marque souvent associée à une région régulatrice (type promoteur).
- H3K27ac : acétylation sur le résidu Lysine 27 de la sous-unité H3. Marque souvent associée à une région régulatrice active.
- H3K27me3 : tri-méthylation sur le résidu Lysine 27 de la sous-unité H3. Marque souvent associée à une région inactive (réprimée).
- H3K36me3 : tri-méthylation sur le résidu Lysine 36 de la sous-unité H3. Marque souvent associée à la transcription ou à des enhancers géniques.

Pour amener plus de robustesse dans mon annotation, j'ai également récupéré ces mêmes marques d'histone dans deux autres lignées cellulaires : U266 (lignée de myélome multiple issue d'un autre patient; données disponibles sur BluePrint) et GM12878 (lignée de lymphome; données disponibles sur ENCODE). En plus de ces marques, j'ai utilisé les données de la DNase-Seq disponnibles sur ENCODE. Cette technique permet de séquencer l'ensemble des régions accessibles sur le génome grâce à l'utilisation d'une enzyme particulière : La DNase. L'enrichissement de la DNase pour les différents états annotés pourra ainsi être évalué.



(a) (b) (c)

Figure 12: Résultats obtenus après utilisation de *ChromHMM* sur les lignées MM1S, U266 et GM12878 pour les marques H3K4me1, H3K4me3, H3K27ac, H3K27me3 et H3K36mes3. (a) Paramètres d'émission relatifs aux trois lignées par rapport aux états et aux marques; (b) Enrichissement des différents états selon la catégorie pour la lignée MM1S, la DNase-Seq est particulièrement enrichie pour les états 4 à 8 (encadré rouge); (c) Distribution des états autour du TSS (Transcription Start Site) à plus ou moins 2kb pour la lignée MM1S

ChromHMM se déroule en deux étapes. La première consiste à binariser les différents fichiers pour les marques d'histone (on utiliseras les fichiers d'alignement bam) puis le logiciel va réaliser un apprentissage et utiliser le modèle de Markov pour élaborer le modèle (200 itérations). Dans cette étude, le nombre de marques d'histone disponibles étant limité, nous rechercherons 10 états de chromatine différents.

Le résultat et l'annotation du génome comporte une grande partie d'interprétation biologique. Pour annoter correctement les états donnés par *ChromHMM*, je me suis servi de trois résultats : les paramètres d'émission (Figure 12a) qui vont décrire comment les marques sont représentées selon les états (exemple : l'absence de toutes les marques utilisées dans l'état 2 peut indiquer un état hétérochromatine), le second est l'enrichissement de différentes régions génomiques (Figure 12b), pour lequel *ChromHMM* va regarder l'enrichissement de chaque état dans des fichiers de coordonnées de références (RefSeq), par exemple, les états 9 et 10 sont particulièrement enrichis au niveau des exons, gènes et transcription end site (TES) ce qui peut indiquer des régions liées à la transcription. Enfin j'ai regardé la distribution de ces états autour d'un site de début de transcription (TSS) à plus ou moins 2kb (Figure 12c), Par exemple les états 8 et 7 sont particulièrement retrouvés au niveau du TSS et peuvent indiquer des régions promotrices.

Table 3: Annotations des différents états trouvés par *ChromHMM*

ETAT	ANNOTATION	ETAT	ANNOTATION
E1	Régions réprimées	E6	Enhancers Actifs
E2	Hétéro-chromatine	E7	Faibles Promoteurs
E3	Régions répétées/CNV	E8	Promoteurs Actifs
E4	Faibles Enhancers	E9	Transcription faible
E5	Enhancers Actifs	E10	Transcription

A l'aide de ces interprétations, j'ai pu assigner aux différents états un sens biologique (Table 3). Les états 4 à 6 sont annotés en tant que faibles enhancers et enhancers actifs et les états 7 et 8 sont annotés respectivement en tant que faibles promoteurs et promoteurs actifs. On remarque également que l'ensemble de ces états montre un enrichissement pour la DNase-Seq dans la MM1S (encadré rouge, Figure 12b).

7.2 Annotation fonctionnelle des pics obtenus

Je me suis servi des résultats obtenus dans la partie précédente pour annoter les pics identifiés pour les facteurs NF-κB et FOXM1. Tout d'abord j'ai créé un fichier de coordonnées pour les enhancers et un autre pour les promoteurs. Pour se faire, les états 4 à 6 (Table 3) seront regroupés dans le fichier enhancer et les états 7 et 8 seront regroupés dans le fichier promoteur.
J'ai ensuite récupéré les positions des sommets de chaque pic obtenu et les ai étendus à plus et moins 100 bases de chaque côté de façon à obtenir des pics uniformes de longueur de 200 bases (longueur moyenne estimée du fragment d'ADN sur lequel est fixé le facteur de transcription). J'ai ensuite évalué le nombre de pics qui sont localisés dans un promoteur ou un enhancer sur au moins 51% de sa longueur (de façon à ce que le sommet du pic soit pris en compte). Cette procédure a été automatisée sous une nouvelle sous-fonctionnalité du pipeline: *AnnoPeaks*
Les résultats obtenus (Figure 13) montrent que, pour toutes les sous-unités NF-κB, la majorité des pics (plus de 80% pour l'ensemble des quatre sous-unités et FOXM1) sont annotés dans des régions régulatrices enhancers ou promoters. On remarque également que le nombre de promoteurs est supérieur au nombre d'enhancers dans chacune des sous-unités et surtout pour les sous-unités RelA, p50 et FOXM1 (respec-

tivement 89%, 81% et 76% de promoteurs). Seul RelB montre également une proportion importante d'enhancers (40%). En revanche, il n'y a quasiment pas de pics RelA trouvés dans les enhancers (2% soit 56 pics).
Les données étant issues de ChIP-seq sur des facteurs de transcription, il est attendu que la majorité des pics identifiés soit localisée dans des régions régulatrices telles que des promoteurs ou des enhancers. En revanche, Nous avons observé pour chacun de ces facteurs qui étaient préférentiellement localisés au niveau de promoteurs plutôt qu'au niveau d'enhancer (seule la sous-unité RelB montre une proportion importante de pics localisés au niveau d'enhancers), ces observations diffèrent des résultats obtenus dans la littérature pour le ChIP-Seq des facteurs NF-κB dans la lignée cellulaire GM12878 où la majorité des pics obtenus étaient situés au niveau des enhancers [3].

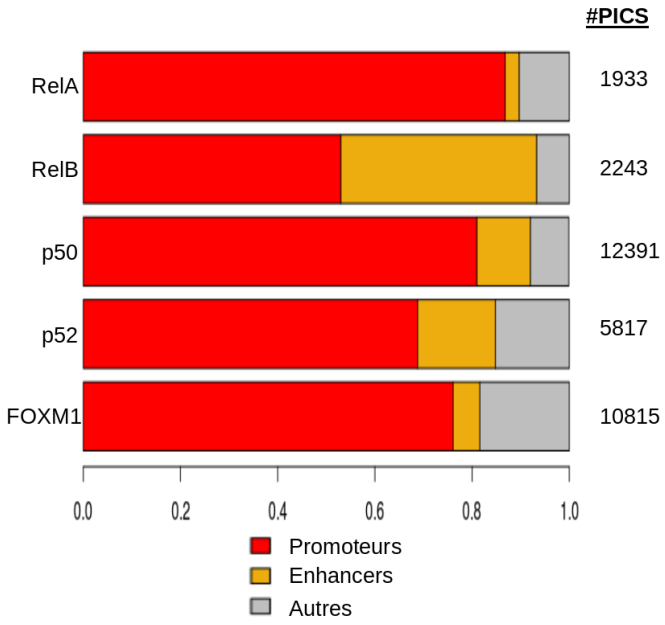


Figure 13: Annotation des pics NF-κB obtenus et correspondant à un promoteurs (rouge) ou à un enhancer (jaune) donné en fraction du nombre de pics total

8 Recherche de patterns de co-binding : étude des deux voies

Une fois que les pics ont été annotés, j'ai effectué un clustering de l'ensemble des régions de promoteurs, puis sur l'ensemble des régions d'enhancers de NF-κB. Le but est de découvrir des de co-binding différents selon les sous-unités qui s'y fixent (SBP: subunit binding pattern).
Pour se faire, j'ai utilisé *SeqMiner* (v1.2) en appliquant la méthode des k-means et en effectuant une normalisation par rapport au rang de l'intensité du signal ChIP-Seq (utilisation des fichiers d'alignement). Ce clustering se fera sur toutes les régions sélectionnées par la recherche de pics et qui comportent au moins un pic NF-κB.
J'ai obtenu par cette technique 7 clusters pour les promoteurs indiquant 6 SBPs différents et 5 pour les

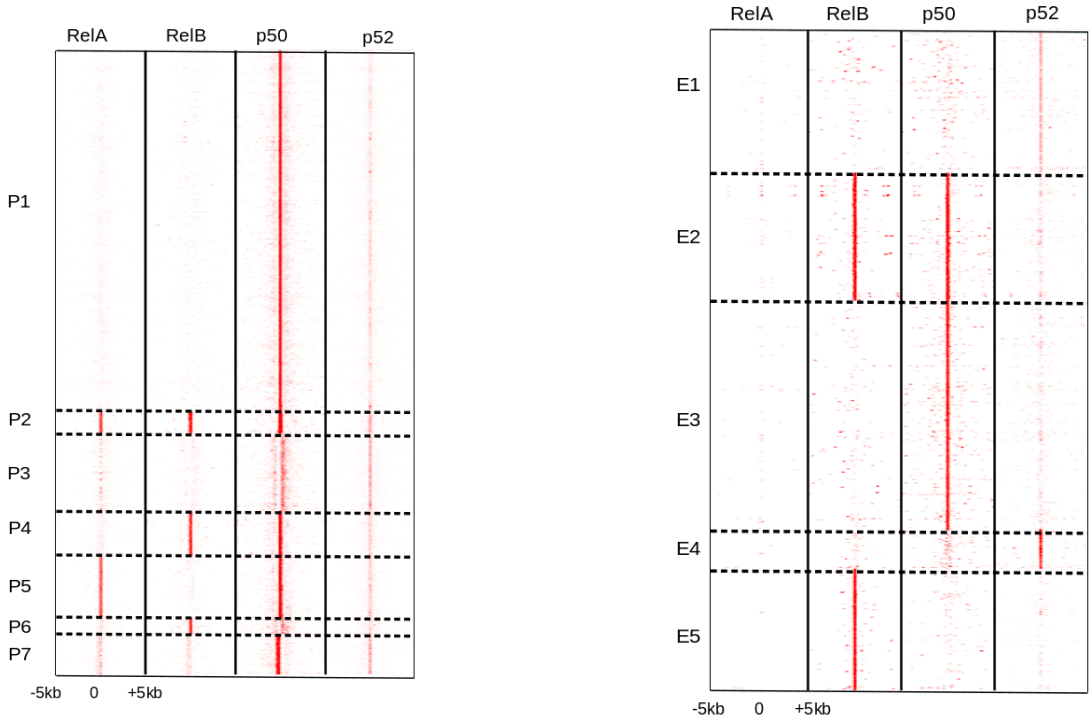


Figure 14: Recherche de sites de fixations parmi les différents pics NF-κB dans les promoteurs (a) et dans les enhancers (b)

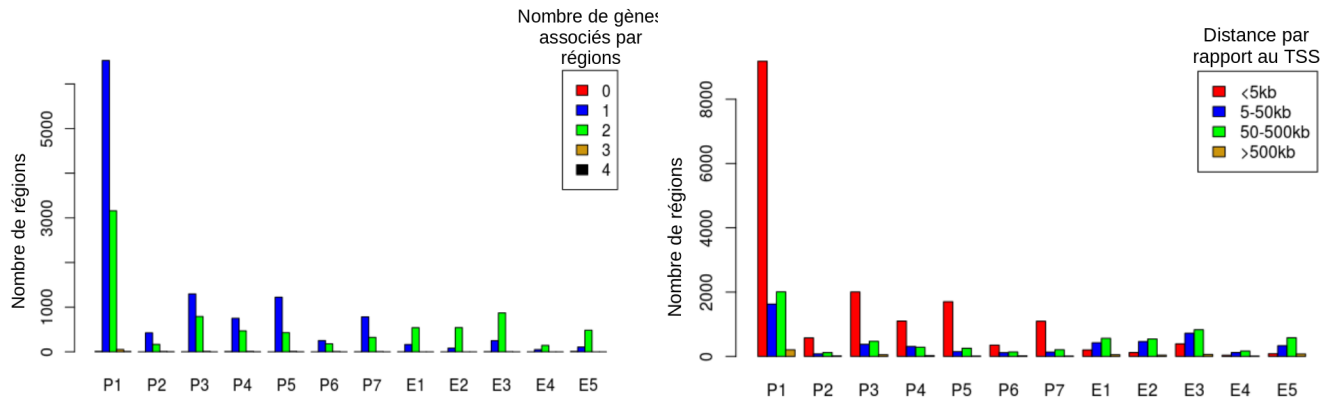
enhancers (6 SBPs différents) (Figure 14).
Pour ce qui est des promoteurs (Figure 14a), la sous-unité p50 est majoritairement présente dans 5 des 7 SBPs trouvés (sauf P3 et P6). RelA est majoritairement retrouvé dans P2 et P5, RelB est présent dans P2, P4 et p6. Pour ce qui est de p52, son signal est plus faible comparé aux autres mais on distingue une légère augmentation dans P3 et P6. On a un seul SBP qui montre un signal pour les quatres sous-unités NF-κB : P2. D'après la littérature [1], seul le SBP P5 semble correspondre à la voie canonique avec les possibles dimères RelA:RelA, RelA:p50 et p50:p50. Pour ce qui est de la voie non canonique, les SBPs P3 (p52:p52), P4 (RelB:RelB, RelB:p50) et P6 (RelB:RelB) pourraient être spécifiques de l'activation de cette voie. Pour ce qui est de P1,P2 et P7, ces SBPs pourraient résulter de l'activation simultanée des deux voies et de la formation de dimères hybrides entre ces deux voies (par exemple RelA:RelB). Dans P1 et P7 il semble y avoir majoritairement présence de l'homo-dimère p50:p50.
Au regard des enhancers (Figure 14b), la sous-unité RelA est quasiment absente de tous les SBP (seulement 56 régions enhancers). p52 est présent dans les clusters E1 et E4 (seul dans ces deux clusters mais avec une intensité légèrement différente), RelB est fortement présent avec p50 dans le cluster E2. p50 est la seule sous-unité présente dans le cluster E3 et RelB lui est seul dans E5. Ces résultats pourraient indiquer des enhancers spécifiques de la voie non canonique de NF-κB pour l'ensemble des SBPs et induirait l'hypothèse qu'il n'y a pas de dimères issus de la voie canonique (avec RelA ou cRel) qui se fixeraient au niveau des enhancers. On retrouverait spécifiquement des homo-dimères p52:p52 (E1 et E4), p50:p50 (E3), RelB:RelB (E5) et un seul type d'hétéro-dimère RelB:p50 (E2) au niveau des enhancers
Ces hypothèses seront testées prochainement *in vitro* par des techniques de western blot afin de vérifier

biologiquement quels sont les dimères présents dans le noyau de la lignée MM1S.

9 Associations des SBPs à des processus biologiques

9.1 Association gènes-SBPs

L'application en ligne *Great* (version 3.0.0) est utilisée pour effectuer l'annotation fonctionnelle des clusters de promoteurs et d'enhancers obtenus, cette partie de l'analyse n'est donc pas automatisable dans le pipeline. Pour utiliser *Great*, j'ai sélectionné les mêmes paramètres par défaut que ceux proposés par McLean et al [9] : La recherche se fera sur le génome hg19 avec le génome entier en tant que background, chaque gène est étendu à 5kb en aval et 1kb en amont par rapport au site de début de transcription (=TSS) (proximal), si une région spécifiée n'est associée à aucun gène avec cette règle, une autre recherche (distal) est effectuée et va permettre de regarder jusqu'à 1Mb autour de la région pour lui associer un gène. Cette règle s'applique pour les promoteurs et enhancers car il existe des promoteurs alternatifs qui peuvent être situés à plus de 5kb d'un gène de la même façon que les enhancers distaux.
On remarque ainsi que pour l'association des gènes aux régions (Figure 15a), on remarque que la majorité des régions dans les promoteurs (P1 à p7) sont associées à un seul gène. Pour ce qui est des clusters d'enhancers (E1 à E5), il y a généralement deux gènes d'associés à chaque régions. Ces différences s'expliquent par le fait qu'un promoteur est le plus souvent situé au niveau du gène tandis qu'un enhancer peut être situé beaucoup plus loin et donc impacter plusieurs gènes.
Ces résultats sont confirmés lorsque l'on s'intéresse à la distance de l'association entre le gène (TSS) et la région qui lui est associée(Figure 15b). On peut voir que la plupart des distances des associations est inférieure à 5kb pour les clusters de promoteurs. Cette distance augmente dans les clusters d'enhancers puisqu'une importante part à une distance entre 5 et 50kb (enhancers proximaux) mais la plus grande partie des enhancers se situent à une distance d'entre 50 et 500kb d'un TSS.



(a) (b)

Figure 15: (a) Nombre de gènes associés par région pour chaque SBP; (b) Distance de l'association gène-région par rapport au TSS

Cet outil va également permettre de récupérer la liste de tous les gènes qui ont été associés à chacun des SBPs, on observe ainsi que certains SBPs sont peuvent être associés avec le même gène(Figure 16). De

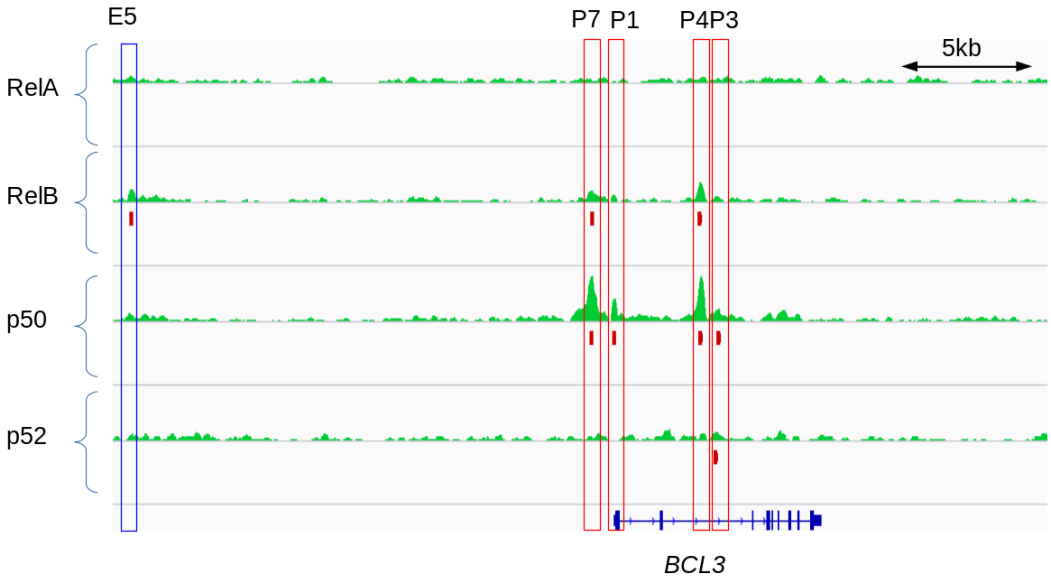


Figure 16: Association de 5 SBPs différents (E5, P1, P3, P4 et P7) avec le même gène BCL3 par l'outil *Great* et visualisation des pistes dans *IGV*. Les pistes en vert représentent le signal ChIP-seq pour chaque sous-unité et les pistes rouges représentent les pics identifiés par le pipeline

façon générale, une même région d'un SBP peut être associée à plusieurs gènes tout comme un même gène peut être associé a plusieurs SBPs. Je travaillerais ensuite avec des données de transcriptome issues de puces pour vérifier quels gènes associés à des sites de fixation NF-κB sont exprimés pour identifier potentiellement des SBPs qui auraient tendance à activer l'expression ou au contraire à la réprimer.

9.2 Association processus biologiques-SBPs

Great permet également, en associant les gènes sélectionnés avec des processus biologiques, d'identifier les fonctions et processus les plus enrichis pour chaque cluster.
Dans les bases de données, chaque gène est associé à un ou plusieurs processus biologiques. En observant les gènes associés à un SBP, *Great* peut calculer l'enrichissement de chacun des processus biologiques. Pour ce calcul, il va y avoir deux tests utilisés: un test binomial et un test hypergéométrique. Pour cette analyse, j'ai conservé uniquement les "GO biological Process" qui avaient un FDR supérieur à 0.01 (1%) pour les deux tests (Figure 17). Pour l'ensemble des SBPs testés, seuls deux clusters(P6 et E4) n'ont pas d'enrichissement significatif pour un processus biologique. Deux hypothèses peuvent expliquer cette absence d'enrichissement: soit les gènes associés au régions de ces SBPs impactent des processus biologiques différents et aucun de ceux-ci n'est suffisamment représenté comparé aux nombre de gènes sélectionnés ou alors le nombre de régions contenues par ces SBPs est trop faible et aucun enrichissement significatif ne peut être mis en évidence. La deuxième hypothèse est la plus probable au vue de la taille de ces deux clusters. On obtient cependant pour ces deux clusters la liste des gènes associés aux régions. Les résultats montrent également que les processus biologiques enrichis ne sont pas les mêmes en fonction du SBPs (Figure 17), on peut voir que les SBPs identifiés dans les promoteurs (à l'exception de P2) semblent plus intervenir dans la régulation des fonctions de "housekeeping" de la cellule (épissage,

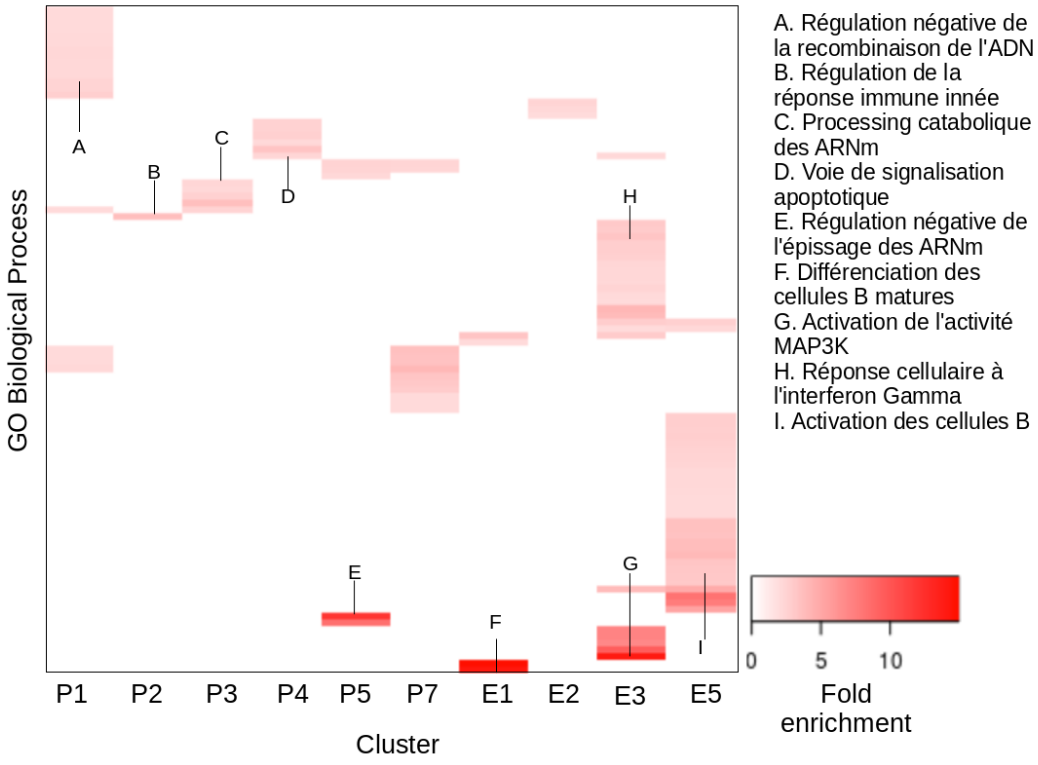


Figure 17: Enrichissement des processus biologiques GO dans chacun des SBPs, seuls deux cluster (E4 et P6) n'ont pas d'enrichissement significatifs pour aucun processus biologiques. Chaque ligne correspond à un processus biologique unique avec un FDR > 0.01

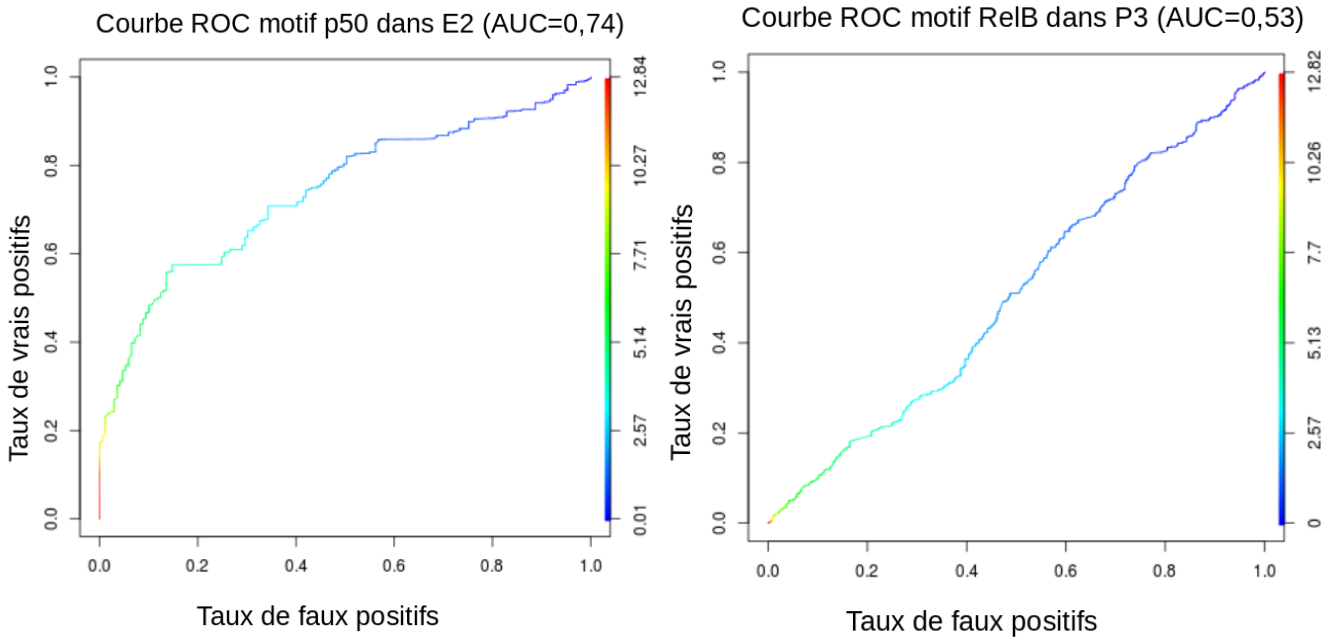
apoptose, recombinaison de l'ADN). En revanche pour les processus mis en évidences pour les SBPs localisés au niveau d'enhancers et P2, semblent concerner des processus plus spécifiques de la cellule B et de l'immunité en générale (activation des cellules B, réponse aux cytokines et à l'interferon gamma...).
Il semblerait donc que le rôle des sous-unités de la voie NF-κB ne soit pas le même selon la fixation de ces facteurs sur les séquences régulatrices (promoteurs ou enhancers). De cette fixation va dépendre les gènes régulés et donc les fonctions biologiques impactées. Une fois encore, les données de transcriptome seront importantes pour vérifier l'expression des gènes impactant certaines fonctions biologiques ciblées.

10 Analyse des motifs κB

10.1 Contruction des courbes ROC et calculs d'aire sous la courbe

J'ai ensuite analysé l'enrichissement des motifs consensus κB trouvés par *Meme* pour les sous-unités RelB et p50 dans chaque SBP.
Pour cela, j'ai récupéré pour chacun de ces motifs la matrice de probabilité de chaque nucléotide à chaque position dans le motif (PWM) qui est fournit par *Meme* dans les résultats.
Pour réaliser cette analyse, j'ai utilisé les séquences de 200 bases autour du sommet de chaque pic pour

chaque SBP puis pour chaque fichier de séquence, j'ai créé aléatoirement un fichier de "background", ce fichier contient des séquences prises à des positions aléatoires dans le génome sans que ces séquences ne soient présentes dans des régions "blacklist" ou dans des exons codants.
J'ai ensuite utilisé une partie de la suite logicielle *Homer* afin de scanner l'ensemble des séquences à tester et des séquences du background et de réaliser tous les alignements possibles dans toutes les orientations du motif consensus. A chaque fois qu'un alignement sera effectué, le logiciel va lui attribuer un score d'appariement selon son orientation, la présence de gap... J'ai obtenu tous les scores d'appariement pour les deux jeux de données.
Si le motif est significativement enrichi dans un jeu de donnée, alors il devrait y avoir un seuil de score d'appariement pour lequel il y a plus de séquences avec un score supérieur à ce seuil dans la liste de séquences à tester que dans celles prises à des positions aléatoires.
Avec ces résultats, j'ai pu finalement établir pour chaque SBP et pour chaque motif consensus des courbes ROC (Figure 18). En calculant ensuite l'aire sous la courbe (AUC) de cette courbe ROC, on peut définir si le SBP étudié est significativement enrichis pour un motif particulier. une valeur d'AUC de 0.5 signifie qu'il n'y a pas de différence entre les deux sets de données analysées (Figure 18b) et qu'il n'y a pas de différence avec un enrichissement aléatoire. Si l'AUC est supérieure à cette valeur (Figure 18a), alors on peut dire que le motif étudié est significativement enrichi pour le cluster étudié.



(a) (b)

Figure 18: Exemple de courbes ROC réalisées pour déterminer l'enrichissement de motif de façon significative; (a) exemple d'un SBP (E2) spécifiquement enrichis pour le motif consensus p50 (AUC=0.74); (b) exemple d'un SBP (P3) non significativement enrichis pour le motif consensus RelB (AUC=0.53)

10.2 Enrichissement des SBPs pour les motifs κB

J'ai donc compilé les valeurs d'AUC pour l'ensemble des SBPs et pour les deux motifs consensus étudiés afin de déterminer quels SBPs dépendent spécifiquement de la fixation des sous-unités NF-κB sur leurs sites consensus (Figure 19).

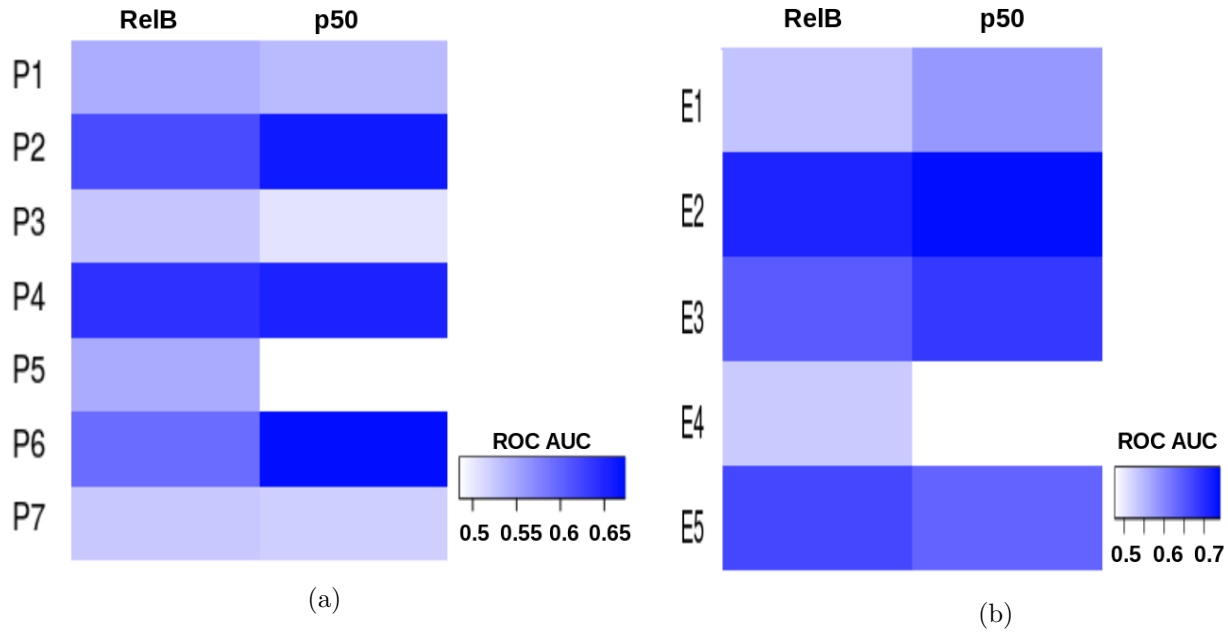


Figure 19: Enrichissement des motifs consensus κB trouvés pour RelB et p50 au niveau des clusters de promoteurs (a) et d'enhancers (b)

Pour les promoteurs, les AUC déterminées sont généralement plus faibles que celles obtenues pour les enhancers. On constate qu'il y a trois clusters qui montrent un enrichissement spécifique du motif κB: P2, P4 et P6. Ces trois clusters ont un enrichissement plus fort pour le motif consensus de p50 que pour le motif consensus de RelB (surtout P2 et P6). Les quatre autres clusters ne semblent pas être spécifiquement enrichi pour un motif κB.
Pour ce qui est des enhancers, on a une nouvelle fois trois clusters significativement enrichis pour un motif κB: le cluster E2 qui est enrichi pour les deux motifs consensus de RelB et p50, le cluster E3 qui est significativement enrichi pour le motif consensus de p50 et le cluster E5 qui montre un enrichissement plus important pour le motif consensus de RelB. Pour les autres clusters, il ne semble pas y avoir d'enrichissement signifactifs d'un motif κB.
Ces résultats correspondent à ceux obtenus précédemment (Figure 14), où l'on avait dans P2, P4, P6 et E2 la présence conjointe de la sous-unité RelB et de p50. pour E3, on avait juste p50 de présent alors que E5 ne montrait la présence que de RelB.
Ce résultats sous-entend que ces six clusters sont spécifiques de la fixation de RelB et p50 sur leurs séquences consensus, au contraire des autres clusters qui semblent requérir la fixation de co-facteurs.

11 Recherche de motifs autres que κB

Il y a donc six clusters dont quatre SBPs annotés dans les promoters(P1, P3, P5 et P7) et deux SBPs annotés dans les enhancers(E1 et E4) pour lesquels il semblerait que la fixation des sous-unités NF-κB requiert la fixation d'un co-facteur sur sa propre séquence consensus.
Pour déterminer quels sont ces co-facteurs, j'ai effectué une découverte de motif *de novo* avec *Meme*, ainsi qu'une recherche d'enrichissement de motifs connus avec *Homer*.
Pour les deux clusters d'enhancers E1 et E4 qui correspondent à deux clusters où se fixe la sous-unité P2, j'ai trouvé un enrichissement correspondant au motif de la E-box dans ces deux clusters (Figure 20). Il s'agit d'un motif court de 6 nucléotide: 5'-CANNTG-3' qui peut fixer de nombreux facteurs. Dans le cas de ces deux sous-unités il semblerait que les facteurs impliqués dans la fixation du dimère p52:p52 sur la séquence consensus E-box soient ITF2 qui est un activateur de transcription ou ID4 qui est un répresseur de transcription. Des données de transcriptome ou de ChIP-seq seront nécessaires afin dé vérifier quel facteur est spécifiquement co-localisé sur ces sites dans ces deux clusters.

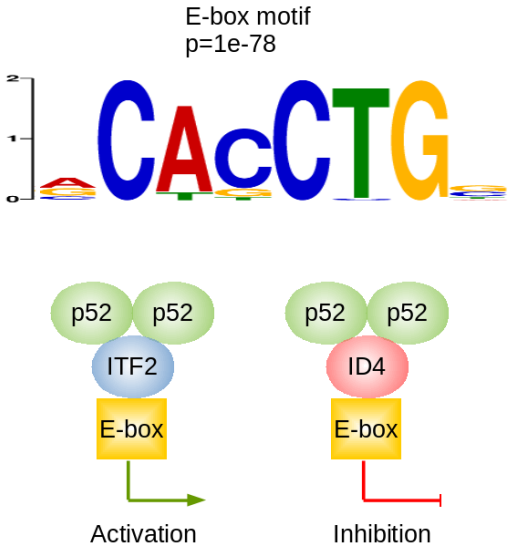


Figure 20: Motif consensus E-box trouvé au niveau des séquences des SBPs E1 et E4

12 Conclusion

References

[1] Louis M Staudt. Oncogenic activation of nf-κb. *Cold Spring Harbor perspectives in biology*, 2(6):a000109, 2010.

[2] Gareth J Morgan, Brian A Walker, and Faith E Davies. The genetic architecture of multiple myeloma. *Nature Reviews Cancer*, 12(5):335–348, 2012.

[3] Bo Zhao, Luis A Barrera, Ina Ersing, Bradford Willox, Stefanie CS Schmidt, Hannah Greenfeld, Hufeng Zhou, Sarah B Mollo, Tommy T Shi, Kaoru Takasaki, et al. The nf-κb genomic landscape in lymphoblastoid b cells. *Cell reports*, 8(5):1595–1606, 2014.

[4] Stephen G Landt, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, Peter Bickel, James B Brown, Philip Cayting, et al. Chip-seq guidelines and practices of the encode and modencode consortia. *Genome research*, 22(9):1813–1831, 2012.

[5] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002.

[6] Ivan V Kulakovskiy, VA Boeva, Alexander V Favorov, and VJ Makeev. Deep and wide digging for binding motifs in chip-seq data. *Bioinformatics*, 26(20):2622–2623, 2010.

[7] Timothy L Bailey, James Johnson, Charles E Grant, and William S Noble. The meme suite. *Nucleic acids research*, 43(W1):W39–W49, 2015.

[8] Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Noam Shoresh, Lucas D Ward, Charles B Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011.

[9] Cory Y McLean, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger, and Gill Bejerano. Great improves functional interpretation of cis-regulatory regions. *Nature biotechnology*, 28(5):495–501, 2010.