Université de Montréal

Développement d'outils pour l'analyse de données de ChIP-seq et l'identification des facteurs de transcription

par
Eloi Mercier

Département de bioinformatique
Faculté de médecine

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en bioinformatique

05, 2011

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

Développement d'outils pour l'analyse de données de ChIP-seq et l'identification des facteurs de transcription

présenté par:

Eloi Mercier

a été évalué par un jury composé des personnes suivantes:

Sylvie Mader ,	président-rapporteur
Raphael Gottardo ,	directeur de recherche
François Robert,	membre du jury

Mémoire accepté le: . . . . . .

RÉSUMÉ

La méthode ChIP-seq est une technologie combinant la technique d'immuno-précipitation de chromatine avec le séquençage haut-débit et permettant l'analyse *in vivo* des facteurs de transcription à grande échelle. Le traitement des grandes quantités de données ainsi générées nécessite des moyens informatiques performants et de nombreux outils ont vu le jour récemment. Reste cependant que cette multiplication des logiciels réalisant chacun une étape de l'analyse engendre des problèmes de compatibilité et complique les analyses. Il existe ainsi un besoin important pour une suite de logiciels performante et flexible permettant l'identification des motifs. Nous proposons ici un ensemble complet d'analyse de données ChIP-seq disponible librement dans R et composé de trois modules PICS, rGADEM et MotIV. A travers l'analyse de quatre jeux de données des facteurs de transcription CTCF, STAT1, FOXA1 et ER nous avons démontré l'efficacité de notre ensemble d'analyse et mis en avant les fonctionnalités novatrices de celui-ci, notamment concernant le traitement des résultats par MotIV conduisant à la découverte de motifs non détectés par les autres algorithmes.

Mots clés: Génétique, Régulation, Facteur de transcription, ChIP-seq

ABSTRACT

ChIP-seq is a technology combining the chromatin immunoprecipitation method with high-throughput sequencing and allowing the analysis of transcription factors *in vivo* on a genome wide scale. The treatment of such amount of data generated by this method requires strong computer resources and new tools have been recently developed. Though this proliferation of software performing only one step of the analyze leads to compatibility problems and complicates the analysis. Thus, there is a real need for an integrated, powerful and flexible pipeline for motifs identification. Here we proposed a complete pipeline for the analysis of ChIP-seq data freely available in R and composed of three R packages PICS, rGADEM and MotIV. Analyzing four data sets for the human transcription factors CTCF, STAT1, FOXA1 and ER we demonstrated the efficiency of or pipeline and highlighted its new features, especially concerning the processing of the results by MotIV that led to the identification of motif not detected by other methods.

Keywords: Genetics, Regulation, Transcription Factors, ChIP-seq

TABLE DES MATIÈRES

LISTE DES ANNEXES

LISTE DES SIGLES

ADN	Acide désoxyribonucléique
ARN	Acide ribonucléique
AP1	Activator Protein 1
ChIP-chip	Chromatin Immunoprecipitation on chip
ChIP-seq	Chromatin Immunoprecipitation sequencing
CRM	Cis-Regulatory Module
ER	Estrogen Receptor
FOXA1	Forkhead Box A1
FDR	False Discovery Rate
GABP	Growth-Associated Binding Protein
PSSM	Position-Specific Scoring Matrix
PWM	Position Weight Matrix
STAT1	Signal Transducers and Activators of Transcription 1
TF	Transcription Factor
TSS	Transcription Start Site

REMERCIEMENTS

Raphael Gottardo et Arnaud Droit pour la patience dont ils ont fait preuve tout au long de ma maitrise. L'IRSC pour sa participation financière par l'intermédiare de la bourse d'excellence bIT. L'IRCM pour les locaux et la qualité des services. Tous les collaborateurs extérieurs au projet qui nous ont fournis les données sans lequelles il n'aurait pas pu aboutir.

CHAPITRE 1

INTRODUCTION

1.1 Régulation des gènes

L'information génétique est encodée dans l'ADN (Acide Désoxyribonucléique) présent dans les noyaux de nos cellules. Ce code est constitué de quatre nucléotides que sont l'adénine (notée A), la cytosine (C), la guanine (G) et la thymine (T). C'est l'ordre de ces nucléotides qui définit les caractéristiques de l'ensemble de l'organisme. Notre génome est composée de près de 3,4 milliards de paires de bases. Cependant, moins de 2% de notre génome code pour une protéine ou de l'ARN et sont donc le point de départ de la machinerie cellulaire de l'organisme.

Les deux principaux mécanismes amenant à la synthèse d'une protéine fonctionnelle sont la transcription et la traduction. La transcription est le processus de copie de l'ADN en ARN (Acide ribonucléique) dans le noyau cellulaire grâce à l'ARN polymérase. Après plusieurs étapes de maturation, l'ARN messager (ARNm) est exporté dans le cytoplasme de la cellule où il est traduit en protéine par les ribosomes. Chaque triplet de nucléotides, nommé codon, porté par l'ARNm code pour un acide aminé différent et va définir la nature de la protéine ainsi synthétisée.

Ces deux étapes primordiales pour l'organisme sont fortement encadrées. De nombreux mécanismes cellulaires permettent à l'organisme de réguler précisément le moment et les conditions de l'expression des gènes.

1.2 Régulation de la transcription et facteurs de transcription

La régulation de la transcription est la phase du contrôle de l'expression des gènes agissant au niveau de la transcription de l'ADN. Elle s'effectue principalement par la modulation du taux de transcription grâce aux éléments cis-régulateurs et trans-régulateurs

appelés facteurs de transcription(TF). Le terme cis signifie qu'ils agissent sur la même molécule d'ADN c'est à dire le chromosome contrairement aux éléments trans qui sont capables d'agir sur des gènes distants de celui qui les a transcris.

Un facteur de transcription est une protéine possédant une structure 3D spécifique lui permettant de se fixer à une séquence d'ADN particulière. C'est ainsi que grâce à son ou ses domaines de liaison à l'ADN, il peut s'attacher à l'ADN proche de l'endroit du gène à réguler. Par la suite, d'autres domaines du facteur de transcription vont permettre de favoriser ou de réduire la transcription de la séquence d'ADN du gène cible en ARN messager. Un des principaux mécanismes d'action des facteurs de transcription est la modification de la stabilité de la liaison entre l'ADN et l'ARN polymérase. Cela peut aussi passer par une modulation de l'accessibilité de la molécule d'ADN par une acétylation ou une désacétylation des histones.

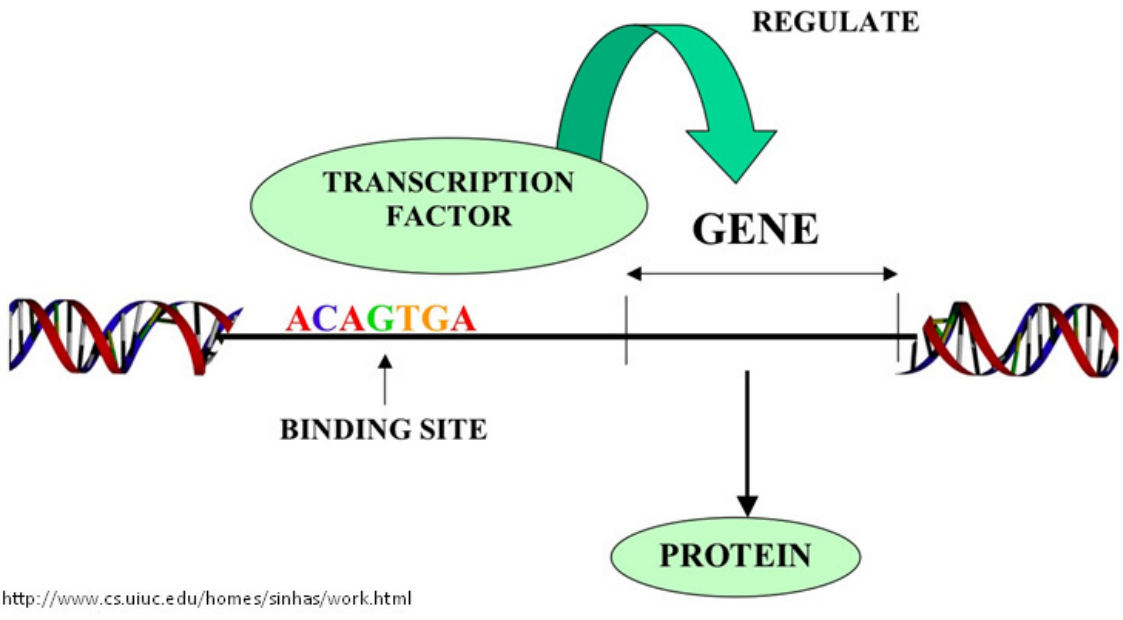


Figure 1.1 – Une représentation de l'action d'un facteur de transcription, se liant à l'ADN au niveau du site de liaison et agissant sur le taux de transcription du gène en ARNm et de ce fait réguler la production de la protéine.

Les facteurs de transcription agissent généralement de concert, formant des modules de régulation en cis (CRM). Pour que la transcription du gène ait lieu, il est nécessaire que chacun des facteurs de transcription se fixe à la molécule d'ADN [7]. Pour que la transcription puisse s'initialiser, il est parfois nécessaire qu'un certain nombre de co-facteur ou de protéines jouant le rôle d'intermédiaire soit présent. Chaque facteur de transcription se doit donc d'être dans une conformation qui lui permette de recevoir les autres protéines. Ainsi, la régulation de la transcription peut être modulée finement grâce à l'action coordonnée de plusieurs facteurs de transcription [8].

Notez enfin qu'un facteur de transcription peut agir sur sa propre régulation. Par exemple, dans une boucle de rétroaction négative, le facteur de transcription agit comme son propre répresseur et ainsi contrôler lui même son niveau d'expression au sein de la cellule.

1.2.1 Domaine de liaison à l'ADN des facteurs de transcription

Un des points clé des facteurs de transcription est la conformation spatiale particulière adoptée par leur site de liaison à l'ADN qui leur permet de "reconnaître" une courte séquence d'ADN simple ou double brin. Les domaines de liaison à l'ADN des facteurs de transcription sont classifiés en différents groupes selon leurs agencements 3D (hélice-boucle-hélice, doigts-de-zinc, leucine zipper,...) [11].

Si un domaine de liaison à l'ADN reconnait un motif d'ADN précis, il existe souvent une tolérance pour certains nucléotides dans la séquence d'ADN cible. Ainsi, il est possible d'intervertir un nucléotide sans perdre la reconnaissance du motif par le facteur de transcription correspondant. Chaque changement modifie cependant l'affinité du domaine de liaison avec la séquence d'ADN tant et si bien que certaines séquences sont plus susceptibles d'être reconnues que d'autres. On ne peut dès lors associer un facteur de transcription à une séquence d'ADN unique, mais à un ensemble de motifs, chacune ayant une probabilité de liaison différente selon les nucléotides présents.

On représente donc les domaines de liaison à l'ADN selon la probabilité de chaque nucléotide à chacune des positions de la séquence d'ADN cible. Cette représentation prend le plus souvent la forme d'une matrice, dont les colonnes correspondent aux positions dans le motif et les lignes le taux d'occurrences des nucléotides A, C, G ou T à cette position. Ces matrices sont appelées PWMs, de l'anglais Position Weight Matrix [33]. Il arrive également que ces matrices indiquent non plus le taux d'occurrence, mais le nombre d'occurrence des nucléotides à chaque position à la suite d'expériences en laboratoire, où chaque motif reconnu par le facteur de transcription est rapporté et sert à calculer la matrice. On parle alors de Position-Specific Scoring Matrix (PSSM) et il est possible de passer de celles-ci à une PWM en divisant chaque colonne par le somme de chacune.

On définit la séquence consensus comme celle étant la plus susceptible d'être reconnue. Elle est constituée des nucléotides ayant la plus grande probabilité de présence à chaque position. En cas d'égalité des fréquences, on utilise la lettre N pour signifier la non spécificité du site de liaison pour cette position. Certains logiciels utilisent parfois une représentation plus poussée avec une lettre de l'alphabet associée à chacune des configurations possibles.

position	1	2	3	4	5	6
A	0	0.8	0.25	0.3	0.1	0
C	1	0.1	0.25	0.3	0	0
G	0	0	0.25	0.2	0	1
T	0	0.1	0.25	0.2	0.9	0
consensus	C	A	N	N	T	G

Figure 1.2 – Une PWM représentant la séquence du site de liaison CANNTG. À chaque position est associée la probabilité de la présence d'un nucléotide. Dans cet exemple, on peut observer une probabilité de présence de la cytosine (C) en position 1 du motif de 100%. Cela signifie qu'un tout autre nucléotide (A, G ou T) à cet emplacement empêchera la reconnaissance du motif.

1.2.2 Contenu informatif des PWMs

Nous connaissons ainsi pour chaque position le ou les nucléotides les plus susceptibles d'être présents. Il faut cependant tenir compte des probabilités. Une probabilité de 0.6 pour C et 0.4 pour A ne nous renseigne pas autant qu'une probabilité de 1 et 0 respectivement. Ce dernier cas est plus informatif pour nous, l'information fournie a plus de valeur. Pour représenter cela, on utilise un score, IC de l'anglais Information Content, représentant la quantité d'information contenue à chacune des positions de la PWM. Il reflète donc le degré de conservation des nucléotides à chacune des positions du motif ou en d'autres termes, le degré de tolérance aux substitutions.

Le contenu informatif d'une PWM **P** à la position **i** est défini comme suit :

$$ic_i = 2 + \sum_{j=1}^{4} \begin{cases} P[i,j] * log(P[i,j]) & \text{si } P[i,j] > 0 \\ 0 & \text{sinon} \end{cases}$$

Le contenu informatif est mesuré en bit et, dans le cas de l'ADN, varie de 0 à 2 bits. Un contenu informatif nul signifie qu'aucun nucléotide n'est favorisé pour cette position. En effet, dans le cas où la probabilité d'observation de tous les nucléotides est de 0.25, alors le contenu informatif est de 0 bit. À l'inverse, le contenu informatif maximal de 2 bits est obtenu lorsqu'un nucléotide est fixé, c'est à dire que sa probabilité d'occurrence est de 1 tandis que pour les trois autres ont des probabilités nulles. Ainsi les positions hautement conservées et ayant une faible tolérance aux substitutions correspondent à un haut niveau de contenu informatif, tandis que les positions fortement soumises aux substitutions possède un faible contenu informatif [33].

position	1	2	3	4	5	6
A	0	0.8	0.25	0.3	0.1	0
C	1	0.1	0.25	0.3	0	0
G	0	0	0.25	0.2	0	1
T	0	0.1	0.25	0.2	0.9	0
Contenu Informatif	2	1.08	0	0.03	1.53	2

Figure 1.3 – Une PWM représentant la séquence du site de liaison CANNTG ainsi que le contenu informatif associé à chaque position. Un contenu informatif de 2 signifie une forte conservation d'un nucléotide spécifique tandis qu'un contenu information nul signifie une absence d'information quand aux probabilités d'observation d'un nucléotide à cette position.

1.2.3 Représentation visuelle des PWMs

La représentation des motifs sous forme de matrices n'est pas évidente à visualiser. C'est pourquoi on utilise généralement une représentation visuelle des PWMS sous forme de logo [32]. Un logo est une représentation graphique d'une séquence d'ADN constitué de blocs pour chaque lettre et empilés à chacune des positions du motif. La hauteur de chaque empilement est proportionnelle au contenu informatif de la PWM à cette position. La proportion occupée par chaque lettre représente quant à elle la fréquence relative des nucléotides pour cette position. Ainsi un logo donne un aperçu tout à la fois de la séquence, du contenu informatif et de la fréquence des nucléotides. Il est de ce fait privilégié par rapport à la séquence consensus qui offre moins d'information.

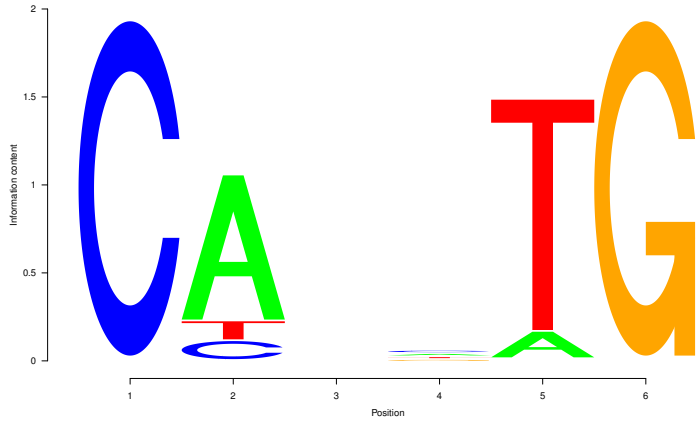


Figure 1.4 – Un exemple de logo correspondant à la séquence des exemples précédent. On peut directement lire le contenu informatif, la fréquence des nucléotides et avoir un aperçu de la séquence consensus.

1.2.4 Base de données de motifs

Plusieurs bases de données répertoriant les motifs d'ADN des sites de liaison connus ont vu le jour ces dernières années. Certaines sont spécifiques à des organismes, d'autres sont plus générales. C'est le cas des deux principales bases de données les plus couramment utilisées : JASPAR [31] et TRANSFAC [26]. Chaque base de données est principalement constituée d'une liste de facteurs de transcription auxquels est rattachée la PWM du site de liaison à l'ADN. Le nombre de motifs dans chacune d'elle ainsi que le contenu des PWMs peuvent différer sensiblement. Les bases de données de motifs permettent ainsi de faire le rapprochement entre un motif d'ADN et le facteur de transcription le plus susceptible de s'y fixer. De part leurs différences, le choix de la base de données est déterminante pour la qualité de l'identification des facteurs de transcription.

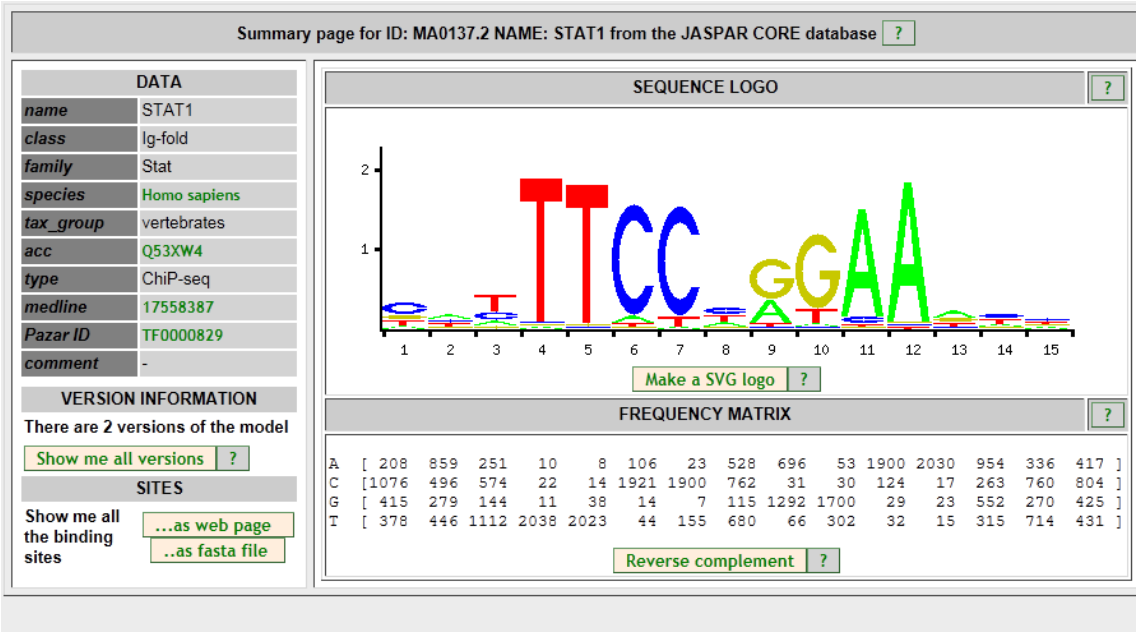


Figure 1.5 – La visualisation des sites de liaison comme proposé par JASPAR. Ici STAT1. Notez la PWM en bas et sa représentation graphique sous forme de logo.

1.3 Principe de l'analyse des données de ChIP-seq

1.3.1 Chip-seq : définition et principes de fonctionnement

La méthode de ChIP-seq (Chromatin ImmunoPrecipitation Sequencing) est une méthode permettant l'étude des protéines interagissant avec l'ADN. Elle combine la technique de Chromatine Immunoprecipitation (ChIP) avec le séquençage haut-débit. Cette méthode est une amélioration et de ce fait remplace de plus en plus la méthode de ChIP-chip (Chromatin ImmunoPrecipitation on chip) qui elle, utilise une puce à ADN pour l'identification des séquences nucléotidiques. De ce fait, la méthode ChIP-seq apporte de nombreux avantages. Le premier et le plus important étant qu'elle offre la possibilité d'étudier l'ensemble du génome d'un individu et ce, pour un coût inférieur à la méthode ChIP-chip et ouvre la voie à l'analyses des interactions inter-géniques à l'échelle du génome. Enfin la méthode ChIP-seq offre une meilleure résolution et permet par là même de détecter des mutations dans les séquences des sites de liaison. La possibilité

d'observer directement les modifications dans les séquences des sites de liaison des facteurs de transcription à donc favoriser l'adoption de cette méthode pour l'analyse de la régulation de l'expression des gènes.[12]

Comme son nom l'indique, la méthode ChIP-seq se décompose en deux parties distinctes. D'un coté l'immuno-précipitation de chromatine et de l'autre le séquençage.

ChIP

La méthode d'immuno-précipitation de chromatine est une méthode permettant la sélection de fragments d'ADN possédant une interaction avec une protéine d'intérêt. Dans un premier temps, la protéine d'intéret est fixé sur son site de liaison grâce à l'utilisation de formaldéhyde. L'ADN est ensuite extrait et découpé en courts brins de 200 à 400 paires de bases, généralement par sonication. Un anticorps spécifique à la protéine étudié est ensuite incorporé afin de former un complexe ADN-protéine-anticorps. L'immunoprécipitation de ces complexes permet ensuite de récupérer les brins d'ADN associé à la protéine après élimination du surnageant (i.e. ADN non associé à la protéine d'intérêt). Pour finir, la protéine d'intéret est séparée de l'ADN, avec de l'ADN ou par chauffage par exemple. Les fragments d'ADN obtenus sont ainsi ceux où la protéine s'est fixée.

Séquençage

La collection de fragments ainsi obtenus est ensuite introduite dans un séquenceur à haut débit. Ce dernier va alors séquencer une ou les deux extrémités des brins d'ADN sur une courte distance, souvent entre 35 et 75 paires de bases. Les séquenceurs actuels permettent de séquencer un génome entier en une seule fois, générant ainsi des millions de séquences. Il est nécessaire ensuite d'utiliser des moyens informatiques afin d'aligner ces séquences sur un génome de référence. Apparait alors des zones fortement séquencées correspondant à une des extrémités des séquences d'ADN collectées. Les sites de liaison des facteurs de transcription sont les zones situées entre deux de ces régions.

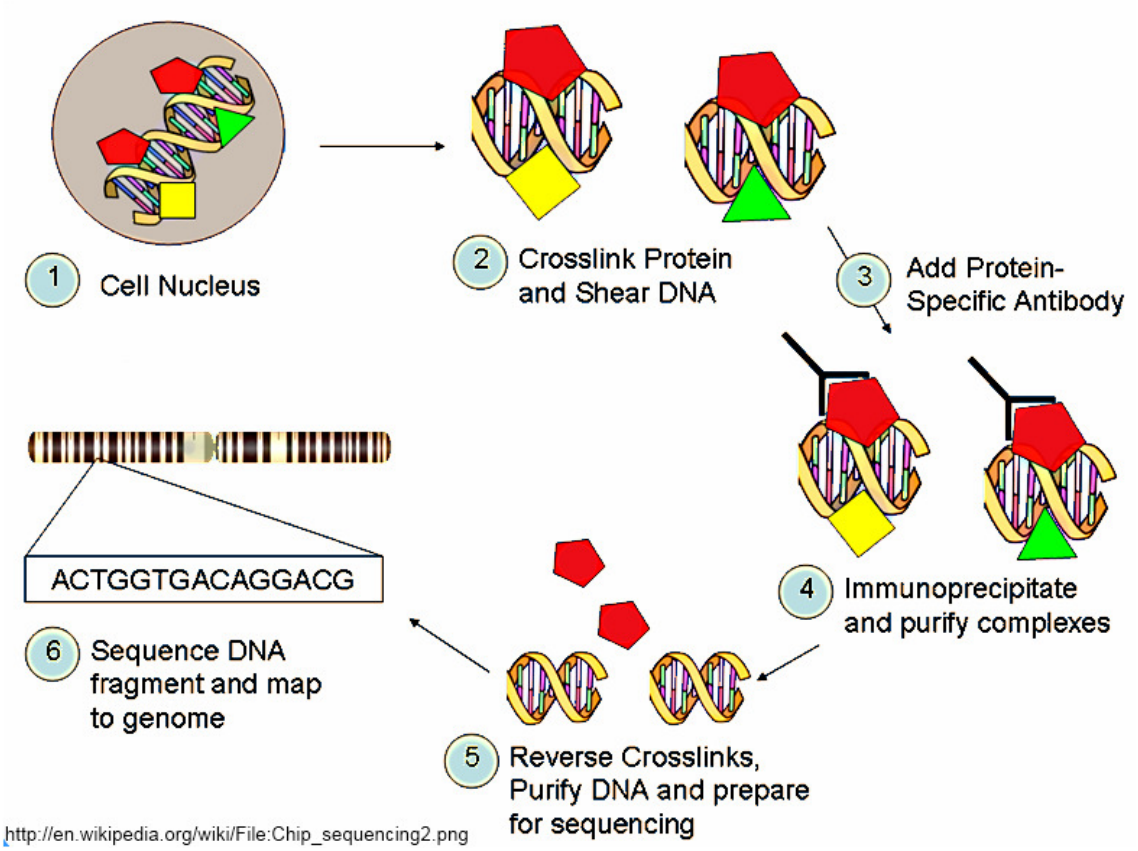


Figure 1.6 – Illustration des principales étapes de la méthode ChIP-seq : 1. Le facteur de transcription est introduit dans le noyau de la cellule 2. Liaison covalente in vivo des protéines à l'ADN et extraction de l'ADN de la cellule puis découpage de l'ADN en courts brins 3. Sélection des fragments associés à la protéine étudiée grâce à un anticorps correspondant 4. Précipitation des complexes ADN-protéine-anticorps, élimination du surnageant 5. Séparation du complexe ADN-protéine pour ne garder que l'ADN 6. Séquençage.

1.3.2 Détection des régions enrichies

Afin de détecter les sites de liaison à l'ADN des facteurs de transcription, il est nécessaire d'identifier les régions fortement enrichies. Comme indiqué précédemment, celles-ci correspondent aux zones avec un fort taux d'alignement de séquences. Ces régions sont souvent appelées *peaks* du fait de l'allure de la courbe de densité à ces endroits, d'où la référence à cette étape de détection des zones enrichies par la dénomination "peak call-

ing" comme illustrer par la figure 1.7. Les analyses sont cependant compliquées par des biais et des artefacts locaux pouvant être introduit lors du séquençage, de l'alignement, par la structure de la chromatine ou encore par des variations dans le nombre de copies d'un gène [1] [39] [30].

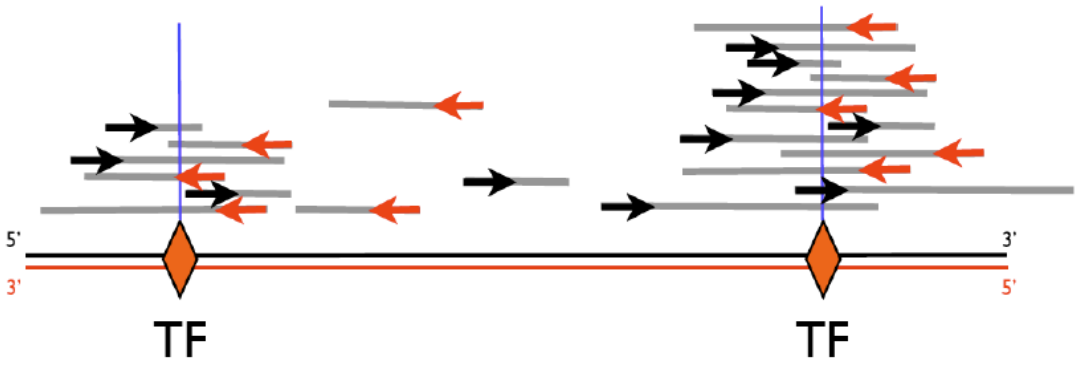


Figure 1.7 – Visualisation de l'alignement des extrémités des brins d'ADN collectés par la méthode ChIP-seq. Les régions densément séquencées permettent de détecter la présence de facteurs de transcription. Les flèches noires et rouges représentent respectivement les extrémités sens et antisens des brins d'ADN (en gris) ayant été séquencés.

Plusieurs logiciels ont été développés ces dernières années pour identifier les régions enrichies. Les principaux étant QuEST [41], PICS [46], MACS [47], CisGenome [18] et FindPeaks [3]. Leur principale différence réside dans la manière de calculer le profile de densités des séquences. En effet, chaque logiciels utilise une méthode de calcul différente : une distribution de Poisson pour MACS, une distribution binomiale pour CisGenome, une densité par noyau (QuEST), PICS fait appel à une distribution de Student alors que FindPeaks n'assume aucune distribution particulière.

Ces différences influent sur la définition de l'étendue des séquences enrichies, chacun des logiciels ne détectant pas le sommet des *peaks* au même endroit. De plus certains algorithmes tirent partis d'expérience contrôles (non enrichies) afin de mieux discerner entre une région fortement séquencés du fait de la présence du site de liaison du facteur

de transcription d'un artefact lié par exemple à une région fortement répétée. Le taux de fausses découvertes FDR (False Discovery Rate) est un autre moyen de contrôler la qualité des régions définies. Il peut se définir basiquement comme étant le rapport du nombre de *peaks* du contrôle et du nombre de *peaks* des tests sélectionnés. Plus il est faible, plus la qualité des résultats est optimale. Cependant, un FDR trop stringant réduit le nombre de *peaks* sélectionnés et peut donc nuire à la qualité de l'analyse. Il faut donc faire un compromis entre FDR et nombre de *peaks*.

Le tableau 1.I résume les différences entre les différents algorithmes [28].

1.3.3 Recherche des motifs surreprésentés

Les logiciels de détection des régions enrichies nous laissent donc avec une liste de séquences dont nous savons qu'elles contiennent préférentiellement le site de liaison à l'ADN du facteur de transcription. Il s'agit donc de faire appel à de nouveaux algorithmes afin d'identifier les motifs les plus présent au sein de la liste des séquences. C'est ce que font les algorithmes tels que MEME [4], GADEM [21], CEAS [45], CisFinder [34], FlexModule [43] et Weeder [29]. Cependant, il ne s'agit pas de détecté une séquence précise puisque le site de liaison du facteur de transcription n'est pas unique mais un patron de celle ci, autrement dit une PWM.

GADEM et MEME comptent sans doute les plus connus et les plus utilisés. Ils utilisent tous deux un algorithme EM (Expectation-Maximization) afin de prédire les PWMs. La différence principale qui les sépare réside dans l'initialisation des PWMs en début de recherche. MEME va en effet procéder à une première analyse en utilisant toutes les sous-séquences possibles contenues dans le jeu de données comme PWMs pour ne garder au final que celles possédant la plus grande vraisemblance d'apparition. Gadem quand à lui travaille avec des dyades espacés, c'est à dire deux motifs séparés par espace variable.

Une autre méthode largement utilisé par d'autres algorithmes est l'échantillonneur de

Logiciel	Profile de densité utilisé	Entendue des séquences enrichies	Utilisation du contrôle	Calcul du FDR
CisGenome	distribution binomial négative ou distribution de Poisson	longueur des séquences définie comme étant égale à la moitié de la distance moyenne entre les *peaks* appareillés	utilisation de l'ensemble des données du contrôle si disponible	contrôle direct par le FDR
FindPeaks	pas de profile de densité spécifique	non disponible	utilise l'ensemble des données du contrôle si disponible	pas de FDR calculé directement
MACS	distribution de Poisson	longueur des séquences définie comme étant égale à la moitié de la distance moyenne entre les *peaks* du brins sens et antisens	utilisation de l'ensemble des données du contrôle si disponible	contrôle direct par le FDR
PICS	distribution t de Student	longueur des séquences pour chaque paires de *peaks* choisie selon la densité du modèle	utilisation de l'ensemble des données du contrôle si disponible	contrôle direct par le FDR
QuEST	estimation par noyau (méthode de Parzen-Rozenblatt)	longueur des séquences définie comme étant égale à la moitié de la distance moyenne entre les *peaks* du brins sens et antisens	une moitié du contrôle est utilisée en tant que référence, une autre moitié est utilisée comme une pseudo analyse ChIP-seq	pas de FDR calculé directement

Tableau 1.I – Résumé des différentes caractéristiques des logiciels de détection des zones enrichies.

Gibbs (Gibbs Motifs Sampler). C'est ce qu'utilise FlexModule implanté dans CisGenome. CisFinder recherche quant à lui les oligonucléotides (courtes séquences d'ADN) surreprésentés, qu'il convertit en PWM, et tente de les étirer en insérant des trous et en étendant les extrémités. Weeder fait lui une recherche exhaustive des k-mers (de 6 à 12) pouvant comporter plusieurs mutations (de 1 à 4) en se basant sur les séquences consensus et retourne les meilleurs occurrences pour chacune de ces longueurs.

Dans chacun des cas, une *E-value* est calculée pour chacune des PWMs trouvées, reflétant sa probabilité d'apparition au sein des séquences par pure chance. Ainsi une faible *E-value* signifie un très faible probabilité d'apparition par chance.

1.3.4 Identification des motifs

La dernière étape de l'analyse consiste à identifier les PWMs trouvées en les comparant à une base de données de sites de liaison de facteur de transcription telles que JASPAR [31] ou TRANSFAC [26]. Cette dernière étape n'a pas bénéficier d'autant de considération que les autres et seuls deux algorithmes d'identification des motifs sont actuellement disponible : TOMTOM [15] et STAMP [25] [24]. Tout deux sont disponible en ligne et bien que des version téléchargables existent, elle ne permettent pas d'exploiter toutes les fonctionnalités nottament les fonctions de visualisation graphique.

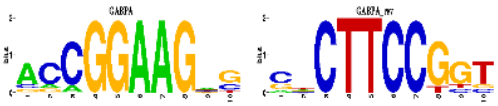
Le principe de fonctionnement est identique pour les deux. Le ou les motifs à identifier sont comparés à chacune des PWMs de la base de donnée et un score d'alignement est calculé pour chaque alignement sous forme d'E-value. Plus l'E-value est faible, plus nous pouvons avoir confiance dans l'alignement. Les 5 meilleurs alignements pour chacun des motifs sont alors renvoyés à l'utilisateur dans une page HTML. A la différence de STAMP, TOMTOM ne permet d'itentifier qu'un seul motif à la fois à moins d'utiliser le fichier de sortie de MEME. De ce fait STAMP s'impose naturellement comme la référence en ce qui concerne l'identification des motifs. Il est rapide (quelques secondes suffisent pour recevoir les résultats), propose un grand nombre de base de données et d'option d'alignement et permet de télécharger le fichier de résultats au format PDF.

Logiciel	**Méthode de détection**	**Classification des résultats**	**Temps de calcul**	**Restriction de la recherche à un motif spécifique**
GADEM	dyades espacées et algorithme EM	Par *E-value*	Plusieurs heures (réduit à quelques minutes pour la recherche d'un motif spécifique)	Oui (v1.3)
MEME	algorithme EM	Par *E-value*	Plusieurs heures à plusieurs jours	Non
Weeder	recherche exhaustive de k-mers	Par longueur de motifs (de 6 à 12) puis les 3 meilleurs motifs selon leur *E-value*	Plusieurs heures	Oui
CisFinder	nombre d'occurrences	100 meilleurs motifs par ordre décroissant d'*E-value*	Quelques secondes	Non
FlexModule	Gibbs Motif Sampler	Par *E-value*	Plusieurs heures	Oui

Tableau 1.II – Résumé des caractéristiques des logiciels de recherche de motifs.

Motif Similarity Matches

GABPA



forward *reverse compliment*

Name	*E value*	*Alignment*	*Motif*
MA0062.1_GABPA	0.0000e+00	SNCTTCCGGT SNCTTCCGGT	
MA0076.1_ELK4	6.6679e-11	SNCTTCCGGT -ACTTCCGGT	
MA0026.1_Eip74EF	9.0176e-10	SNCTTCCGGT --YTTCCGG-	
MA0062.2_GABPA	5.1101e-09	--SNCTTCCGGT NNCACTTCCGG-	
MA0275.1_ASG1	2.2852e-06	SNCTTCCGGT ---WTCCGG-	

Figure 1.8 – Un extrait des résultats d'alignements réalisés par STAMP.

1.4 Motivations

La détection des zones enrichies, la recherche des motifs surreprésentés et l'identification des motifs sont généralement trois étapes réalisées séparément. Cela pose parfois des problèmes de compatibilité notamment parce que chaque programme utilise ses propres formats de fichier. Il est ainsi difficile de passer d'un logiciel à l'autre, par exemple de GADEM à STAMP. Il faut pour cela parcourir le fichier de sortie pour récupérer les PWMs et les transposer dans un format accepté par STAMP. À cela s'ajoute la perte d'information en passant d'une étape à l'autre. Ainsi lorsque l'on fournit les PWMs à STAMP, la localisation des sites de liaison associée à chaque PWM n'est pas conserver.

Il est nécessaire de revenir au fichier de sortie. Des programmes regroupant plusieurs de ces étapes existent bien, tels que MICSA [40], CEAS [45] ou Sole-Search [13] mais ils laissent généralement trop peu de contrôle à l'utilisateur et sont de ce fait peu transparent sur les opérations effectuées. De plus, des analyses complémentaires sont difficile à effectué du fait encore une fois du format de sortie, souvent un fichier texte, HTML ou encore PDF.

Une autre limitation des ces logiciels est qu'ils sont avant tout conçu pour identifier le facteur de transcription principale du jeu de donnée. Or il est connu que les facteurs de transcription possèdent de nombreux cofacteurs. De ce fait, les sites de liaison fonctionnels ont tendances à former des groupes chez les eucaryotes, souvent dénommés module de régulation en cis (CRM pour cis-regulatory modules).

C'est pourquoi nous avons développé un ensemble d'analyse entièrement intégré R [38] autorisant l'analyse et l'identification des sites de liaison des facteurs de transcription et des modules de régulation.

CHAPITRE 2

MÉTHODES

2.1 Ensemble d'analyse

L'ensemble d'analyse que nous avons développé pour l'analyse des sites de liaison des facteurs de transcription peut se décomposer en trois modules réalisant chacun une étape de l'analyse

- PICS permet l'identification des zones enrichies,
- rGADEM recherche les motifs surreprésentés,
- MotIV assure l'identification de motifs et offre des outils facilitant leur validation.

L'ensemble d'analyse est pensé pour mené entièrement l'analyse d'un bout à l'autre mais un utilisateur peut, s'il le souhaite, importer le résultat d'analyses intermédiaires issues d'autres logiciels et peut à tout moment exporter les résultats pour les traiter avec le logiciel de son choix. Ceci rend notre ensemble d'analyse flexible. Néanmoins, nous conseillons d'utiliser chacun des modules PICS, rGADEM et MotIV afin de tirer parti au mieux des fonctionnalités de ceux-ci. Nous avons fortement travaillé à rendre le passage d'un module à l'autre le plus transparent et le plus facile possible pour l'utilisateur.

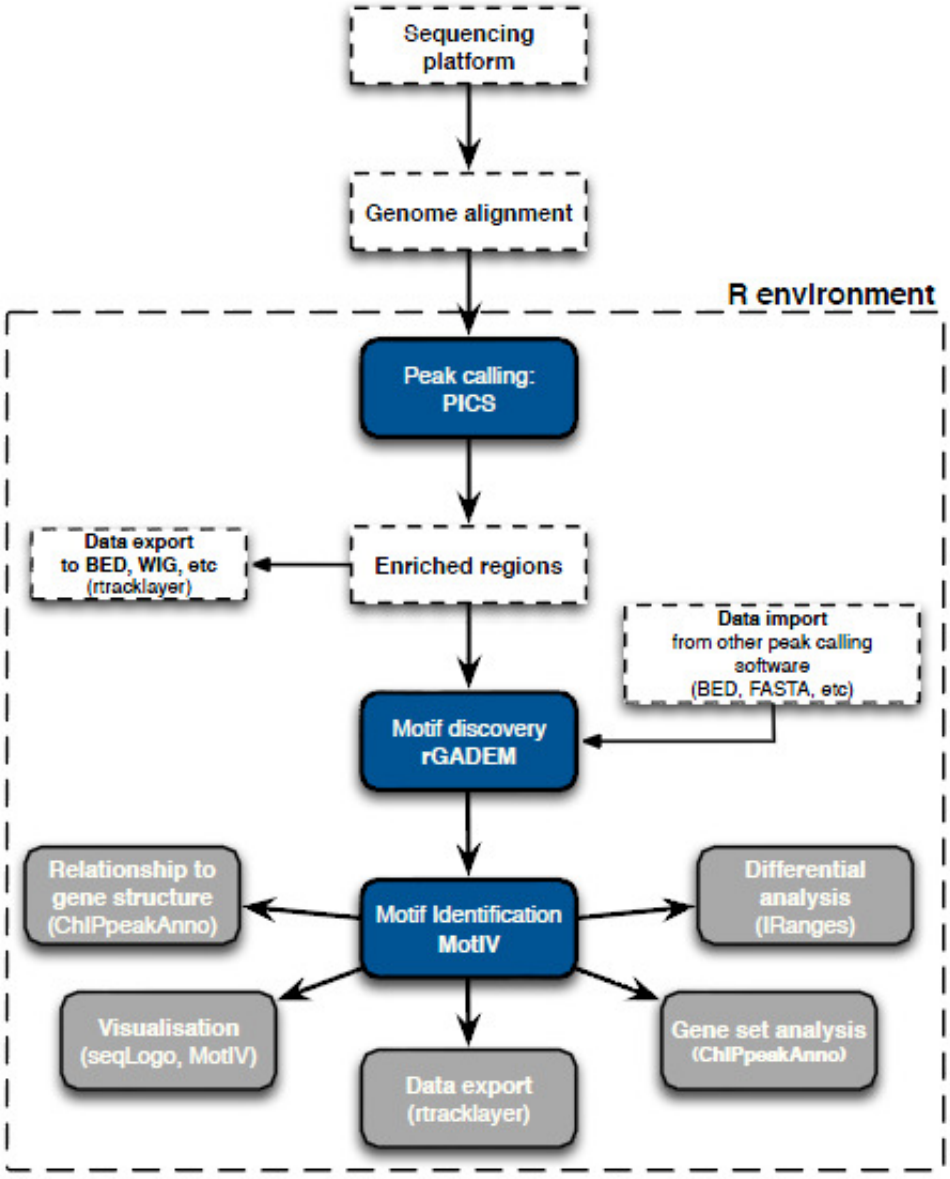


Figure 2.1 – Aperçu globale de l'ensemble d'analyse.

2.2 Architecture et disponibilité des modules

PICS, rGADEM est MotIV ont été développés en langage C et C++ communiquant directement avec le logiciel R. Cela permet de bénéficier d'une interface utilisateur facile d'accès (R) et d'une gestion performante des ressources systèmes à même de gérer les

importants flux de données à traiter. Par ailleurs, rGADEM tire parti de Grand Central Dyspatch (GCD) sur MAC OS X (et plus récent) pour permettre et accélérer le traitement en parallèle un grand nombre de données. PICS autorise le traitement en parallèle grâce aux modules snowfall et multicore.

Chacun des modules adopte les standards de programmation et utilise le system *S4* propre à R pour créer des objets et des méthodes exportables. Cela leur apporte une intégration complète avec les autres modules R et Bioconductor. L'utilisateur peut ainsi exporter les résultats intermédiaires ou finaux dans différent formats tel que un objet *RangedData* utilisable par RtrackLayer, ChIPpeakAnno ou BSgenome ou bien encore des fichiers textes tel qu'un ficher BED reconnu par les outils du site internet UCSC.

Les trois modules PICS, rGADEM et MotIV sont chacun disponibles sur le site de BioConductor BioConductor pour LinuX, MAC OS et MS-Windows. Sous licence GPL, ils sont ainsi librement disponibles à la communauté et peuvent être redistribués gratuitement.

2.3 Jeux de données utilisées

Pour démontrer l'efficacité et la précision que permet notre ensemble d'analyse, nous avons utilisé quatre récents jeux de données de ChIP-seq pour les facteurs de transcription humains suivant :

- le répresseur transcriptionnel CTCF dans des cellules T CD4+ [47],
- STAT1(Signal Transducers and Activators of Transcription) dans des cellules HeLa S3 simulées par interférons gamma [14],
- FOXA1 (Forkhead box protein A1) dans des cellules humaines MCF-7 [47],
- ER (Estrogen Receptor) également dans les cellules MCF-7 [22].

Les jeux de données contiennent :

- 2 947 043 courtes séquences pour les données traitées de CTCF (pas de contrôle disponible)

- 26 731 493 pour STAT1 et 23 435 632 pour le contrôle associé
- 3 909 805 et 5 233 683 pour les données traitées et le contrôle de FOXA1 respectivement
- 3 624 961 et 5 189 798 pour les données traitées et le contrôle d'ER respectivement

2.4 Analyse avec PICS et rGADEM

PICS a été utilisé sur chaque jeu de données afin d'identifier les zones enrichies.

PICS se base sur une loi de Student et fait appel à un modèle bayesien. Il comporte quatre étapes importantes. En premier lieu, il modélise conjointement les profils de densité pour les brins sens et anti-sens (correspondant chacun à une extrémité des séquences d'ADN récupérées par la méthode ChIP). Puis, faisant appel à différent modèles statistiques, il tente de séparer au mieux plusieurs régions enrichies proches. Troisièmement, il utilise la connaissance *a priori* de la distribution des longueurs des fragments d'ADN collectés pour identifier certaines régions enrichies au profil de densité atypique. Enfin, grâce à une pré-cartographie de l'ensemble du génome, il lui est possible d'ajuster le profil de distribution des séquences qui n'aurait pas pu être alignées du fait de la répétitive du génome. PICS permet de fixer le taux de fausse découvertes (FDR) lorsqu'un contrôle est disponible.

Pour les données FOXA1 et STAT1, nous avons sélectionné les meilleurs 15000 *peaks* à la suite de PICS, tout en gardant un faible taux de fausses découvertes (FDR) d'environ 5 à 10% (Voir Annexe ; figures 5.1 à 5.3). En revanche pour les données ER, nous avons du limiter le nombre de séquences à 8000 afin de conserver un FDR similaire. Ne disposant pas de contrôle pour les données CTCF, nous avons pris le parti d'utiliser le même nombre de séquences que dans le cas des données FOXA1 et STAT1, c'est à dire 15000.

Pour la suite, nous nous sommes appuyés sur l'algorithme GADEM, ou plus pré-

cisément un portage de GADEM sur le logiciel R renommé pour l'occasion rGADEM. Alors que certains algorithmes tel que MEME ont besoin de parcourir chaque base de l'ensemble des séquences, demandant ainsi un calcul conséquent et exponentiel avec l'augmentation de la taille des jeu de données tels qu'il est typique d'avoir suite à des résultats de ChIP-seq (jusqu'à plusieurs jours de calcul), GADEM offre des performances remarquables en performant les analyses en quelques heures, voir quelques minutes dans le cas de la recherche d'un motif précis.

Pour cela GADEM, en plus d'utiliser des dyades espacés et un algorithme EM, emploie un algorithme génétique (GA) afin de guide la formation d'une "population" de dyades espacées. Chaque dyade espacée est alors convertit en PWM qui servira à l'initialisation de l'algorithme d'espérance-maximisation. Celui-ci va alors balayer les séquences avec cette PWM en vue d'identifier les sites de liaison en raffinant au fur et à mesure la PWM.

Nous avons identifié avec rGADEM 10 059, 7 105, 8 711 et 3 947 occurences des sites de liaison de CTCF, STAT1, FOXA1 et ER respectivement.

2.5 Identification des motifs à l'aide de MotIV

MotIV (Motif Identification and Validation) est un outil d'identification des facteurs de transcription disponible sous R [38]. Il permet d'identifier les motifs trouvés par rGADEM et offre de nombreux outils permettant à l'utilisateur de gérer, valider et visualiser les résultats.

2.5.1 Implémentation

MotIV est basé sur l'algorithme C++ de STAMP auquel a été ajouté une surcouche sous R ainsi que de nombreuses méthodes étendant ses possibilités. Pour cela, nous avons tiré parti des objets de type SEXP associés à la fonction R *.Call* comme décrite dans le manuel "Writing R Extension" [37]. Cela autorise un dialogue direct entre les

deux langages et la création d'objets en C++ utilisable directement par R. Cela permet de bénéficier de la flexibilité de R tout en profitant de la vitesse d'exécution du code C++. MotIV fait aussi appel à plusieurs packages R notamment *grid* et *lattice* pour les graphiques, *seqLogo* pour l'affichage des logos et *IRanges* pour une gestion efficace des coordonnées chromosomique.

2.5.2 Format d'entrée

MotIV a été développés pour prendre en compte toutes les informations fournis par rGADEM. Nous avons fait en sorte que cet objet comporte le maximum d'information notamment les PWMs détectées par rGADEM ainsi que leurs coordonnées chromosomiques. Cela nous permet donc à la fois de travailler avec les coordonnées globales (i.e. positions sur les chromosomes) et avec les coordonnées au sein des zones enrichies (i.e. positions relatives). C'est pourquoi, si toutes les fonctionnalités sont disponible avec l'utilisation d'un objet rGADEM, il n'en va pas de même avec les autres types d'entré.

En effet, nous avons rendu MotIV compatible avec plusieurs formats de fichiers. Les fonctions *readPWMfile* et *readGademPWMFile* permettent respectivement de lire un fichier au format TRANSFAC et un fichier standard de sortie de GADEM (généralement 'ObservedPWMs.txt'). Cependant, il n'est alors rendu possible qu'une simple identification des motifs, les fonctionnalités avancées de MotIV réclamant l'utilisation d'un objet *rGADEM*. C'est pourquoi nous encourageons vivement les utilisateurs à préférer rGADEM pour leurs analyses.

Pour être compatible avec tous les formats, la fonction principale de MotIV, `motifMatch`, utilise une simple liste de PWMs à identifier. La lecture des fichiers TRANSFAC ou 'ObservedPWMs.txt' permet de récupérer directement cette liste. Pour un objet rGADEM, il est nécessaire d'utiliser la fonction *getPWM* pour extraire la liste des PWMs.

2.5.3 Format de la base de données

Nous avons utilisé la version 2010 de JASPAR pour nos analyses [31]. Le choix de la base de données à utiliser est cependant laissé libre à l'utilisateur et il est possible de changer celle-ci et d'utiliser sa propre base de données. Dans ce cas, l'utilisateur doit fournir un fichier contenant les matrices des sites de liaison au format TRANSFAC.

2.5.4 Prétraitement de la base de données

Avant l'analyse proprement dites des PWMs, il convient de connaitre le biais naturellement présent dans la base de données. Ce biais représente en quelque sorte la variabilité, la diversité de la base de données. Il importe de le calculer car il permettra de corriger l'E-value final. Un score élevé entre deux motifs témoigne d'une importante similarité entre les motifs et de ce fait, il sera plus difficile pour l'algorithme de discriminer un des motifs en particulier. L'E-value finale en sera donc d'autant diminuer, témoignant de l'incertitude de discerner les motifs.

Pour cela, nous avons utilisé la méthode de Sandelin et Wasserman [2] qui consiste à générer un grand nombre de matrices reprenant les propriétés de la base de données puis de réaliser des alignements entre chacune d'entre elles. Pour cela, l'algorithme récupère au préalable toutes les informations de la base de données concernant la longueur des motifs, la somme de chacune des colonnes des PWMs ou encore la distribution des nucléotides. A partir de ces informations, un grand nombre de matrices sont générés aléatoirement respectant les propriétés de la base de données. Par la suite, tous les motifs de longueur n (n de 1 à 30) sont alignés avec chacun des motifs de longueur m (de 1 à 30) afin de calculer le score de similarité moyen entre les motifs des deux longueurs. Nous obtenons au final une table avec le score moyen pour chaque couple de longueur de séquence qui servira à affiner les E-value.

Nous avons calculé le score pour la base de données JASPAR 2010 avec un alignement Smith-Waterman sans trou (SWU) et le coefficient de corrélation de Pearson (PCC) et en générant 10 000 matrices comme préconiser par Sandelin et Wasserman [2]. Pour

Longueur motifs 1	Longueur motifs 2	Score moyen	Variance	Nombre d'alignements	Score minimum	Score maximum
5	5	2.056510	0.5397963	38612	0.4370425	4.897025
5	6	2.147597	0.5652990	32111	0.2739117	4.900521
5	7	2.243243	0.5821044	36642	0.2954110	4.868492
5	8	2.352984	0.5757546	51811	0.2745500	4.867568
5	9	2.446685	0.5770445	54175	0.4154951	4.984638
5	10	2.524560	0.5649965	69935	0.5878786	4.909503

Tableau 2.I – Extrait du fichier contenant les scores d'alignements de JASPAR 2010 utilisant un alignement SWU et le coefficient de corrélation de Pearson. La ligne 2 contient par exemple le score moyen pour tous les alignements des motifs (générés aléatoirement d'après les caractéristiques de la base de données) de longueur 5 contre ceux de longueur 6.

plus de facilité pour les utilisateurs, nous fournissons les scores avec MotIV. Il faut souligner ici qu'il est nécessaire de recalculer le score à chaque modification de la base de données, du type d'alignement et de la métrique utilisée.

2.6 Identification des motifs

La principale étape de MotIV est l'identification des motifs. Il s'agit de déterminer pour chacun des motifs trouvés à l'étape de la recherche de motifs surreprésenté par rGADEM, quel motif de la base de données est le plus similaire.

Pour cela, MotIV réalise un alignement local entre les sites de liaison détectés par rGADEM avec chacune des PWMs de la base de données ainsi que pour son complément inverse. Chaque motif est donc comparé à l'ensemble de la base de données. Un score d'alignement est alors calculé. Ce dernier est ensuite corrigé en utilisant la table des scores associé à la base de données calculé auparavant. Il est de ce fait nécessaire d'utiliser le même type d'alignement et la même métrique que celle utilisés pour générer les scores de la base de données. Par défaut MotIV utilise un alignement Smith Waterman sans trou et le score correspond au coefficient de corrélation de Pearson mais l'utilisateur peut changer à sa guise le type d'alignement et la métrique utilisée.

MotIV sélectionne alors les 10 meilleurs alignements selon leur E-value pour chacune des PWMs à identifier. Le résultat de l'identification est stocké dans un objet R ainsi que différente information telles que l'E-value de chaque alignement, le sens de l'alignement (sens ou anti sens) et l'alignement exacte des deux motifs : celui à identifié et celui de la base de donnée.

	FOXA1	Foxa2	FOXD1	FOXO3	Foxd3
seq	--NAGYAAACAR	YTGTTTRCTN---	YTGTTTRCTN	NAGYAAACAR	--YTGTTTRCTN
match	NWRWGYAAACA-	-TGTTTACWYWNN	NTGTTTAC--	-TGTAAACA-	NAWTGTTTNTTT
evalue	2.4312e-09	3.6164e-07	5.5865e-07	3.2338e-06	2.5432e-04

Figure 2.2 – Un exemple d'alignement réalisé par MotIV issue du jeu de données FOXA1.

2.7 Filtrage des motifs

L'analyse ChIP-seq génère souvent un grand nombre de motifs. Il s'avère parfois difficile de trier parmi les motifs ceux pertinent de ceux résultant d'une séquence répété, du bruit de fond ou d'un artefact. C'est pourquoi MotIV propose un ensemble de filtres permettant de sélectionner les motifs pertinent selon différent critères.

2.7.1 Sélection des motifs

Afin donc de réduire le nombre de motifs, l'utilisateur peut définir des filtres à appliquer sur les résultats de MotIV. Pour cela, il suffit de créer un filtre par la fonction `setFilter` et de choisir les paramètres à appliquer.

Parmi les paramètres disponible, l'utilisateur à la possibilité de sélectionné les motifs selon :

- l'identifiant du motif (d'après son nom dans la liste des PWMs identifiées par rGADEM),
- la présence d'un facteur de transcription parmi les alignements,
- les motifs ne dépassant pas un certain seuil d'E-value,

– la variance maximale de la distribution du site de liaison au sein des séquences,
– et le nombre d'alignements sur lesquels porteront ces paramètres.

Le premier paramètre permet de pointer rapidement un motif particulier, par exemple le motif *m1*. Le second recherche dans la liste des alignements réalisés pour une PWM s'il correspond au nom du facteur de transcription indiqué. Cela permet ainsi de sélectionner les motifs qui auraient été associé au facteur de transcription STAT1 par exemple, sans nécessairement que STAT1 soit le meilleur alignement trouvé. Il est aussi possible de définir des filtres portant sur l'E-value des alignements ou la variance de la distribution des sites de liaison identifié au sein des séquences (voir plus loin) permettant de trier les motifs selon des critères qualitatifs : une faible E-value et une faible variance témoigne d'un motifs correctement identifié. Définir un filtre fixant une E-value maximale est souvent un bon moyen d'éliminer les motifs non pertinents (poly-A, motifs peu conservés,...). Enfin, l'utilisateur peut choisir à combien d'alignements il souhaite restreindre la recherche. Cela permet par exemple de limiter la recherche d'un nom de motifs aux 3 premiers alignements. Les filtres sont appliqués au fur et à mesure aux alignements selon leur E-value. Si un alignement satisfait à toutes les conditions, alors le motif est sélectionné. Sinon, les filtres sont utilisés sur l'alignement suivant et ainsi de suite jusqu'à atteindre le seuil définie par l'utilisateur.

Par exemple nous avons choisi dans un premier temps de ne sélectionner que les motifs dont les trois meilleurs alignements réalisés correspondent à un des facteurs de transcription étudié et possédant une E-value maximale de 10^{-4}. Pour cela, nous avons donc créé deux filtres. Le premier filtrant les motifs n'ayant pas un score d'alignement suffisant. Le second ne gardant que les motifs dont l'un des trois meilleurs alignements correspond à STAT1, CTCF, FOXA1 ou ER dans les jeux de données respectifs. Nous avons en effet définie un motif comme étant le facteur de transcription recherché lorsque le score de son alignement est supérieur à 10^{-4} et que le facteur de transcription voulu apparait dans les 3 premiers alignements.

Il est également possible de créer des filtres plus complexes en créant combinant

plusieurs filtres à l'aide de règles "&" (ET) et "||" (OU). Cela permet de définir des règles plus complexes et autorise à sélectionner des motifs correspondant à un, deux ou plusieurs critères spécifiques.

2.7.2 Regroupement des motifs similaires

Du fait de la variabilité des motifs des sites de liaison des facteurs de transcription, il arrive que les logiciels de recherche de motifs surreprésentés dissocient deux motifs qui au final seront identifiés comme étant le même site de liaison. rGADEM n'échappe pas à cela. L'utilisateur ne veut généralement pas avoir à traiter différent motifs pour un même facteur de transcription. Pour cela MotIV dispose d'une fonction permettant de regrouper différent motifs afin qu'ils apparaissent en tant qu'une seule entité pour l'utilisateur. Cela passe tout d'abord par la création de filtres comme vu précédemment afin de sélectionner les motifs à regrouper. La fonction `combineMotifs` permet alors de fusionner les motifs. Il ne s'agit cependant qu'une fusion virtuelle, les motifs étant traité et apparaissant comme un seul pour l'utilisateur mais restant distinct dans l'objet R. Ceci afin de laisser à tout moment le choix à l'utilisateur de les analyser séparément. Il n'y a donc pas de perte d'information.

2.8 Visualisation des résultats

La visualisation des résultats est une étape importante de l'analyse puisque c'est grâce à cela que l'utilisateur prendra connaissance des résultats. Pour cela MotIV dispose de plusieurs façons d'afficher les résultats, apportant des informations complémentaires et permettant à l'utilisateur d'avoir une vision globale autant que précise des résultats.

2.8.1 Visualisation des alignements

La fonction principale de visualisation de MotIV offre à l'utilisateur un aperçu des alignements réalisés pour chacun des motifs grâce à la fonction `plot`. La fenêtre d'affichage est subdivisée selon le nombre de motifs à afficher. Chaque case correspond à un motif. En haut de celle-ci l'utilisateur retrouve le motif trouvé par rGADEM sous

forme de logo avec à sa droite, son complément inverse. En dessous sont répertorié les logos des meilleurs sites de liaison de la base de données identifié par MotIV, le nom du facteur de transcription associé et l'E-value de l'alignement. Plus l'E-value est faible, plus nous pouvons avoir confiance dans l'alignement. L'utilisateur dispose ainsi d'une vue d'ensemble des alignements réalisés et peut ensuite choisir les filtres à appliquer. MotIV propose également la visualisation des alignements au format texte utilisant les séquences consensus. Ce mode d'affichage permet de visualiser l'alignement exact des séquences mais se montre moins ergonomique.

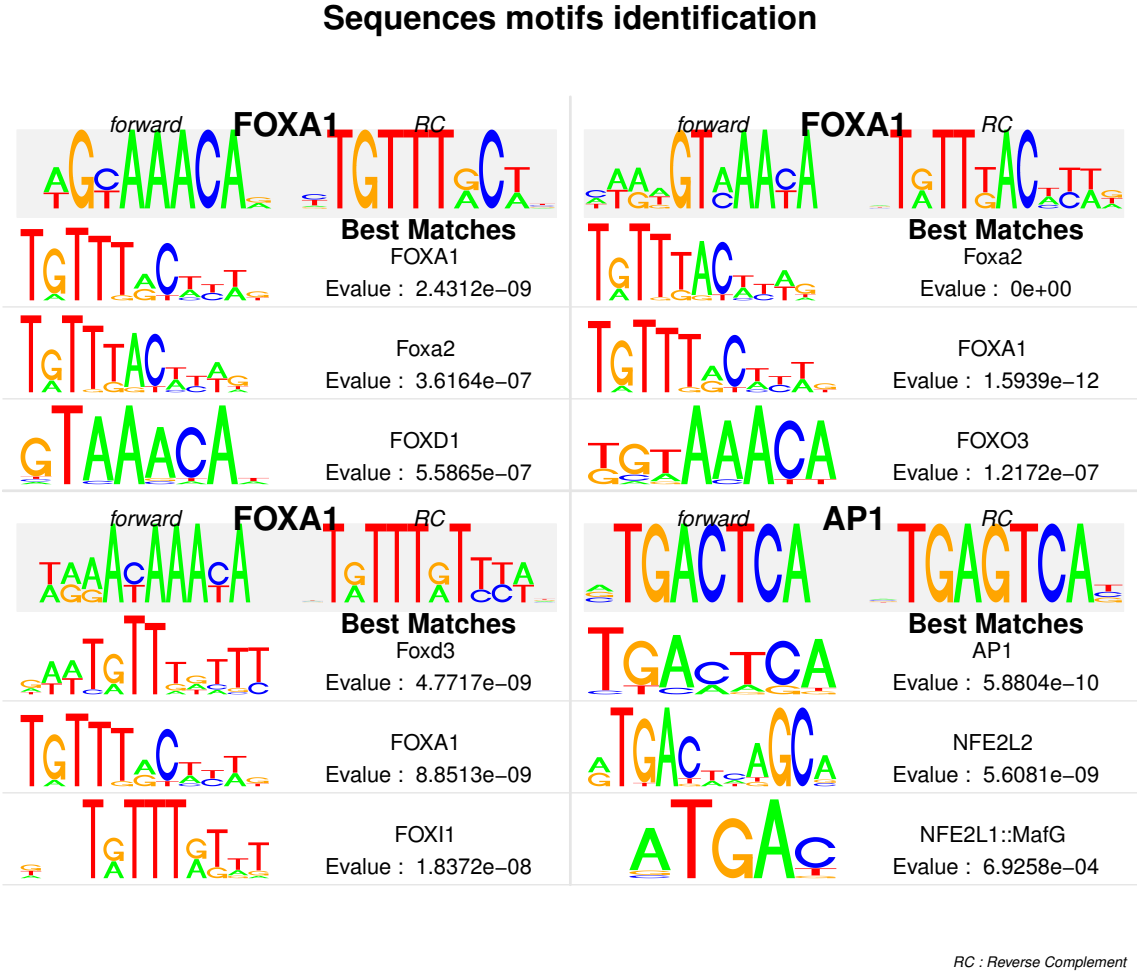


Figure 2.3 – Un exemple de visualisation des alignements en mode graphique tel que le permet MotIV issue du jeu de données FOXA1. Le motif original est dans le cadre gris avec son complément inverse. En dessous est affiché les logos des PWMs de la base de données telles qu'alignées avec le motif. Figure aussi le nom du facteur de transcription auquel correspondent le site de liaison et l'E-value de l'alignement.

2.8.2 Distribution des sites de liaison

Du fait de la répétitivité du génome, la recherche de motifs surreprésentés peut conduire à l'identification de motifs ayant un bonne E-value mais ne représentant pas des motifs intéressant pour l'analyse (poly-A ou séquence peu conservée par exemple). Il apparait que du fait de la spécificité et la précision permise par PICS et rGADEM, les sites

de liaison biologiquement pertinents sont situés majoritairement au centre des séquences dans lesquelles ils ont été identifiés. A l'inverse, les motifs répétés sont répartis tout le long des séquences et possèdent une distribution plate. Ainsi la visualisation de la distribution des sites de liaison est un critère important pour juger de l'intérêt d'un motif. MotIV permet à l'utilisateur d'accéder à cette information très simplement pour chacun des motifs.

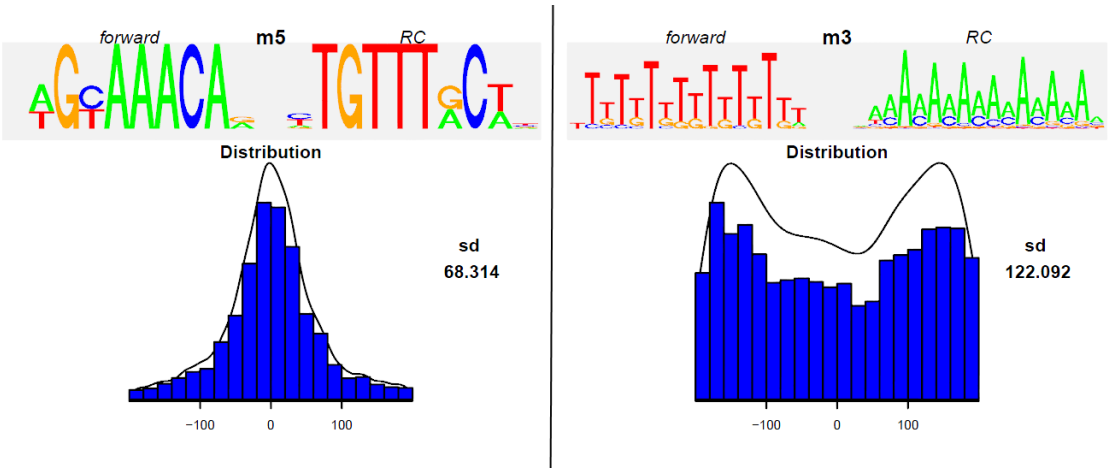


Figure 2.4 – Un exemple de visualisation des distributions des sites de liaison au sein des séquences par MotIV. A droite un motif possédant une distribution piquée au centre, il correspond à un motif identifié comme étant le site de liaison de FOXA1. A gauche, une séquence répétée (poly-A) possède une distribution plate témoignant de la non-spécificité de ce motif.

MotIV utilise à la fois le résultat de l'alignement de motifs avec la base de données et l'objet rGADEM contenant les positions des sites de liaison détecté par rGADEM pour calculer la distribution des motifs. Cette fonction n'est ainsi disponible que lorsque la recherche des motifs surreprésentés a été réalisée par rGADEM. Les résultats sont présentés sous forme graphique, la fenêtre est partagée en cases selon le nombre de motifs présent. Dans chacune l'utilisateur retrouve le logo du motif, la distribution sous forme d'histogramme et de courbe ainsi que la variance de la distribution.

2.8.3 Distance inter-motifs

Finalement, dans le but d'identifier les modules cis-régulateurs, MotIV propose un moyen de visualiser les distances entre les paires de motifs. Ce mode d'affichage indique à la fois le nombre de cooccurrences des deux motifs dans les mêmes séquences ainsi que la distribution de la distance entre ces occurrences. L'idée est de permettre à l'utilisateur de repéré des motifs apparaissant régulièrement côte-à-côte, témoignant d'une possible interaction entre les deux facteurs de transcription.

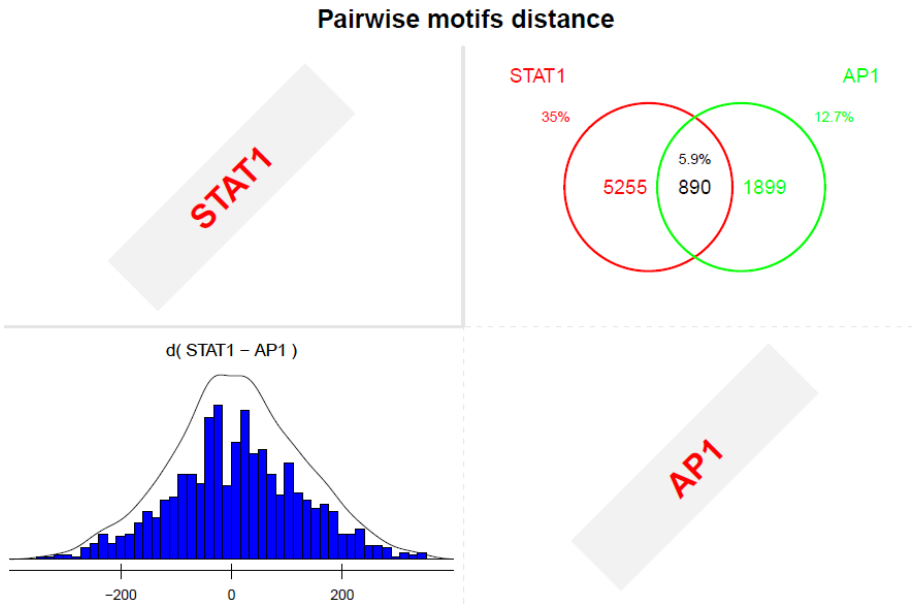


Figure 2.5 – Un exemple de visualisation des distances inter-motifs par MotIV. En haut à droite un diagramme de Venn indique le nombre et la proportion de séquence partageant un des deux sites de liaison ou les deux. En bas à gauche la distribution des distances entre les motifs STAT1 et AP1 lorsqu'ils apparaissent dans les mêmes régions enrichies.

2.9 Comparaison aux autres méthodes

Nous avons comparé notre ensemble d'analyse aux autres logiciels d'analyse de données ChIP-seq disponibles. PICS a déjà bénéficié d'une comparaison favorable par rapport aux autres algorithmes de détection des régions enrichies [46], De ce fait, nous nous sommes concentrés sur la comparaison des étapes de recherche de motifs surreprésen-

tés et d'identification des motifs. STAMP représente actuellement le standard en ce qui concerne l'identification des sites de liaison et puisque MotIV se base sur l'algorithme de STAMP -et approfondissant les outils de validation des résultats- nous avons choisi de conserver MotIV pour la dernière étape des analyses. Au final, nous avons principalement comparé rGADEM aux autres algorithmes de détection des motifs *de novo*. Nous avons décidé d'utiliser MEME, CisFinder, FlexModule et Weeder du fait qu'il s'agit des outils les plus populaires et que leur performances ne sont plus à démontrer. Chacun des logiciels a été utilisé avec les paramètres par défaut suivant les recommandations de leurs auteurs et les analyses ont été faites sur un Mac Pro comportant deux processeurs quadri-coeurs cadencés à 3.2Ghz et 16Gb de mémoire vive.

CHAPITRE 3

RÉSULTATS

Notre ensemble d'analyse a été appliqué aux quatre jeux de données ChIP-seq CTCF, STAT1, FOXA1 et ER. Les 15 000 meilleures *peaks* issues de PICS pour STAT1, FOXA1 et ER et les 8000 meilleures *peaks* pour de CTCF à l'issue de PICS ont ensuite été traitées par rGADEM. La visualisation et l'analyse des résultats ont ensuite été réalisées avec MotIV.

3.1 Identification des motifs primaires

À partir des régions enrichies identifiées par PICS pour chacun des jeux de données, nous avons procédé à la recherche des motifs surreprésentés grâce à rGADEM. rGADEM a identifié 23, 25, 78 et 68 motifs respectivement pour STAT1, FOXA1, ER et CTCF.

Afin de vérifier l'analyse par notre ensemble de l'analyse a été correctement menée, nous avons cherché à identifier dans un premier temps si le site de liaison des facteurs de transcription de chacun des jeux de données a été correctement identifié. Nous avons utilisé pour cela les fonctionnalités de MotIV permettant la sélection de motifs selon différent critères.

Nous avons définie nos propres critères de sélection pour définir qu'un motif correspond à celui recherché. Pour cela, il faut que le site de liaison du facteur de transcription que nous recherchons apparaissent dans l'un des 3 meilleurs alignements et qu'il ait une E-value inférieure à 10^{-4}.

En appliquant ces deux filtres, l'un portant sur le nom du facteur de transcription attendu (STAT1, FOXA1, ER et CTCF) et l'autre fixant l'E-value maximale, nous avons

identifié les motifs d'intérêt dans chacun des jeux de données, avec une très faible E-value (figures 3.5 à 3.8) :

STAT1 1 motif identifié avec une E-value de $3.984^{-}10$

FOXA1 3 motifs identifiés avec une E-value de $2.431^{-}09$

ER 4 motifs identifiés avec une E-value de 0.0

CTCF 1 motif identifié avec une E-value de 0.0

Nous avons ainsi regroupé et renommé les motifs suivant grâce à la fonction `combineMotifs` de MotIV : STAT1=m1, CTCF=m1, FOXA1=m5,m10,m25 et ER=m4,m22,m33,m48.

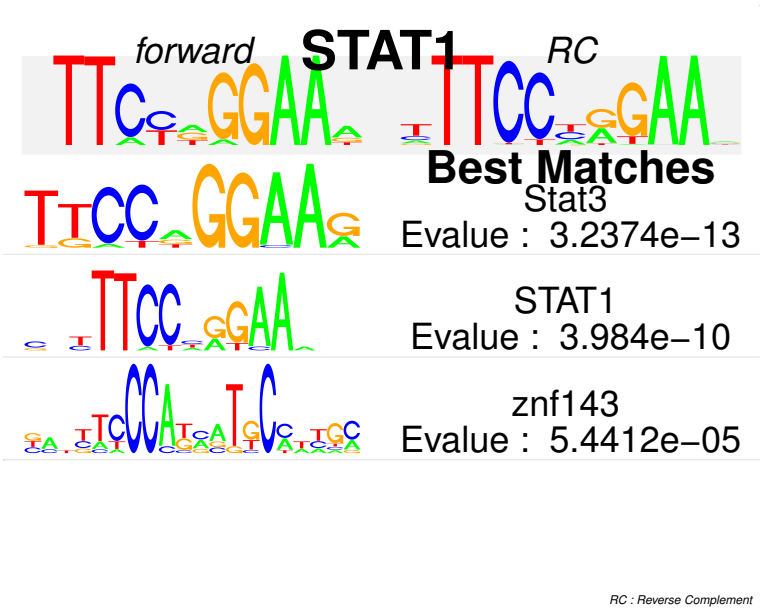


Figure 3.1 – Motif du site de liaison de STAT1 identifié par rGADEM et visualisé par MotIV.

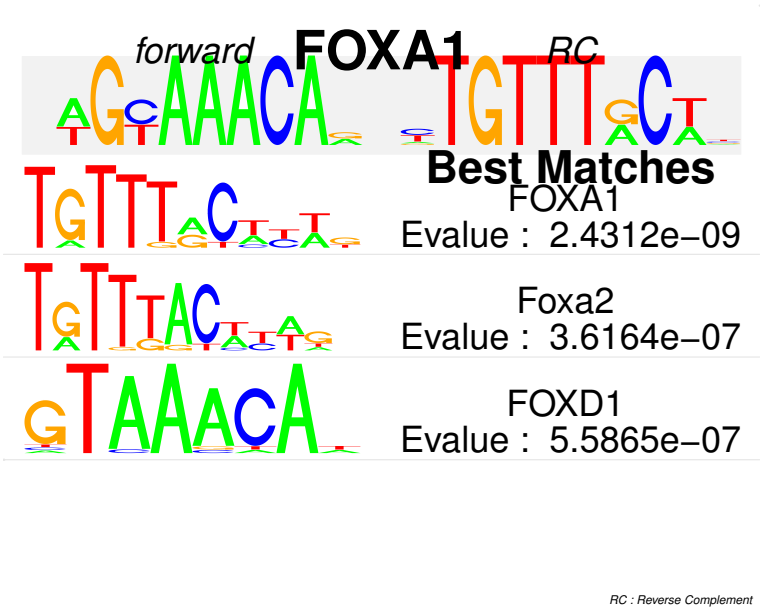


Figure 3.2 – Motif du site de liaison de FOXA1 identifié par rGADEM et visualisé par MotIV.

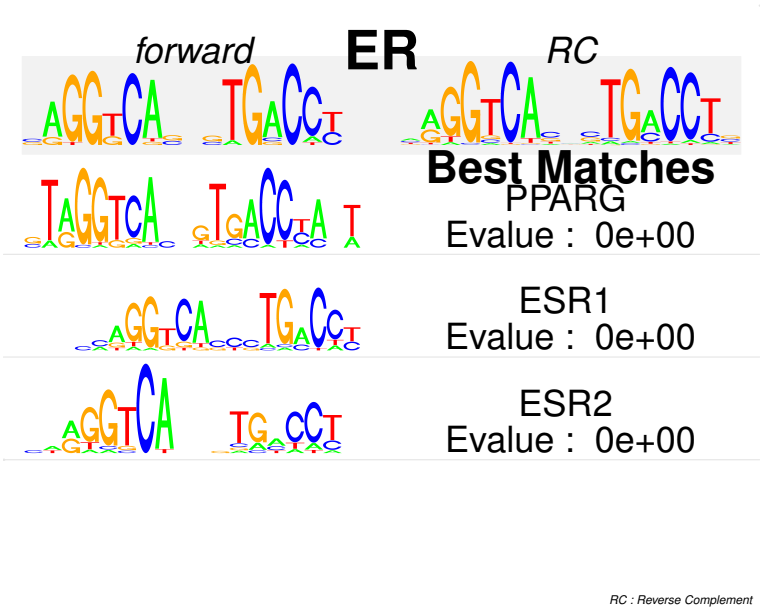


Figure 3.3 – Motif du site de liaison d'ER identifié par rGADEM et visualisé par MotIV.

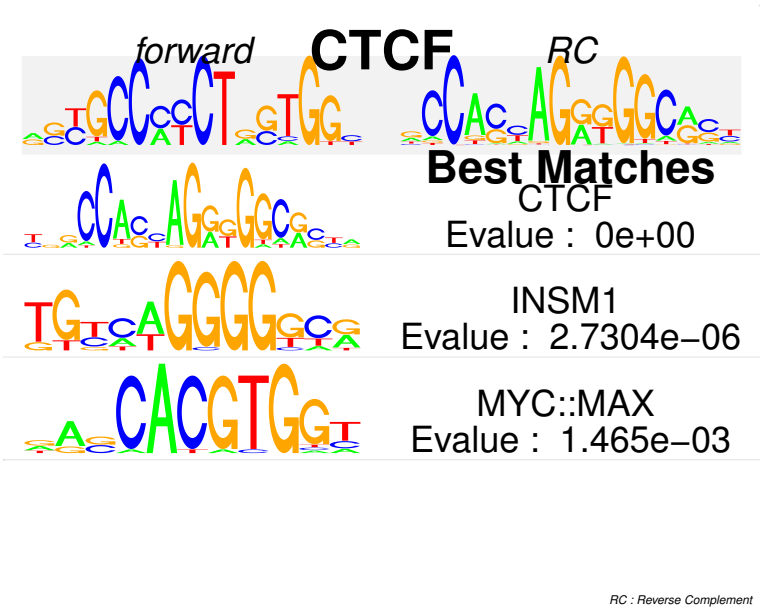


Figure 3.4 – Motif du site de liaison de CTCF identifié par rGADEM et visualisé par MotIV.

La distribution de la position des sites de liaison au sein des séquences nous a également confirmé ces résultats. Il apparait en effet que les motifs identifié comme correspondant au facteur de transcription recherché avaient dans chacun des cas une distribution exceptionnellement piquée en son centre (3.5-3.8).

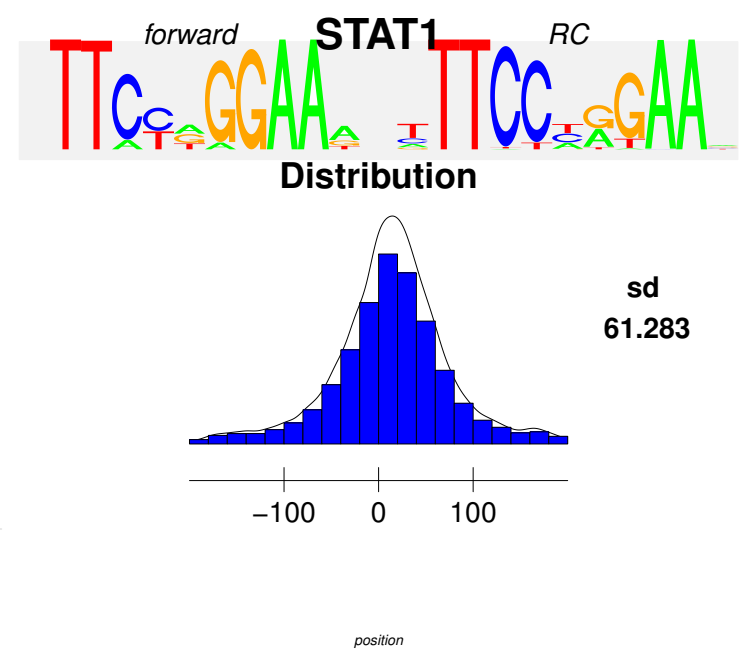


Figure 3.5 – Distribution du motif du site de liaison de STAT1 au sein des séquences et visualisé par MotIV.

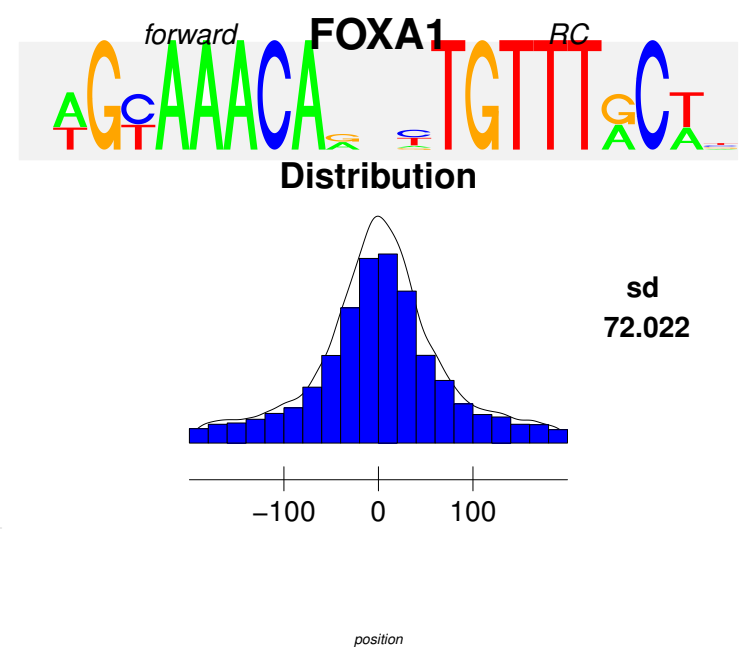


Figure 3.6 – Distribution du motif du site de liaison de FOXA1 au sein des séquences et visualisé par MotIV.

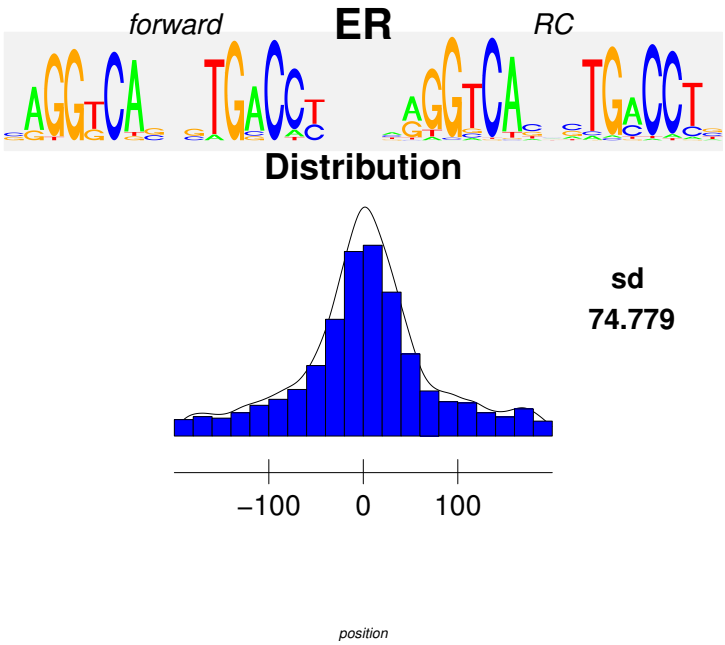


Figure 3.7 – Distribution du motif du site de liaison de ER au sein des séquences et visualisé par MotIV.

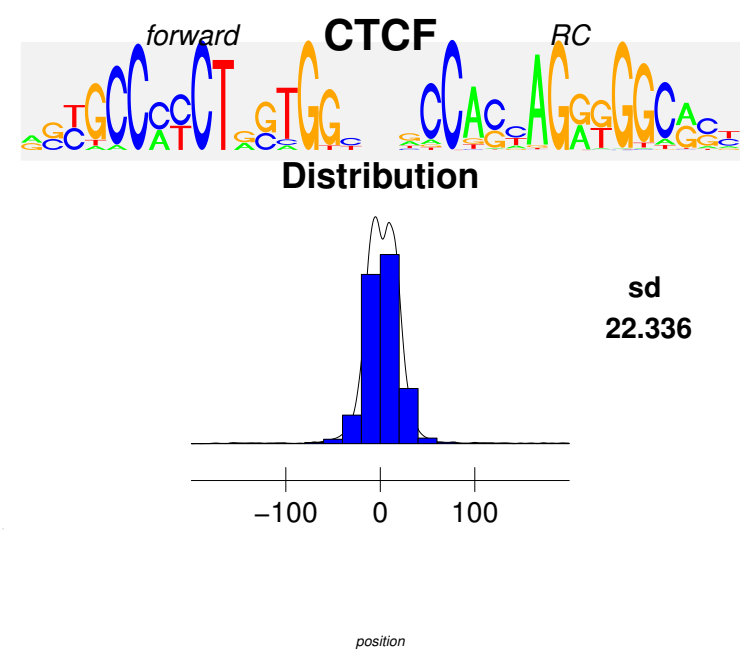


Figure 3.8 – Distribution du motif du site de liaison de CTCF au sein des séquences et visualisé par MotIV.

3.2 Identification des motifs secondaires

Après avoir identifié les motifs primaires correspondant au facteur de transcription attendus, nous avons recherché si d'autre motifs biologiquement pertinents avaient pou être identifié par rGADEM et MotIV.

N'ayant pas de connaissances *a priori* concernant ces autres motifs, nous avons du utiliser une approche différente. Pour cela nous avons à nouveau fait appel à la fonction de visualisation des distributions des sites de liaison. Nous nous attendons en effet à ce que les motifs biologiquement pertinents soient proche du centre des séquences identifiées par PICS. Si un site de liaison se trouve proche de celui de la protéine d'intérêt, cela suggère une possible interaction avec le facteur de transcription principale.

En observant ainsi la distribution des motifs ainsi que le score des alignements, nous sommes parvenus à identifier trois motifs intéressants dans le jeu de données STAT1 que nous avons associé aux facteurs de transcription STAT1, AP-1 et CTCF. Une approche similaire révèle 3 motifs pour les données ER : ER, FOXA1 et AP-1. De même pour FOXA1 nous avons identifié les motifs correspondant à FOXA1 et AP1 et pour CTCF : CTCF et Myf.

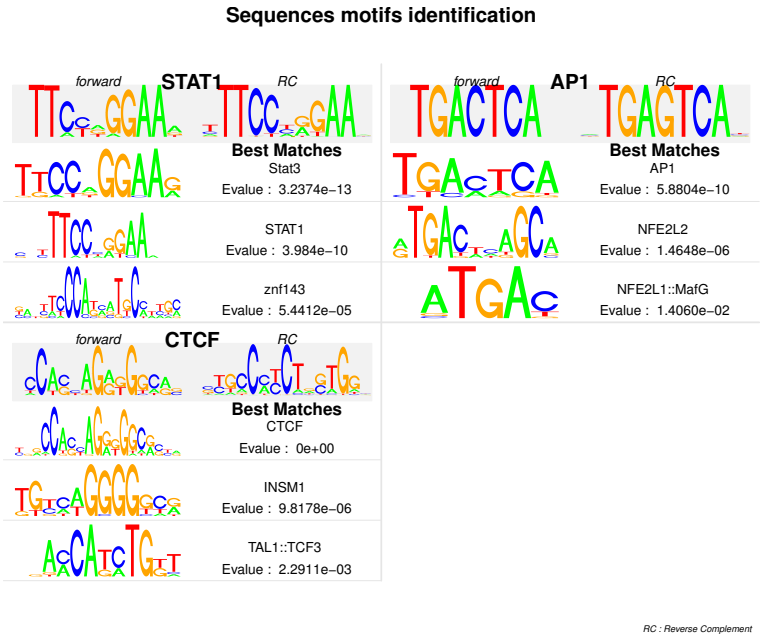


Figure 3.9 – Logo des sites de liaison des meilleurs motifs identifiés dans les données STAT1 par rGADEM et visualisé par MotIV.

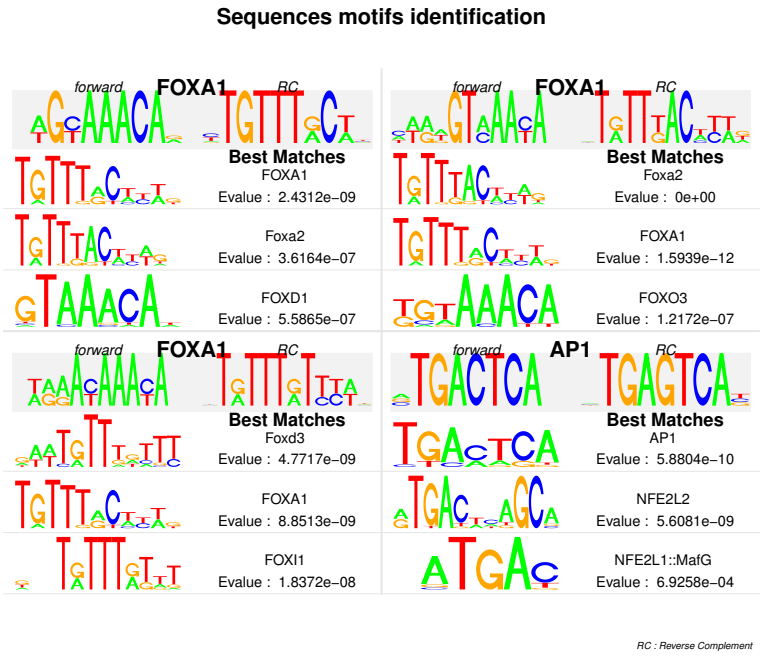


Figure 3.10 – Logo des sites de liaison des meilleurs motifs identifiés dans les données FOXA1 par rGADEM et visualisé par MotIV.

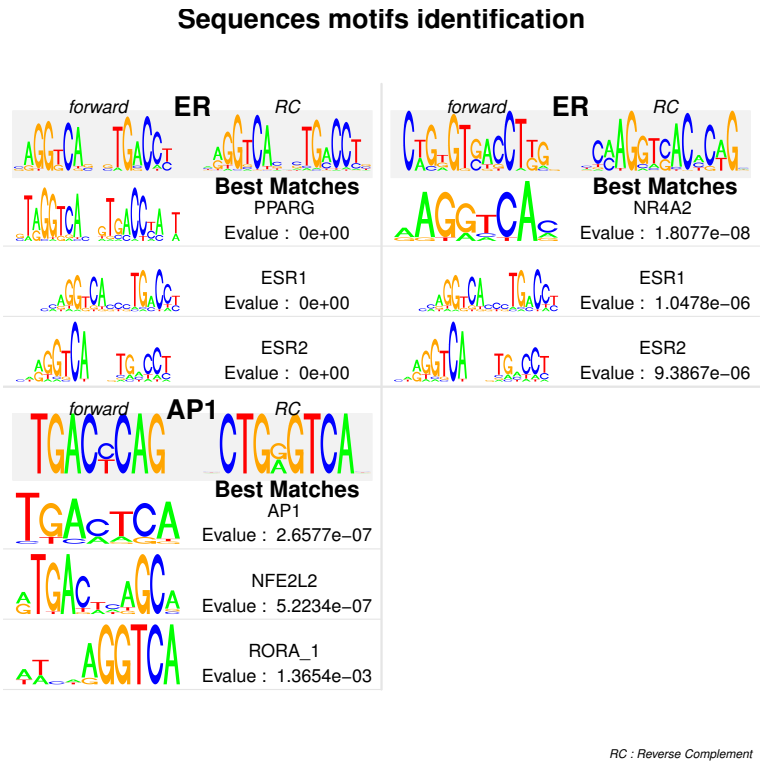


Figure 3.11 – Logo des sites de liaison des meilleurs motifs identifiés dans les données ER par rGADEM et visualisé par MotIV.

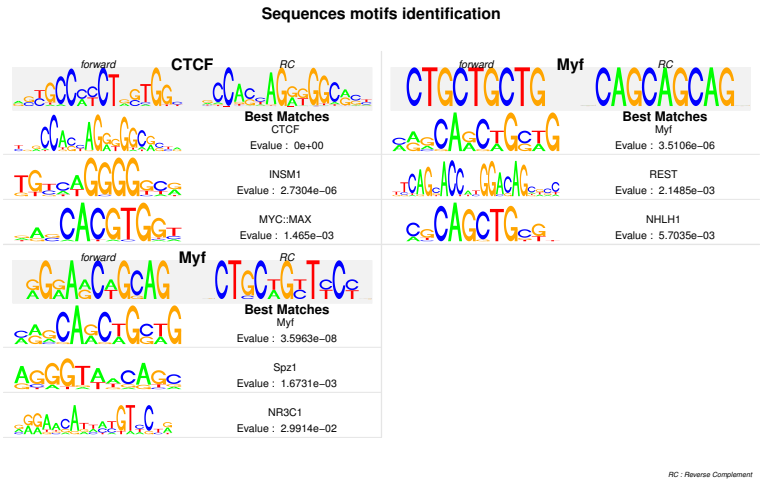


Figure 3.12 – Logo des sites de liaison des meilleurs motifs identifiés dans les données CTCF par rGADEM et visualisé par MotIV.

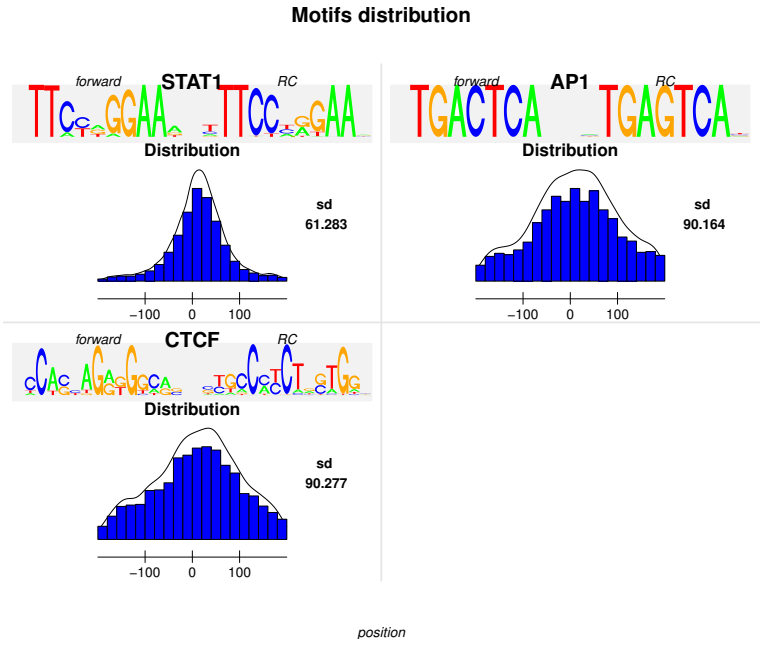


Figure 3.13 – Distribution des sites de liaison des meilleurs motifs dans les données STAT1 visualisées par MotIV.

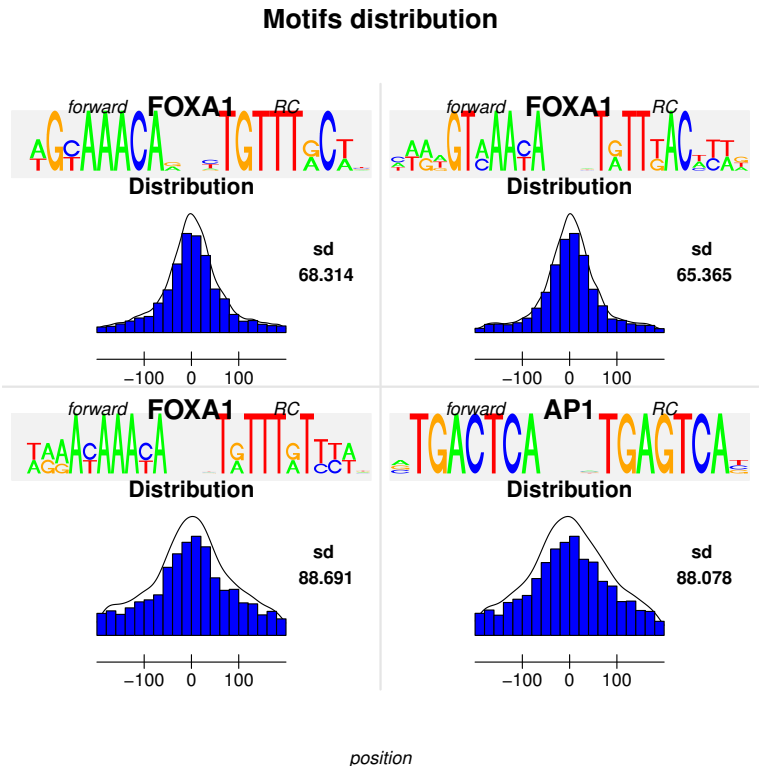


Figure 3.14 – Distribution des sites de liaison des meilleurs motifs dans les données FOXA1 visualisées par MotIV.

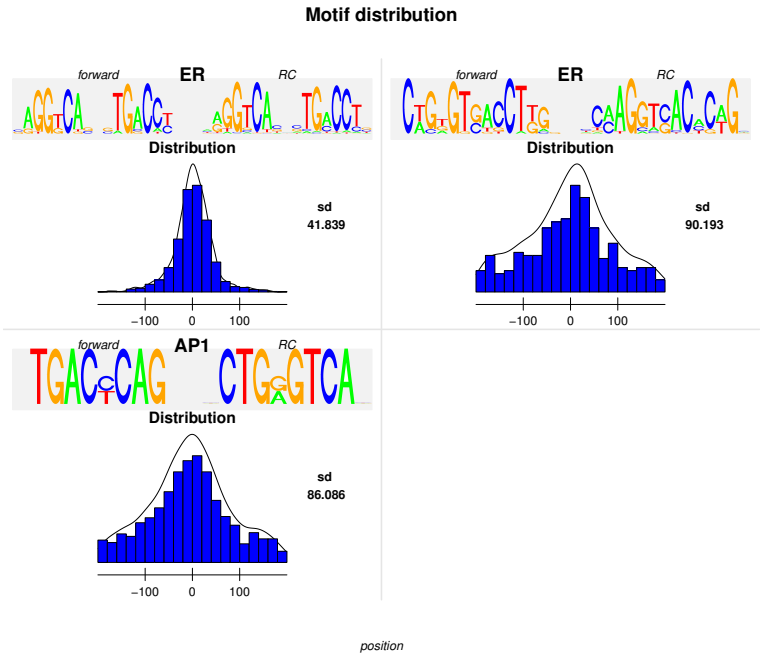


Figure 3.15 – Distribution des sites de liaison des meilleurs motifs dans les données ER visualisées par MotIV.

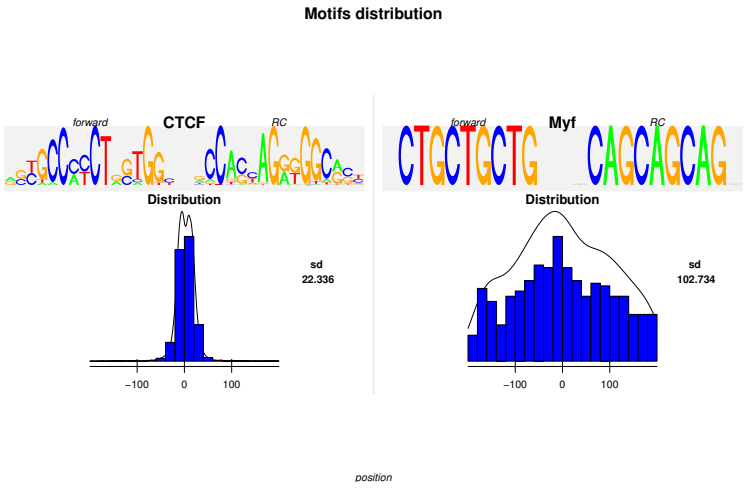


Figure 3.16 – Distribution des sites de liaison des meilleurs motifs da s les données CTCF visualisées par MotIV.

Grâce aux fonctions de MotIV , nous pouvons connaitre le nombre de séquences dans lesquelles deux motifs spécifiques apparaissent ainsi que la distance séparant les deux motifs. Il apparait que le nombre d'occurrences des motifs secondaires est bien moindre que celui des motifs primaires. On remarque cependant que les motifs secondaires co-occurrents près de la moitié du temps avec le motif primaire. Par ailleurs, les sites de liaison des motifs secondaires apparaissent préférentiellement proches (entre 50 et 100 paires de bases) des sites de liaison du motif primaire suggérant une interaction proche des protéines respectives (3.17-3.20),

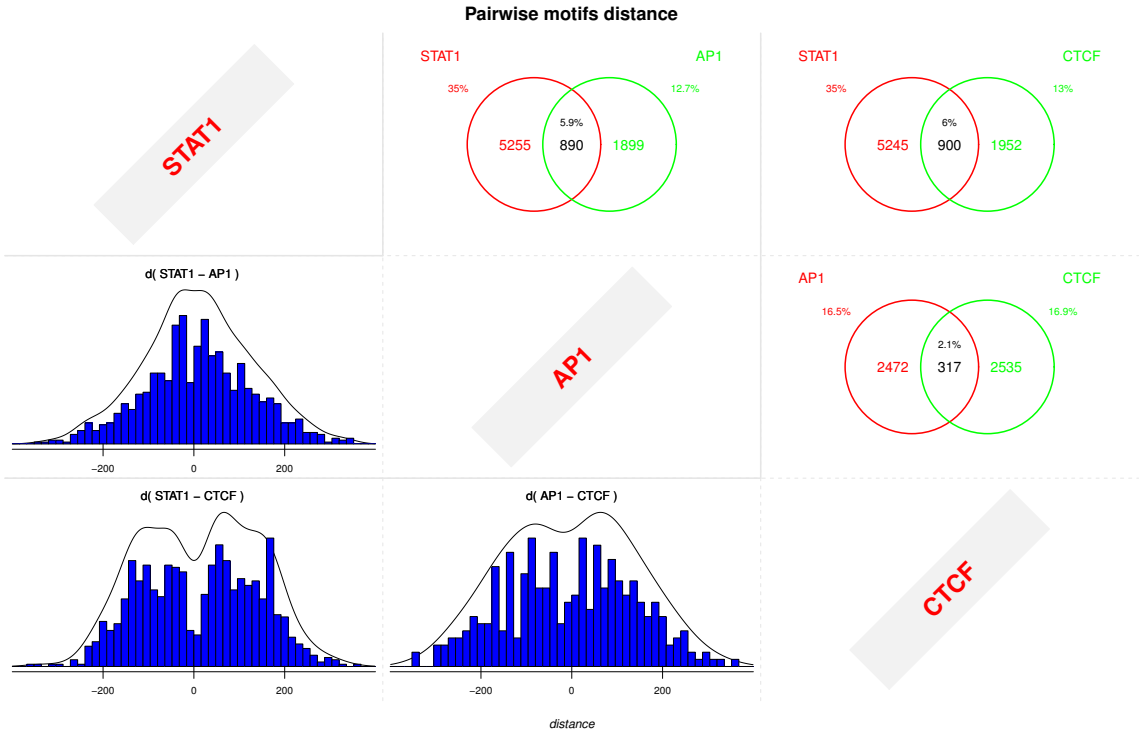


Figure 3.17 – Nombre d'occurrences des sites de liaison de STAT1, AP1, CTCF et distribution des distances entre les motifs STAT1-AP1 et STAT1-CTCF tel que visualisé par MotIV.

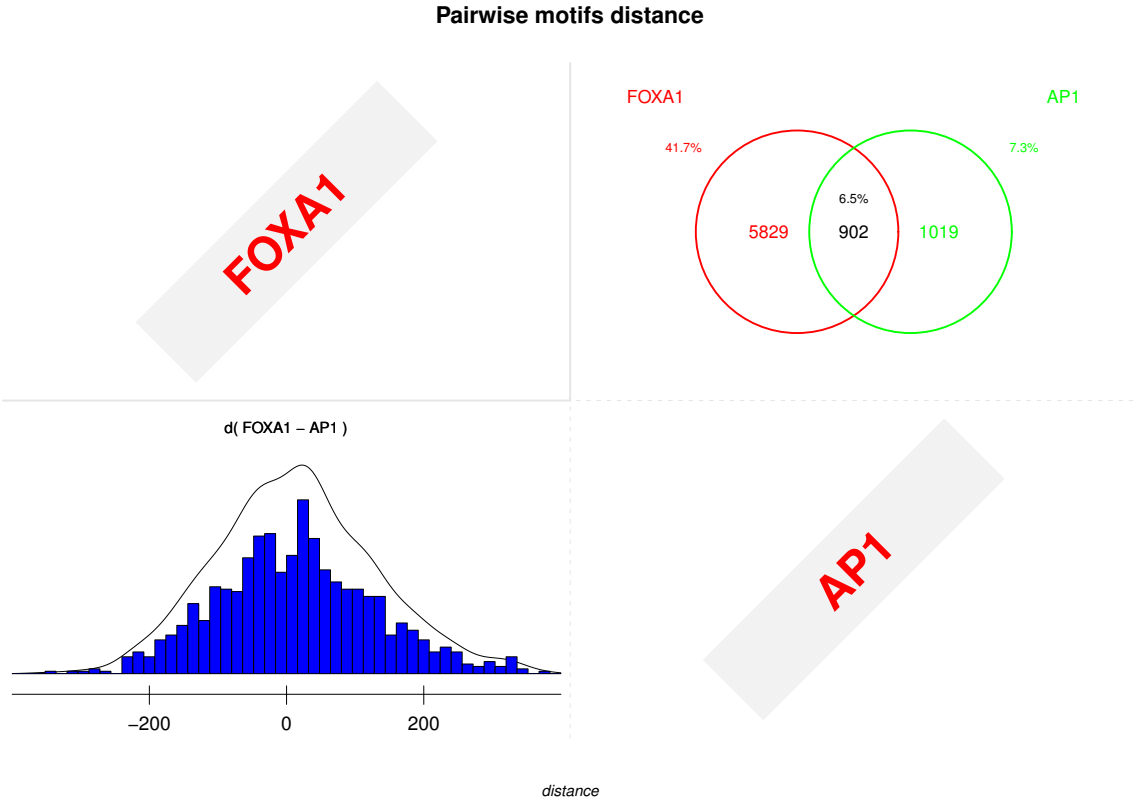


Figure 3.18 – Nombre d'occurrences des sites de liaison de FOXA1 et AP1 et distribution des distances entre les motifs FOXA-AP1 tel que visualisé par MotIV.

-

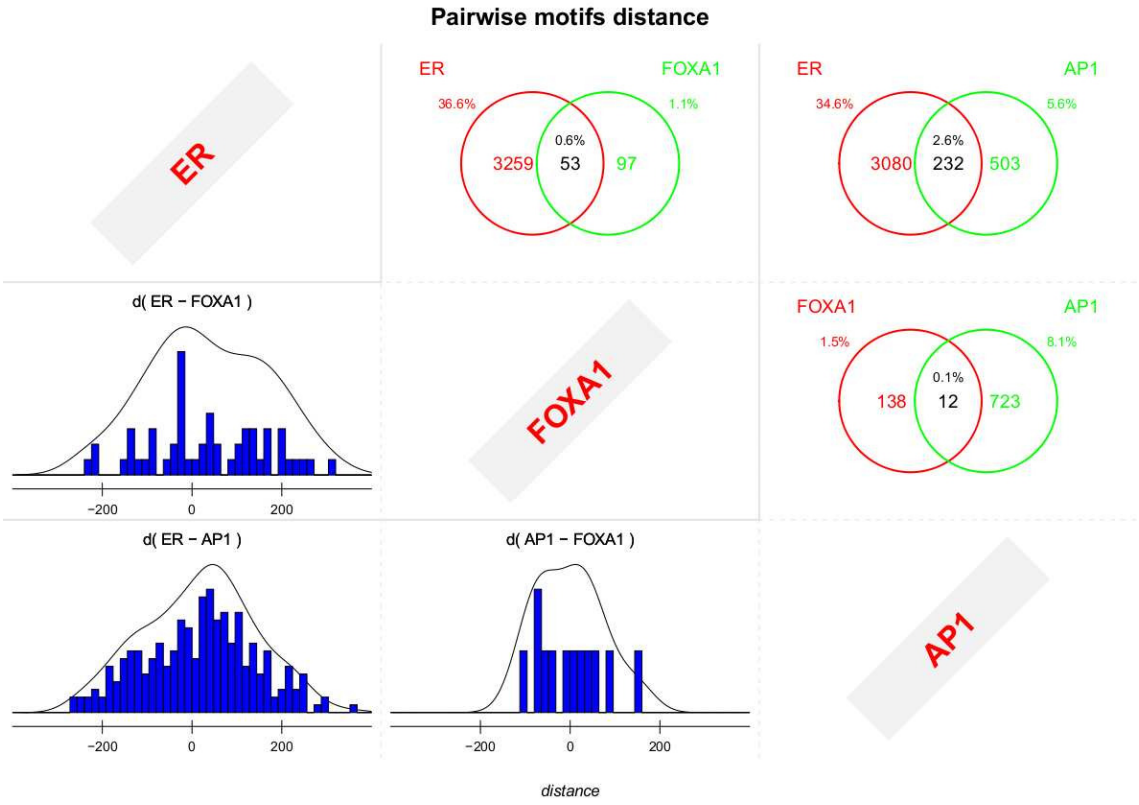


Figure 3.19 – Nombre d'occurrences des sites de liaison de ER, FOXA1, AP1 et distribution des distances entre les motifs ER-FOXA1 et ER-AP1 tel que visualisé par MotIV.

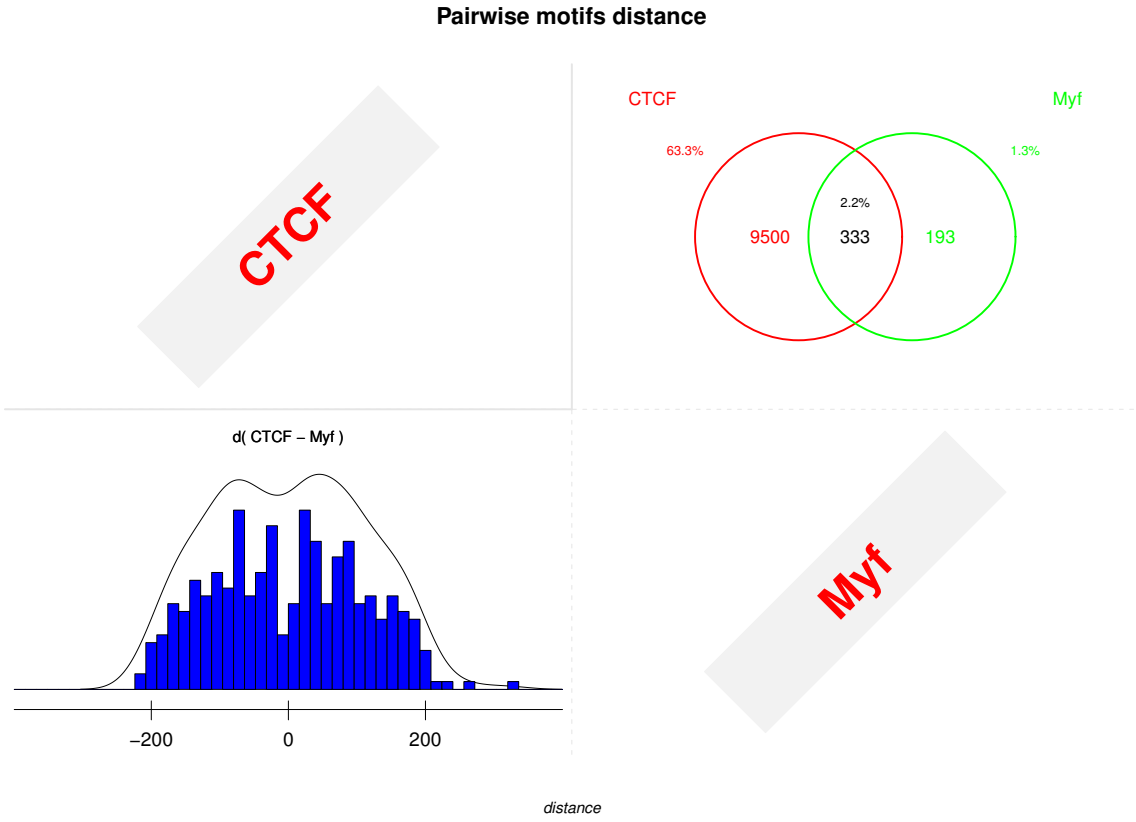


Figure 3.20 – Nombre d'occurrences des sites de liaison de CTCF, Myf et distribution des distances entre les motifs CTCF-Myf tel que visualisé par MotIV.

3.3 Annotation des motifs et des modules

Nous avons effectué des analyses complémentaires afin d'étudier les fonctions régulées par les protéines identifiés. Nous avons pour cela utilisé la base de données de Gene Ontology (GO) [10] à l'aide du module R ChIPpeakAnno [48]. Pour cela nous avons sélectionné les gènes ayant les sites d'initiation de la transcription les plus proches des sites de liaison identifiés et réalisé une étude des fonctions surexprimés par l'ensemble de ses gènes.

Il apparait que concernant le motif primaire de FOXA1, nous avons trouvés préférentiellement des gènes liés à la régulation et au mouvement cellulaire. Pour ER, nous avons avant tout détecté des gènes participant à la régulation positive de la réplication d'ADN et au développement des cellules musculaire. Les gènes associé au motif STAT1 gère principalement le développement des vaisseaux sanguins et des la régulation lymphocytes T. Quand à ceux proches des sites de liaison de CTCF, ils participent à la régulation de l'apoptose et à la communication intercellulaire (3.I-3.IV).

La même analyse a été réalisée pour les modules identifiés. Pour le module formé par ER et FOXA1 nous avons identifié des gènes associés à l'activité des facteurs de transcription tandis que pour les gènes proches du module ER-AP1 contrôlent les processus cellulaire notamment l'activité des facteurs de croissance. En ce qui concerne le module FOXA1-AP1, les fonctions surexprimées concernent la régulation de la transcription de l'ARN polymérase. Enfin, l'analyse des termes GO associé aux motifs STAT1-CTCF révèle une activité localisée dans les vaisseaux sanguins et le développement vasculaire ainsi que dans les composants de la membrane cellulaire.

Tout cela concorde avec les rôles biologiques connus pour les facteurs de transcription STAT1, FOXA1, ER et CTCF. L'association de CTCF avec STAT1 se révèle particulièrement intéressante car elle n'apparait pas ailleurs dans la littérature. Il faudrait cependant des études plus poussées pour comprendre les interactions de ces deux pro-

téines.

	STAT1	STAT1 – AP1	STAT1 – CTCF
Biological process	• blood vessel and vasculature development • regulation of T cell proliferation, T cell tolerance induction		• blood vessel development, vasculature development
Cellular component	• cell-substrate junction, adherens junction, cell leading edge	• plasma membrane part, catenin complex	• membrane fraction

Tableau 3.I – Analyse des fonctions surexprimées pour les gènes proches des sites de liaison de STAT1.

	FOXA1	FOXA1 – AP1
Biological process	• regulation of cell migration and cellular component movement	• regulation of transcription from RNA polymerase II promoter
Molecular function		• transcription factor activity • sequence-specific DNA binding

Tableau 3.II – Analyse des fonctions surexprimées pour les gènes proches des sites de liaison de FOXA1.

	ER	ER – AP1	ER – FOXA1
Biological process			• cell death • regulation of muscle tissue development • regulation of retinoic acid receptor signaling pathway
Molecular function	• transcription factor activity • muscle cell development and activityt • regulation of DNA replication	• growth factor activity	• transcription factor activity • estrogen receptor activity
Cellular component	• contractile fiber, myofibril, actin cytoskeleton • axon, cell projection part	• cytoplasm, mitochondrion • cell projection membrane	• neuron projection terminus • vesicle, Golgi membrane

Tableau 3.III – Analyse des fonctions surexprimées pour les gènes proches des sites de liaison d'ER.

	CTCF	CTCF – Myf
Biological process	• regulation of apoptosis • regulation of cell cummunication	• blastocyst development • regulation of cyclase activity
Molecular function		• growth hormone-releasing hormone receptor activity • protein complex binding

Tableau 3.IV – Analyse des fonctions surexprimées pour les gènes proches des sites de liaison de CTCF.

3.4 Signification biologique des modules

Nous avons par la suite étudié la littérature afin de confirmer les interactions misent en évidence par notre ensemble d'analyse.

Il apparait ainsi que plusieurs études mettent en avant le rôle important que joue FOXA1 pour qu'ER puisse se fixer à l'ADN [7] [16] [23]. Il faut noter également que FOXA1 est lui même exprimé suite à un traitement par estrogène [7] [16]. Cette interaction entre les deux protéines confirme donc que la découverte du site de liaison de FOXA1 par rGADEM parmi les données d'ER n'est pas un hasard.

AP1 est un facteur de transcription hétérodimérique composée de deux protéines de la famille c-Fos et c-Jun [9] [42] [35]. AP1 est connus pour être surexprimés dans les protéines sensible aux estrogène dont fait parti la lignée MCF7 et pour pouvoir interagir directement avec le facteur de transcription ER [36] [20]. Ceci explique la présence du motif AP1 identifié par rGADEM dans les données ER. Une possible explication de la présence du site de liaison d'AP1 dans les séquences enrichies de FOXA1 par une interaction intermédiaire avec ER.

Étant données que nous avons identifié le motif du site de liaison de FOXA1 dans les données ER, nous nous attendions à retrouver inversement le motif d'ER parmi les séquences enrichies de FOXA1 ce qui n'a pas été le cas. Une analyse avait déjà été mené pour tenté d'identifié le motif ER dans le même jeu de données sans succès [47]. Une recherche spécifique utilisant la PWM d'ESR1 issue de JASPAR a permis d'identifier seulement 723 motifs correspondant à ER. Une possible explication de ces résultats est qu'ER a besoin de FOXA1 pour se fixer à l'ADN mais que l'inverse n'est pas vrai, ce qui est cohérent avec la littérature. Il est intéressant de noter que ER est associé à la fois à FOXA1 et AP1 dans le jeu de données de ER. FOXA1 et lui même associé à AP1 dans le jeu de donné de FOXA1. Nous n'avons cependant pas pu observer d'interaction significative entre FOXA1 et AP1 dans les données de ER.

La découverte du motif AP1 dans les données de STAT1 trouve une explication dans la littérature. Des études rapportent que le facteur de transcription STAT1 est capable d'activer leur propre phosphorylation suite à une stimulation des cytokines [27] [6] [5]. Par la suite, il migre vers le noyau cellulaire et se lie à l'ADN afin d'activer la transcription des gènes [17] [19]. Parallèlement les cytokines activent différentes voies de signalisation intracellulaires incluant notamment le facteur de transcription AP1 formé de Fos et Jun et activant une interaction directe entre AP1 et STAT1 [44].

Nous n'avons pas trouvé d'évidence d'interaction entre CTCF et les autres protéines STAT1 et Myf, le facteur de transcription CTCF n'ayant pas une littérature le concernant très fournie.

3.5 Comparaison de notre méthode

Dans le but de démontrer les performances de notre ensemble d'analyse, nous avons comparé notre méthode rGADEM aux autres outils de recherche de motifs cisFinder, Weeder, MEME et FlexModule pour chacun des quatre jeux de données. Nous avons conservé PICS pour la détection des régions enrichies et MotIV pour l'identification et l'analyse des résultats.

Le nombre de motifs identifié varie grandement d'un algorithme à l'autre comme le montre le tableau 3.V. Comme attendu, chaque méthode a identifié avec succès le site de liaison de la protéine d'intérêt de chaque jeu de données. De plus une majorité des motifs identifiés sont communs à ceux détecté par rGADEM (3.21-3.24). De même le nombre de motifs primaires identifié croit globalement de la même façon au fur à mesure de l'analyse des séquences enrichies (Annexe ; figures 5.4-5.7). La principale différence concerne en fait le nombre de motifs secondaires identifiés par les différents algorithmes, Weeder et CisFinder identifiant moins de motifs que les autres. On remarque ici que rGADEM identifie le plus de motifs dans chacun des jeux de données.

En termes de temps de calcul, CisFinder s'est montré le plus rapide en réalisant les analyses en quelques secondes. Cela contraste fortement avec Weeder et FlexModule pour qui les analyses ont duré jusqu'à plusieurs jours. rGADEM bénéficie quant à lui d'un temps de calcul de quelques heures grâce à la possibilité de gérer les calculs en parallèles sur plusieurs processeurs.

	CTCF	ER	FOXA1	STAT1
rGADEM	CTCF (0)	ER (0)	FOXA1 (2e-12)	STAT1 (3e-13)
	Myf (4e-8)	FOXA1 (5e-12)	AP1 (6e-10)	CTCF (0)
		ETS-like (1e-8)		ETS-like (9e-7)
		AP1 (3e-7)		AP1 (6e-10)
cisFinder	CTCF (0)	ER (0)	FOXA1 (4e-13)	STAT1 (2e-10)
		ETS-like (9e-8)		AP1 (9e-8)
		AP1 (8e-3)		
Flexmodule	CTCF (0)	ER (0)	FOXA1 (3e-11)	STAT1 (4e-11)
		FOXA1 (1e-13)	AP1 (4e-8)	SRF (1e-8)
				AP1 (3e-8)
Weeder	CTCF (2e-11)	ER (1e-14)	FOXA1 (1e-12)	STAT1 (2e-11)
				AP1 (1e-10)
				ETS-like (2e-8)
MEME	CTCF (0)	ER (0)	FOXA1 (2e-15)	STAT1 (5e-9)
		AP1 (3e-4)		ETS-like (1e-5)
				AP1 (4e-4)

Motifs identified by all compared methods in the selected PICS enriched regions. The number given between parenthesis is the E-value match to the corresponding JASPAR motif.
doi:10.1371/journal.pone.0016432.t001

Tableau 3.V – Comparaison des motifs identifiés par différentes méthodes de détection des motifs surreprésentés : rGADEM, CisFinder, Weeder, FlexModule et MEME.

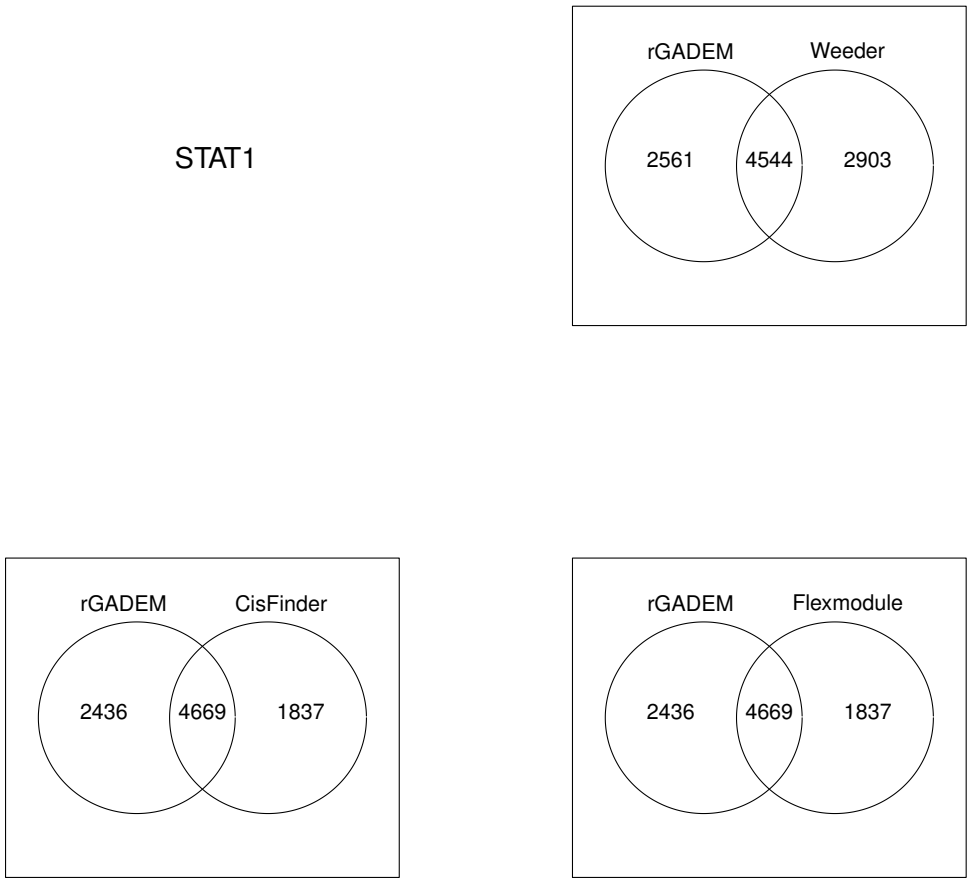


Figure 3.21 – Diagrammes de Venn montrant les motifs STAT1 communément identifiés par rGADEM, CisFinder, FlexModule et Weeder.

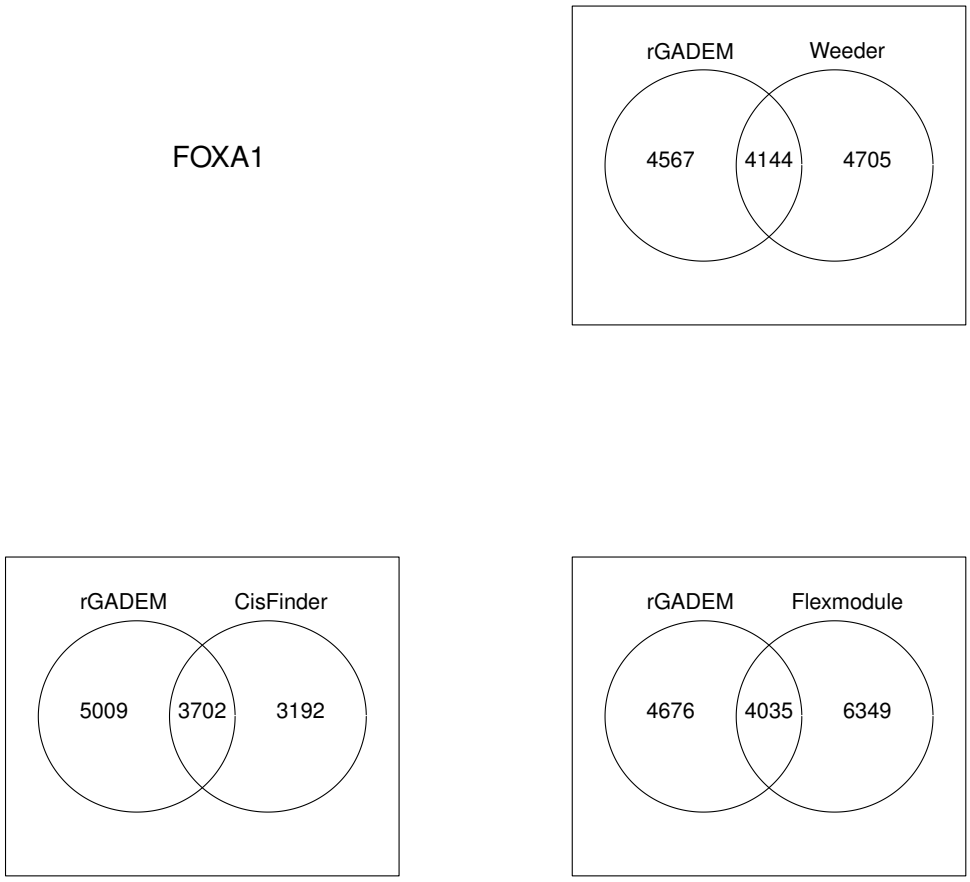


Figure 3.22 – Diagrammes de Venn montrant les motifs FOXA1 communément identifiés par rGADEM, CisFinder, FlexModule et Weeder.

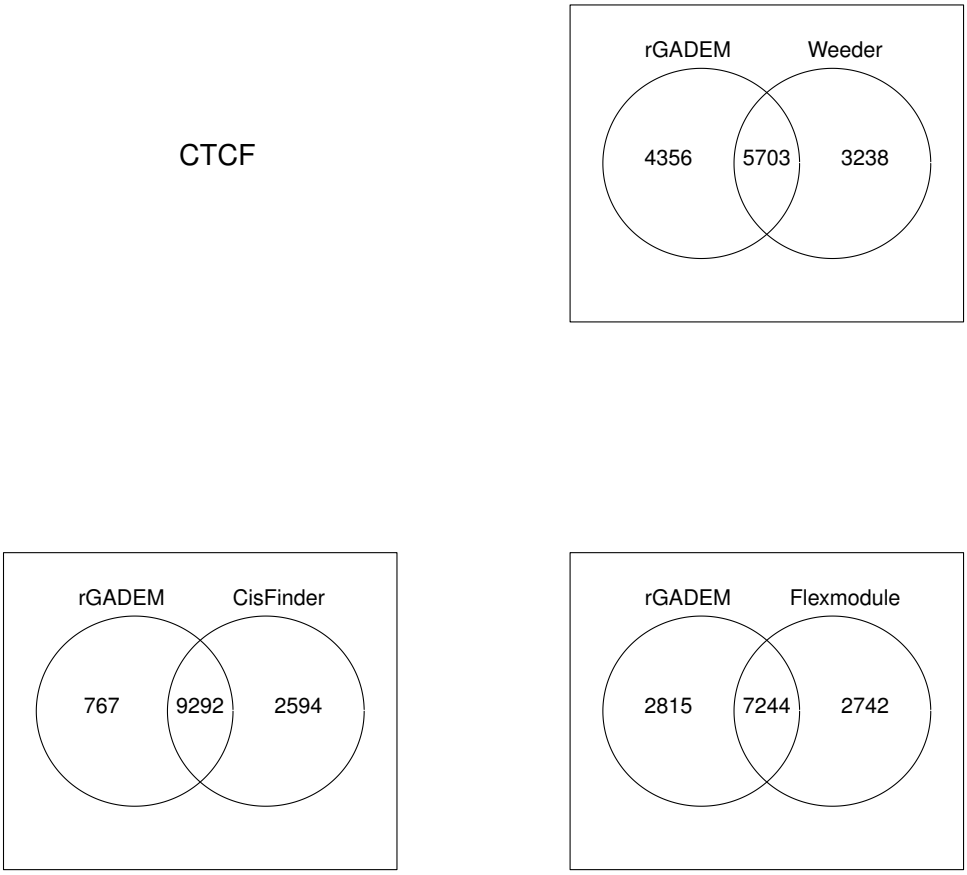


Figure 3.23 – Diagrammes de Venn montrant les motifs CTCF communément identifiés par rGADEM, CisFinder, FlexModule et Weeder.

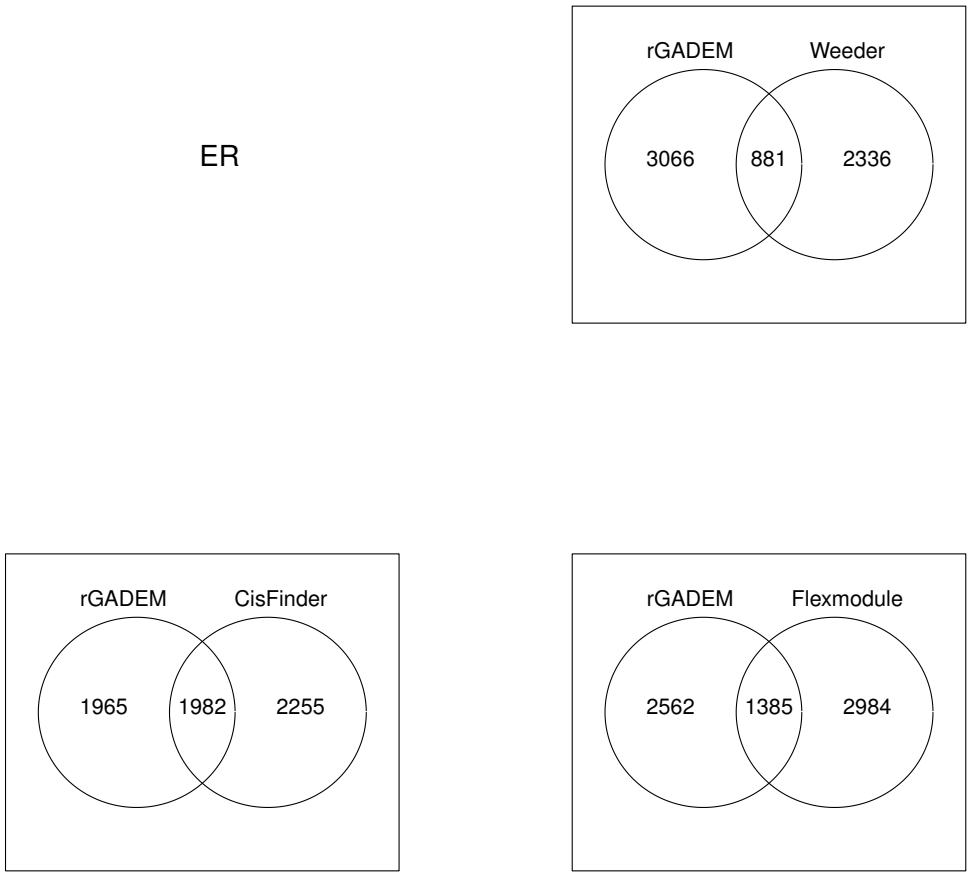


Figure 3.24 – Diagrammes de Venn montrant les motifs ER communément identifiés par rGADEM, CisFinder, FlexModule et Weeder.

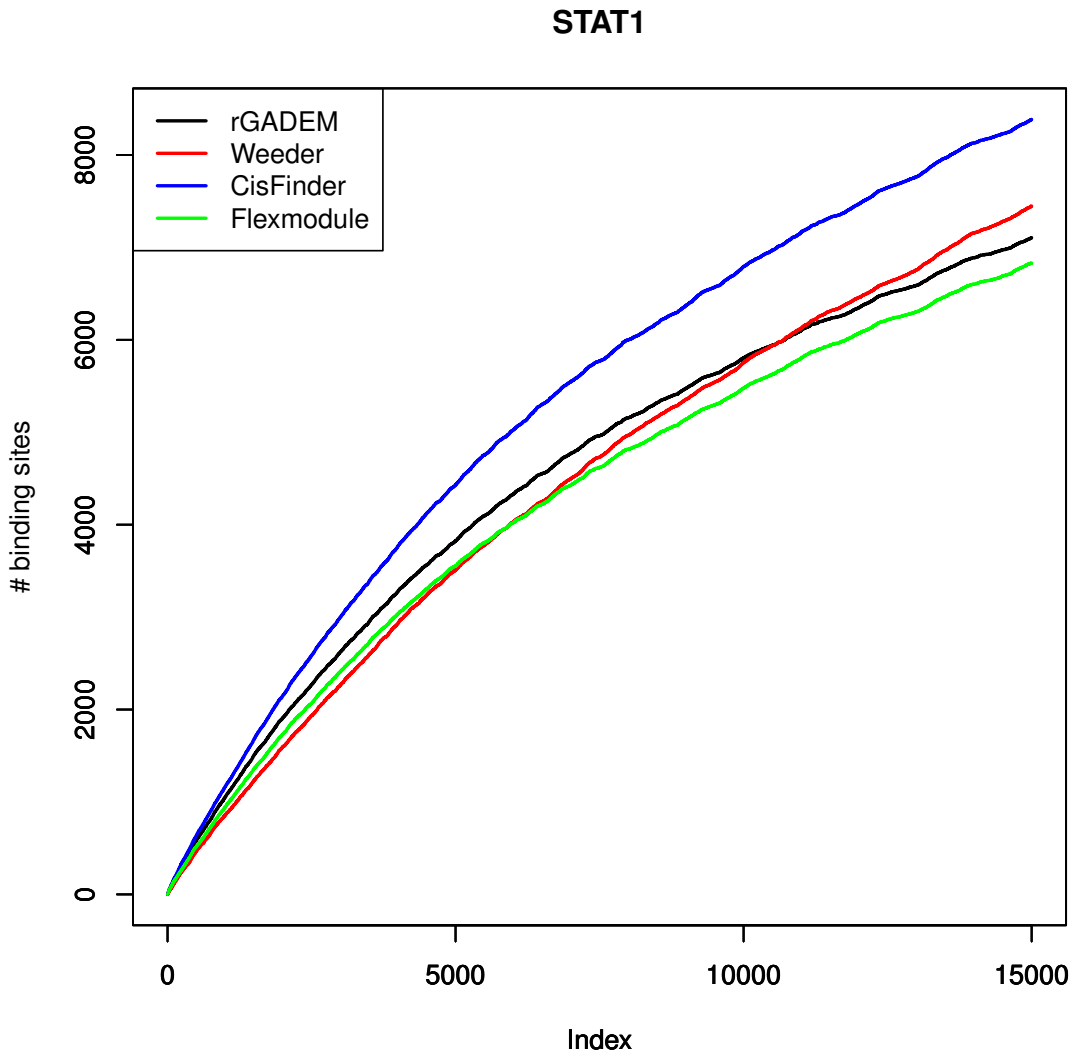


Figure 3.25 – Comparaison du nombre de sites de liaison du motif STAT1 identifiés par rGADEM, CisFinder, FlexModule et Weeder selon le rang des régions enrichies.

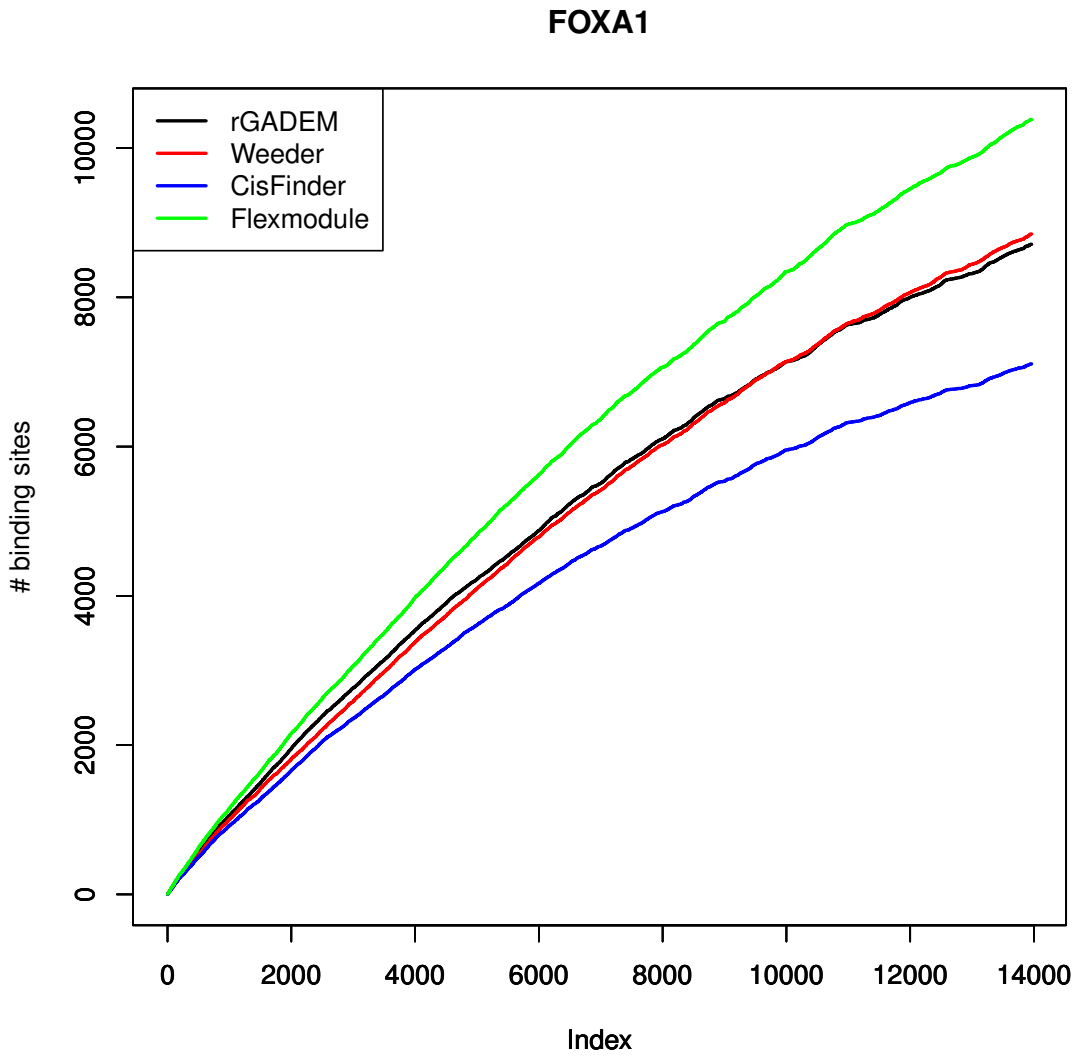


Figure 3.26 – Comparaison du nombre de sites de liaison du motif FOXA1 identifiés par rGADEM, CisFinder, FlexModule et Weeder selon le rang des régions enrichies.

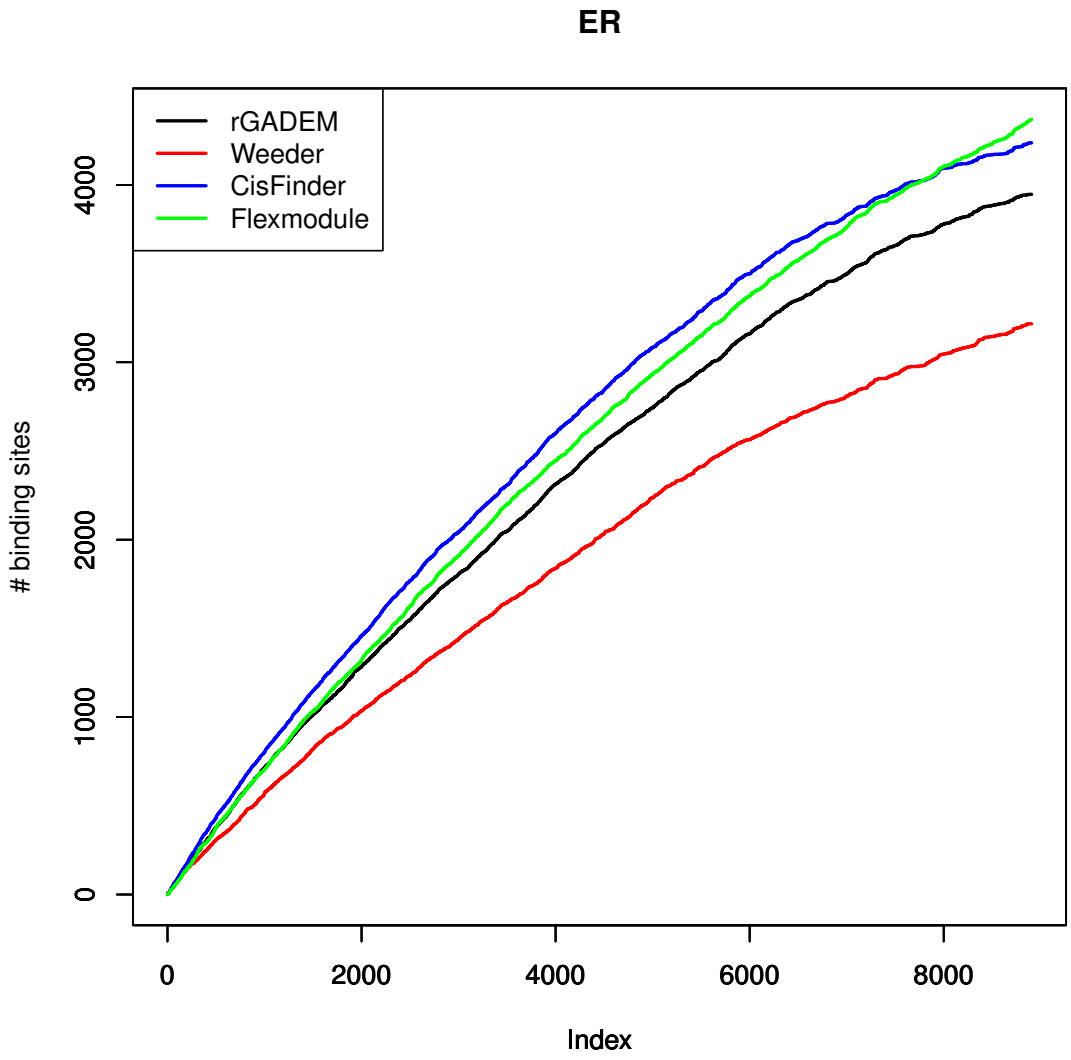


Figure 3.27 – Comparaison du nombre de sites de liaison du motif ER identifiés par rGADEM, CisFinder, FlexModule et Weeder selon le rang des régions enrichies.

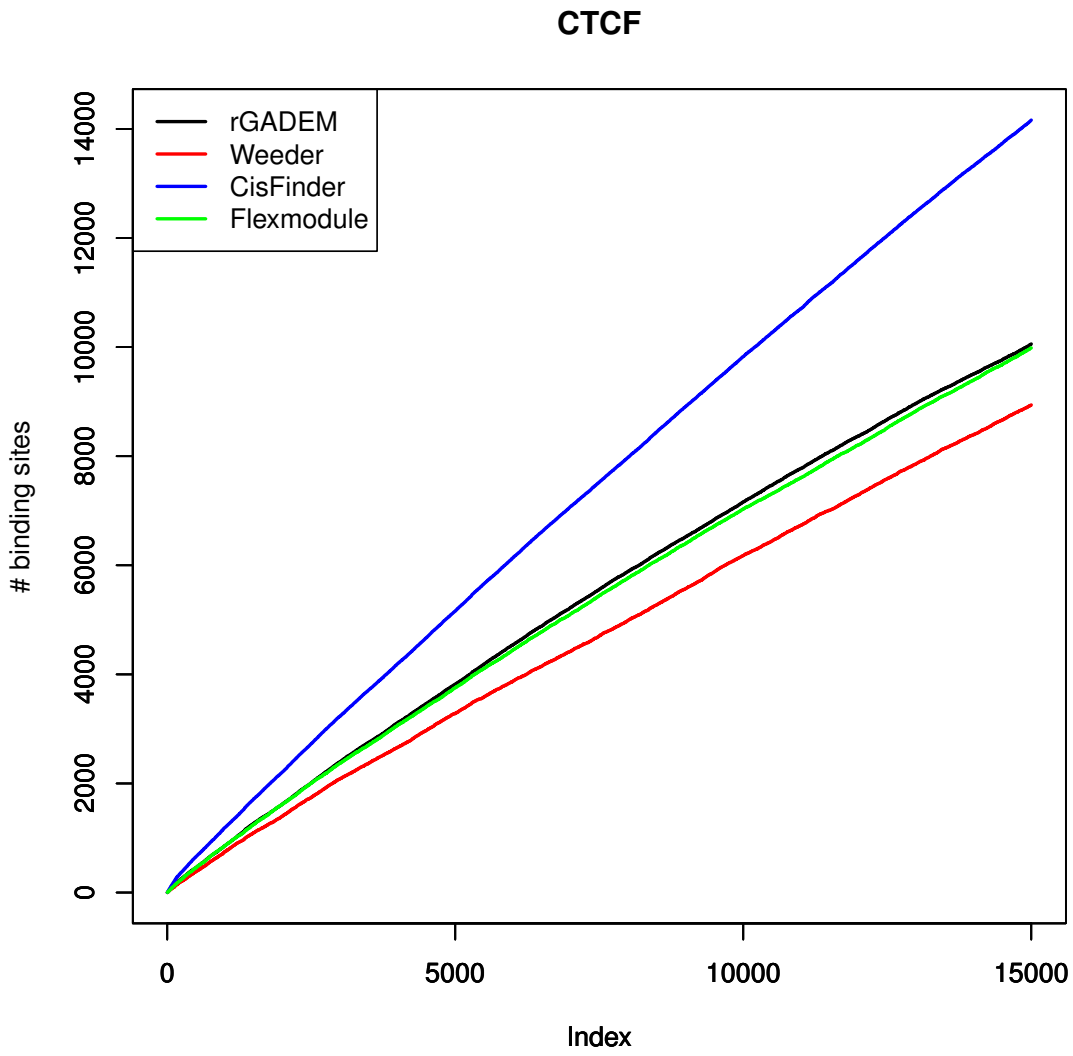


Figure 3.28 – Comparaison du nombre de sites de liaison du motif CTCF identifiés par rGADEM, CisFinder, FlexModule et Weeder selon le rang des régions enrichies.

CHAPITRE 4

DISCUSSION ET CONCLUSION

Nous avons développé un ensemble d'analyse pour l'analyse de facteurs de transcripotion par la méthode ChIP-seq et se basant sur trois modules complémentaires : PICS, rGADEM et MotIV. Nous avons utilisé quatre jeux de données -CTCF, FOXA1, STAT1 et ER- disponible librement afin de démontrer les performances de notre ensemble d'analyse par rapport aux autres logiciels d'identification des motifs et de module de régulation. Nous avons ainsi retrouvé, dans chacun des jeux de données, le site de liaison des facteurs de transcription et avons également réussi à identifier des paires de motifs en accord avec la littérature mais non détectées par les autres algorithmes.

Il existe également plusieurs autres ensembles d'analyse de données ChIP-seq tels que MICSA [40], CEAS [45] ou Sole-Search [13] mais souffrant, selon nous, de plusieurs limitations. MICSA par exemple a été développé afin d'améliorer l'analyse des données ChIP-seq en favorisant les régions enrichies pour un site de liaison d'un facteur de transcription précis. Pour ce faire, il utilise l'algorithme MEME sur quelques premiers milliers de séquences afin d'identifier un motif et réapplique l'algorithme au reste des séquences en recherchant le motif précédemment détecté. Si cela a pour effet d'accélérer la recherche de motifs, cela compromet la découverte des autres motifs que celui du site de liaison du facteur de transcription principal. CEAS et Sole-Search, bien que performant, ne permettent pas autant de fonctionnalité ni de flexibilité que notre ensemble d'analyse. Nous avons tenté de comparer notre ensemble d'analyse à CEAS et ClusterBuster mais le peu de contrôle sur les résultats ne nous a pas permis de les comparer efficacement.

L'ensemble d'analyse décrit dans ce rapport offre des fonctionnalités uniques qui ne se retrouvent pas dans les nombreux autres outils d'analyse des données CHIP-seq. Il offre ainsi de nombreux moyens de visualiser les résultats, les alignements, la distribu-

tion des sites de liaison et la distance inter-motifs. Les fonctions de filtrages des motifs sont également novatrices et permettent d'épurer les résultats de motifs peu significatifs, tout comme la fonction de regroupement des motifs similaires. Pour cette raison il est inutile de masquer les séquences répétées avant l'analyse comme il est recommandé de le faire pour CEAS et MICSA par exemple et qui est susceptible de retirer des motifs intéressants. L'approche que nous avons mise en place combine le taux d'enrichissement calculé par rGADEM et les informations sur la distribution des motifs afin de discriminer les motifs de valeurs.

Contrairement à PICS et GADEM (sur lequel se base rGADEM) qui ont été développés séparément, MotIV a été développé spécifiquement pour notre ensemble d'analyse. Bien que nous ayons travaillé avec des données ChIP-seq, les modules rGADEM et MotIV sont aussi susceptible d'être utilisé sur des données ChIP-chip. La flexibilité de notre ensemble d'analyse permet de remplacer facilement PICS par un autre algorithme permettant l'analyse des données ChIP-chip. De plus, n'importe quel utilisateur ayant des connaissances en R peut modifier nos modules à leur guise, ceux-ci étant sous licence-libre. Tout comme il est possible d'utiliser d'autres modules de Bioconductor pour approfondir ou apporter des fonctionnalités supplémentaires concourant ainsi à fournir un environnement complet pour l'analyse des données ChIP-seq.

BIBLIOGRAPHIE

[1] **A, B., C. S, C. K, R. TY, S. DE, W. Z, W. G, C. I, and Z. K**. 2008. High-resolution profiling of histone methylations in the human genome. Cell **129** :823–37.

[2] **A, S., and W. WW**. 2004. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. J Mol Biol **338** :207–15.

[3] **AP, F., R. G, B. M, V. R, B. M, and J. SJ**. 2008. Findpeaks 31 : a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics **24** :1729–30.

[4] **Bailey, T. L., N. Williams, C. Misleh, and W. W. Li**. 2006. Meme : discovering and analyzing dna and protein sequence motifs. Nucleic Acids Research .

[5] **Bonni, A., D. A. Frank, C. Schindler, and M. E. Greenberg**. 1993. Science **262** :1575.

[6] **C, L., W. UM, Y. J, B. J, S. C, Z. A, H. AG, W. AF, Y. K, T. T, and et al**. 1994. Association of transcription factor aprf and protein kinase jak1 with the interleukin-6 signal transducer gp130. Science **263** :89–92.

[7] **Carroll, JS, Meyer, CA, Song, J, Li, W, Geistlinger, TR, Eeckhoute, J, Brodsky, AS, Keeton, EK, Fertuck, KC, Hall, GF, Wang, Q, Bekiranov, S, Sementchenko, V, Fox, EA, Silver, PA, Gingeras, TR, Liu, XS, and B. M**. 2006. Genome-wide analysis of estrogen receptor binding sites. Nature Genetics **38** :1289Ű97.

[8] **CH, Y., B. H, and D. EH**. 1998. Genomic cis-regulatory logic : experimental and computational analysis of a sea urchin gene. Science **279** :1896–902.

[9] **Chinenov, Y., and T. Kerpolla**. 2001. Close encounters of many kinds fosŰjun interactions that mediate transcription regulatory specificity. Oncogene **20** :2438–52.

[10] **Consortium, T. G. O.** 2000. Gene ontology : tool for the unification of biology. Nature Genetics **25** :25–9.

[11] **DS, L.** 1997. Transcription factors : an overview. Int J Biochem Cell Biol **12** :1305–12.

[12] **et al, G. R.** 2007. Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. Nature Methods **31** :374–378.

[13] **et al, K. R. B.** 2009. Sole-search : an integrated analysis program for peak detection and functional annotation using chip-seq data. Nucleic Acids Research **38**.

[14] **G, R., H. M, B. M, and B. M. andZhao Y et al.** 2007. Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods **4** :651Ű7.

[15] **Gupta, S., J. Stamatoyannopolous, T. Bailey, and W. S. Noble**. 2007. Quantifying similarity between motifs. Genome Biology **8**.

[16] **J, E., C. JS, G. TR, T.-A. MI, and B. M**. 2006. A cell-type-specific transcriptional network required for estrogen regulation of cyclin d1 and cell cycle progression in breast cancer. Genes Dev **20** :2513–26.

[17] **JE, D., K. IM, and S. GR**. 1994. Jak-stat pathways and transcriptional activation in response to ifns and other extracellular signaling proteins. Science **264** :1415–1421.

[18] **Ji, H., H. Jiang, W. Ma, D. S. Johnson, R. M. Myers, and W. H. Wong**. 2003. An integrated software system for analyzing chip-chip and chip-seq data. Nature Biotechnology **26** :1293–1300.

[19] **JN, I.** 1995. Cytokine receptor signalling. Science **377** :591–4.

[20] **L, C., S. C, A. L, C. M, N. G, F. A, I. G, C. R, B. N, D. B. M, S. C, S. P, B. F, and W. A**. 2004. A genomic view of estrogen actions in human breast cancer cells by expression profiling of the hormone-responsive transcriptome. J Mol Endocrinol **32** :719–75.

[21] **Li, L.** 2009. Gadem : a genetic algorithm guided formation of spaced dyads coupled with an em algorithm for motif discovery. Journal of Computational Biology .

[22] **M, H., Y. J, T. J, C. A, and Q. Z**. 2010. On the detection and refinement of transcription factor binding sites using chip-seq data. Nucleic Acids **38** :3154–67.

[23] **M, L., E. J, M. CA, W. Q, Z. Y, L. W, C. JS, L. XS, and B. M**. 2008. Foxa1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell **132** :958–70.

[24] **Mahony, S., P. Auron, and P. Benos**. 2007. Dna familial binding profiles made easy : comparison of various motif alignment and clustering strategies. PLoS Computational Biology **3**.

[25] **Mahony, S., and P. Benos**. 2007. Stamp : a web tool for exploring dna-binding motif similarities. Nucleic Acids Research **35** :253–258.

[26] **Matys, V., E. Fricke, R. Geffers, E. GöSSling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D.-U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Münch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender**. 2003. Transfac : transcriptional regulation and from patterns to profiles. Nucleic Acids Research **31** :374–378.

[27] **N, S., F. TJ, B. TG, Z. Z, D. J. Jr, and Y. GD**. 1995. Choice of stats and other substrates specified by modular tyrosine-based motifs in cytokine receptors. Science **267** :1349–53.

[28] **Ning, K.** 2007. Biological Insights of Transcription Factor through Analyzing ChIP-Seq Data. Master's thesis, University of British-Columbia.

[29] **Pavesi, G., F. Zambelli, and G. Pesole**. 2007. Weederh : an algorithm for finding conserved regulatory motifs and regions in homologous sequences. BMC Bioinformatics .

[30] **Pepke, S., B. Wold, and A. Mortazavi**. 2009. Computation for chip-seq and rna-seq studies. Nature Methods **6** :S22 – S32.

[31] **Sandelin, A., W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard**. 2003. Jaspar : an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Research .

[32] **Schneider, D., and R. Stephens**. 1990. Sequence logos : A new way to display consensus sequences. Nucleic Acid Research **18** :6097Ű6100.

[33] **Schneider, D., G. D. Stormo, L. Gold, and A. Ehrenfeuch**. 1986. Dinformation content of binding sites on nucleotide sequences. Journal of Molecular Biology **3**.

[34] **Sharov, A. A., and M. S. Ko**. 2009. Exhaustive search for over-represented dna sequence motifs with cisfinder. DNA Research **16** :261–273.

[35] **Shaulian, E., and M. Karin**. 2001. Ap-1 in cell proliferation and survival. Oncogene **20** :2390–2400.

[36] **Shema, E., I. Tirosh, Y. Aylon, and et al**. 2008. The histone h2b-specific ubiquitin ligase rnf20/hbre1 acts as a putative tumor suppressor through selective regulation of gene expression. Genes Dev **22** :2664–2676.

[37] **Team, R. D. C.** 2004. Wrinting r extension. URL `cran.r-project.org/manuals.html`.

[38] **Team, R. D. C.** 2011. R : A language and environment for statistical computing. URL `http://wwwR-projectorg/`.

[39] **TS, M., K. M, J. DB, I. B, and et al.** 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature **448** :553–60.

[40] **V, B., S. D, G. N, T. F, F. AP, D. O, and B. E**. 2010. De novo motif identification improves the accuracy of predicting transcription factor binding sites in chip-seq data analysis. Nucleic Acids Res **38**.

[41] **Valouev, A., D. S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. M. Myers, and A. Sidow**. 2008. Genome-wide analysis of transcription factor binding sites based on chip-seq data. Nature Methods .

[42] **van Dam, H., and M. Castellazzi**. 2001. Distinct roles of jun : Fos and jun : Atf dimers in oncogenesis. Oncogene **20** :2453–2464.

[43] **W, T., R. EC, , and L. CE**. 2003. Gibbs recursive sampler : finding transcription factor binding sites. Nucleic Acids Research **31** :3580–3585.

[44] **W, X., C. SAA, Z. S, C. SC, M.-K. J, M. J, H. SJ, and E. SC**. 2003. Cooperative transcription activation of nitric oxide synthase 2 trough stat-1 and c-fos interaction. Am J Physiol **285** :L137–48.

[45] **X, J., L. W, S. J, W. L, and L. XS**. 2006. Ceas : cis-regulatory element annotation system. Nucleic Acids Research **34** :551–554.

[46] **Zhang, X., G. Robertson, M. Krzywinski, K. Ning, A. Droit, S. Jones, and R. Gottardo**. 2009. Pics : Probabilistic inference for chip-seq. URL `http://wwwcitebaseorg/abstract?id=oai:arXivorg:09033206`.

[47] **Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nussbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu**. 2008. Model-based analysis of chip-seq (macs). Genome Biology **9**.

[48] **Zhu, L. J., H. Pages, C. Gazin, S. L. Nathan Lawson, D. Lapointe, and M. Green**. 2010. Chippeakanno : a bioconductor package to annotate chip-seq and chip-chip data. BMC Bioinformatics **11** :237.

Chapitre 5

Annexe 1

5.1 Estimation du FDR par PICS

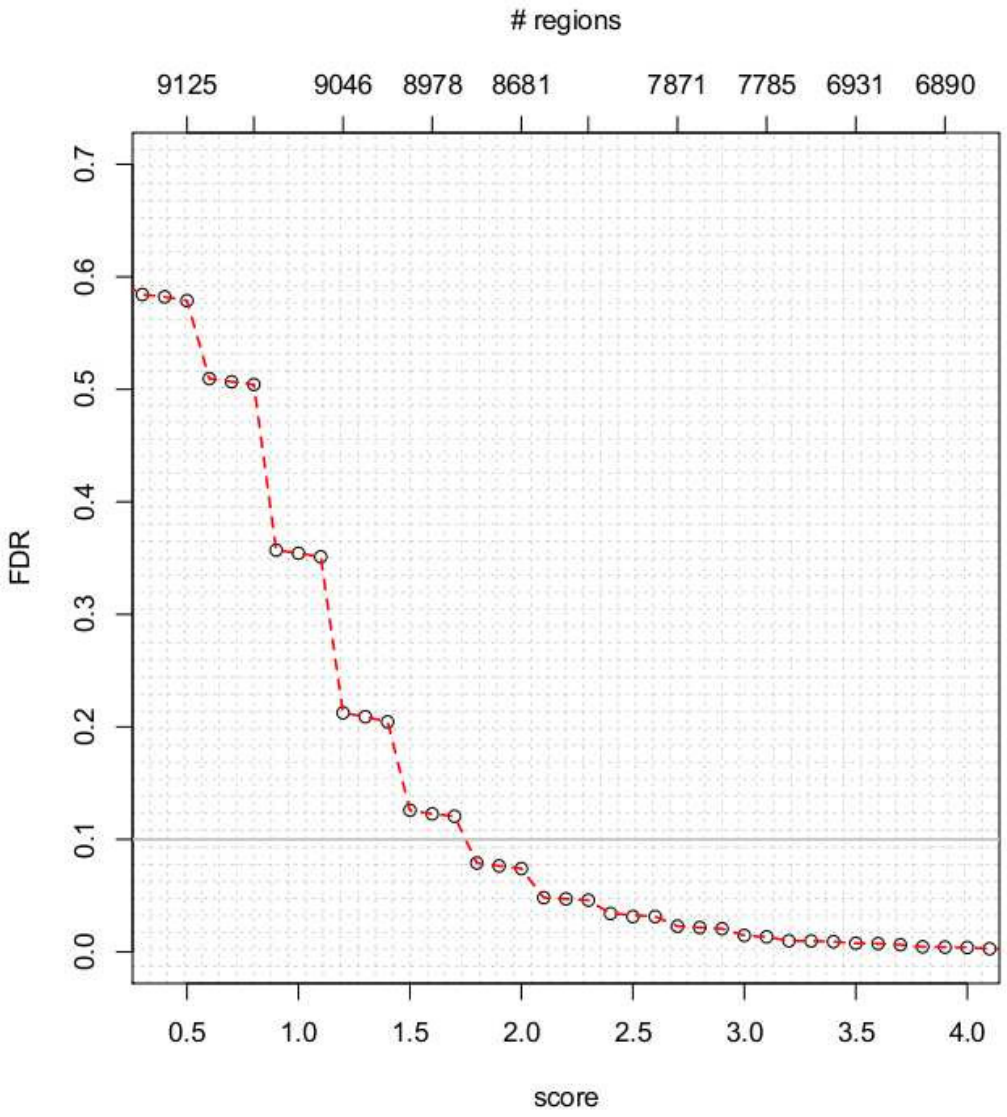


Figure 5.1 – FDR estimé en fonction du score d'enrichissement pour les données d'ER.

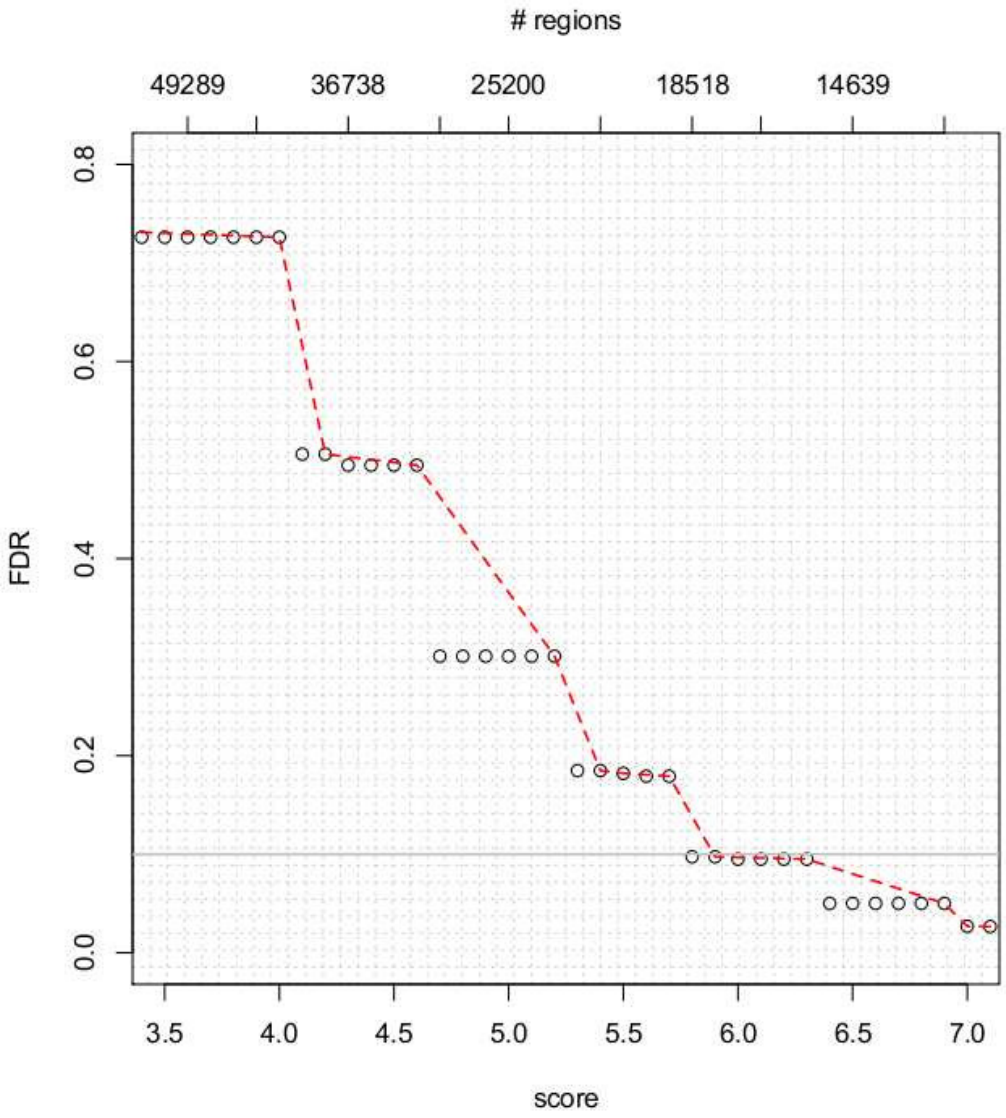


Figure 5.2 – FDR estimé en fonction du score d'enrichissement pour les données de STAT1.

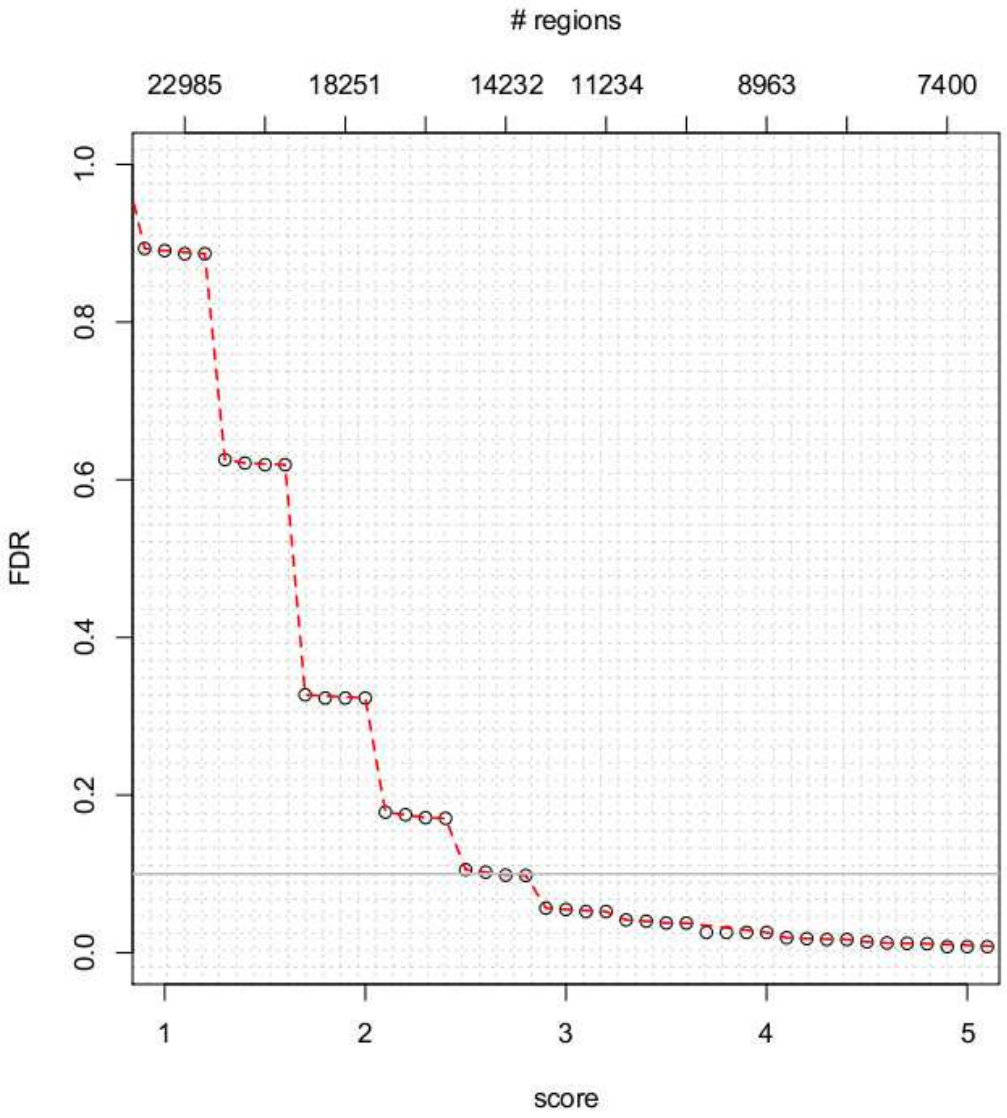


Figure 5.3 – FDR estimé en fonction du score d'enrichissement pour les données de FOXA1.

5.2 Proportion du nombre de site de liaison identifiés

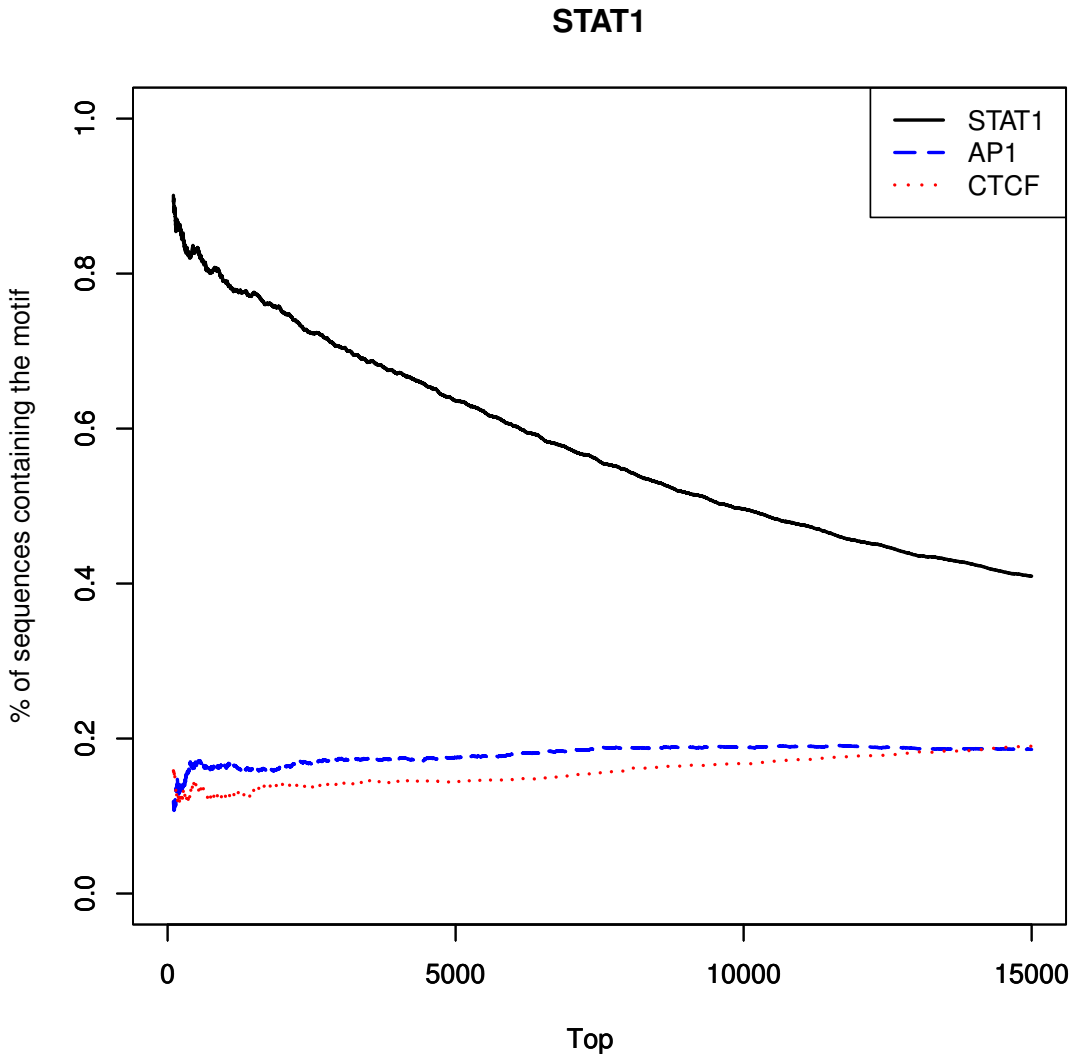


Figure 5.4 – Distribution du nombre de sites de liaison de STAT1 en fonction du rang des séquences.

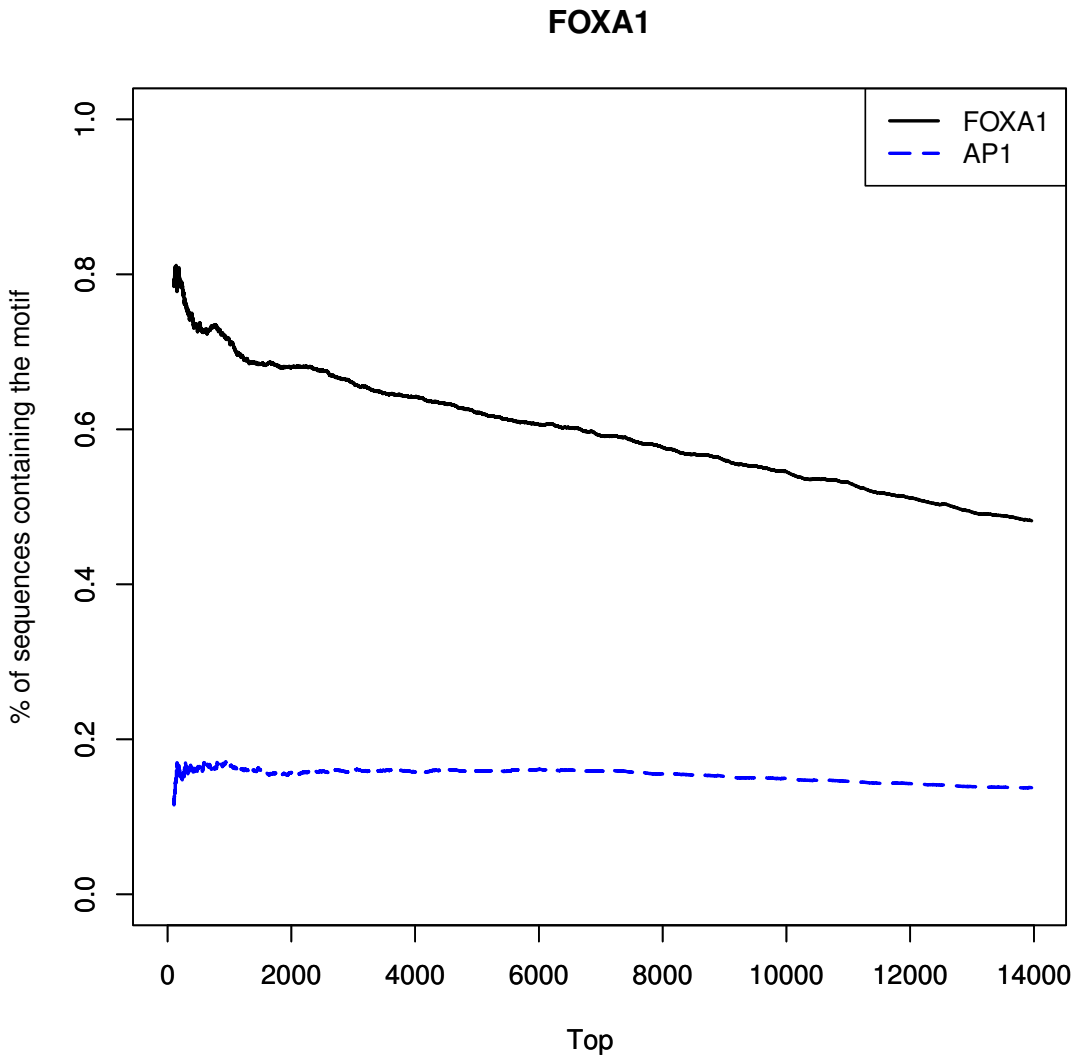


Figure 5.5 – Distribution du nombre de sites de liaison de FOXA1 en fonction du rang des séquences.

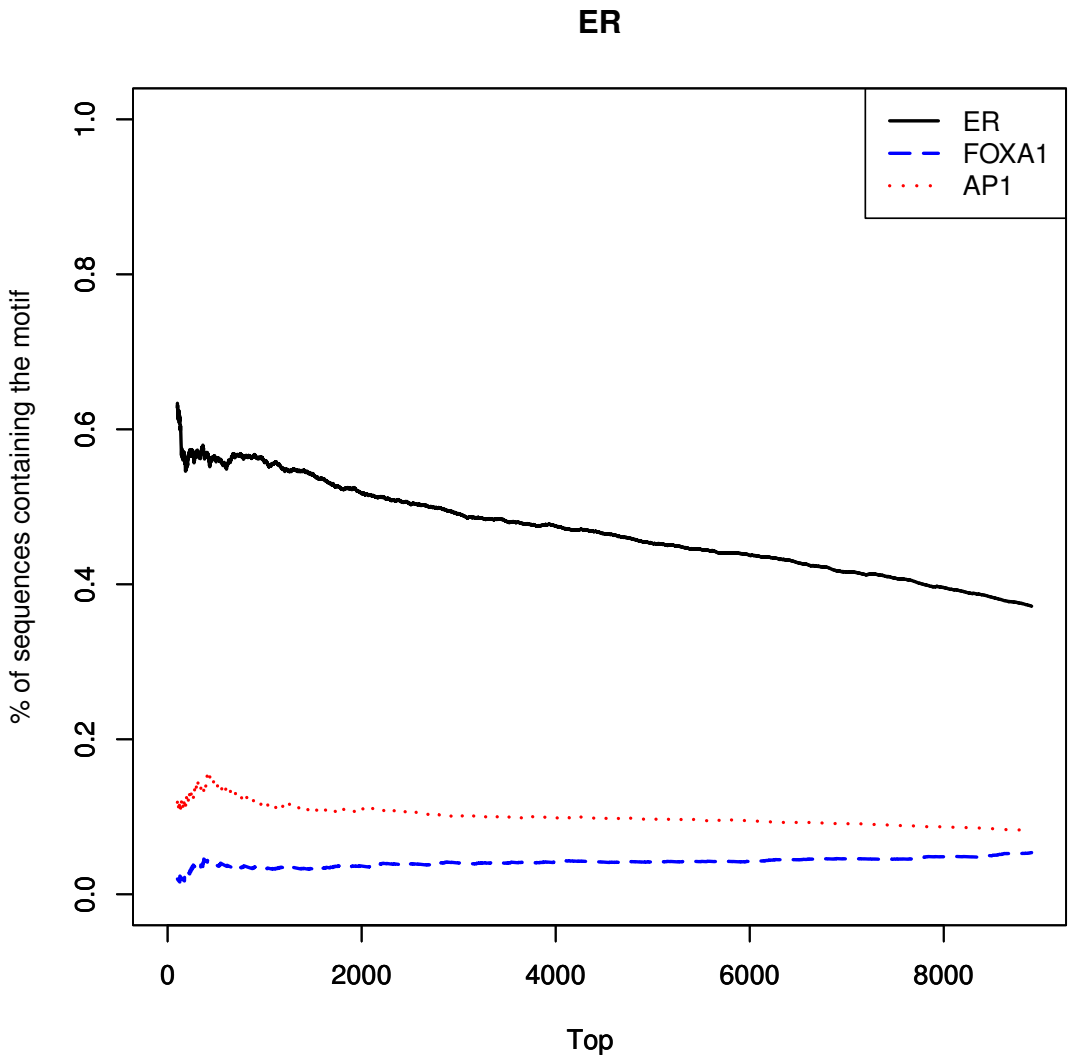


Figure 5.6 – Distribution du nombre de sites de liaison de ER en fonction du rang des séquences.

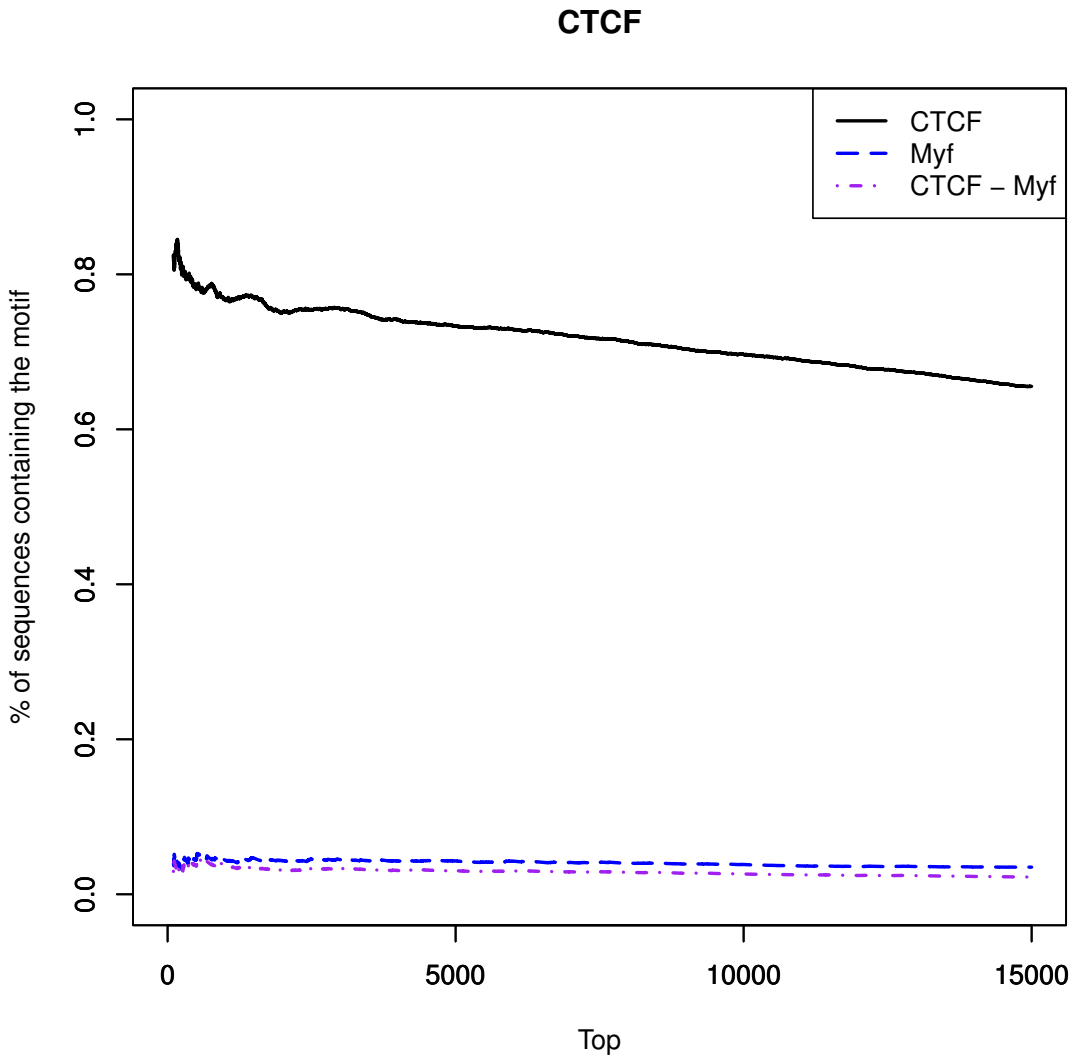


Figure 5.7 – Distribution du nombre de sites de liaison de CTCF en fonction du rang des séquences.