Université de Montréal

Une nouvelle approche computationnelle pour la découverte des sites de fixation de facteurs de transcription à l'ADN, adaptée aux données de ChIP-chip et de ChIP-séquençage

Par
Malika Aid

Département de Biochimie, Institut de Recherche en Immunologie et Cancérologie

Faculté de Médecine

Thèse présentée à la Faculté des études supérieures
en doctorat en bio-informatique

Septembre, 2012

Université de Montréal
Faculté des études supérieures et postdoctorales

Cette thèse est intitulée :

Une nouvelle approche computationnelle pour la découverte des sites de fixation de facteurs de transcription à l'ADN, adaptée aux données de ChIP-chip et de ChIP-séquençage

Présentée par :
Malika Aid

A été évaluée par un jury composé des personnes suivantes :

François Robert Ph.D., président-rapporteur
Sylvie Mader Ph.D., directeur de recherche
François Major Ph.D., co-directeur
Guillaume Bourque Ph.D., membre de jury
Mathieu Lupien Ph.D., examinateur externe
Franz Lang Ph.D., représentant du doyen de la FES

Résumé

Les facteurs de transcription sont des protéines spécialisées qui jouent un rôle important dans différents processus biologiques tel que la différenciation, le cycle cellulaire et la tumorigenèse. Ils régulent la transcription des gènes en se fixant sur des séquences d'ADN spécifiques (éléments cis-régulateurs). L'identification de ces éléments est une étape cruciale dans la compréhension des réseaux de régulation des gènes. Avec l'avènement des technologies de séquençage à haut débit, l'identification de tout les éléments fonctionnels dans les génomes, incluant gènes et éléments cis-régulateurs a connu une avancée considérable. Alors qu'on est arrivé à estimer le nombre de gènes chez différentes espèces, l'information sur les éléments qui contrôlent et orchestrent la régulation de ces gènes est encore mal définie. Grace aux techniques de ChIP-chip et de ChIP-séquençage il est possible d'identifier toutes les régions du génome qui sont liées par un facteur de transcription d'intérêt. Plusieurs approches computationnelles ont été développées pour prédire les sites fixés par les facteurs de transcription. Ces approches sont classées en deux catégories principales: les algorithmes énumératifs et probabilistes. Toutefois, plusieurs études ont montré que ces approches génèrent des taux élevés de faux négatifs et de faux positifs ce qui rend difficile l'interprétation des résultats et par conséquent leur validation expérimentale.

Dans cette thèse, nous avons ciblé deux objectifs. Le premier objectif a été de développer une nouvelle approche pour la découverte des sites de fixation des facteurs de transcription à l'ADN (SAMD-ChIP) adaptée aux données de ChIP-chip et de ChIP-séquençage. Notre approche implémente un algorithme hybride qui combine les deux stratégies énumérative et probabiliste, afin d'exploiter les performances de chacune d'entre elles. Notre approche a montré ses performances, comparée aux outils de découvertes de motifs existants sur des jeux de données simulées et des jeux de données de ChIP-chip et de ChIP-séquençage. SAMD-ChIP présente aussi l'avantage d'exploiter les propriétés de

distributions des sites liés par les facteurs de transcription autour du centre des régions liées afin de limiter la prédiction aux motifs qui sont enrichis dans une fenêtre de longueur fixe autour du centre de ces régions.

Les facteurs de transcription agissent rarement seuls. Ils forment souvent des complexes pour interagir avec l'ADN pour réguler leurs gènes cibles. Ces interactions impliquent des facteurs de transcription dont les sites de fixation à l'ADN sont localisés proches les uns des autres ou bien médier par des boucles de chromatine. Notre deuxième objectif a été d'exploiter la proximité spatiale des sites liés par les facteurs de transcription dans les régions de ChIP-chip et de ChIP-séquençage pour développer une approche pour la prédiction des motifs composites (motifs composés par deux sites et séparés par un espacement de taille fixe). Nous avons testé ce module pour prédire la co-localisation entre les deux demi-sites ERE qui forment le site ERE, lié par le récepteur des œstrogènes ERα. Ce module a été incorporé à notre outil de découverte de motifs SAMD-ChIP.

Mots-clés : *ChIP-chip, ChIP-séquençage, réseau de régulation des gènes, facteurs de transcription, découverte de motifs d'ADN, fonctions de score, éléments cis-régulateurs, modules cis-régulateur, cancer du sein, récepteur des œstrogènes α, ERE, TF, TFBS.*

Abstract

Transcription factors (TF) play important roles in various biological processes such as differentiation, cell cycle progression and tumorigenesis. They regulate gene expression by binding to specific DNA sequences (TFBS). Identifying these cis-regulatory elements is a crucial step to understand gene regulatory networks. Technological developments have enhanced DNA sequencing at genomic scale. On the basis of the resulting sequences, computational biologists now attempt to localize the most important functional regions, starting with genes, but also importantly the whole genome characterization of transcription factor binding sites and allow the development of several computational DNA motif discovery tools.

Although these various tools are widely used and have been successful at discovering novel motifs, they are not adapted to ChIP-chip and ChIP-sequencing data. The main drawback of these approaches is that most of the predicted motifs represent artifacts due to an inefficient assessment of their enrichment.

This thesis is about transcription factor proteins and statistical analysis of their binding sites in ChIP-chip and ChIP-sequencing data. The first objective was to develop a new *do novo* DNA motif discovery tool adapted to ChIP-chip and ChIP-sequencing data. SAMD-ChIP combines enumerative and stochastic strategies to predict enriched motifs in the vicinity of the ChIP peak summits. Our approach is an automated pipeline that includes motif discovery, motif clustering, motif optimization and finally motif identification using transcription factor (TF) databases. SAMD-ChIP outperforms state-of-the-art motif discovery tools in term of the number of predicted motifs and the prediction of rare and degenerate motifs. In particular, SAMD-ChIP efficiently identifies gapped motifs such as

inverted or direct repeats bound by nuclear receptors and composite motifs resulting from the association of different single TF binding sites.

The underlying assumption of the second objective is that in regulatory regions, binding sites of interacting transcription factors co-occur more often than expected by chance in the vicinity of the ChIP-peak summits. We proposed an approach to predict transcription factor binding sites co-localization based on the prediction of single motifs by *do novo* motif discovery tools or by using TFBS models from TF data bases.

Keywords : *ChIP-chip, ChIP-sequencing, gene regulatory network, transcription factors, DNA motifs discovery, scoring functions, cis-regulatory module, cis-regulatory elements, estrogens receptor ERα, breast cancer, TF,TFBS.*

Table des matières

Liste des sigles et des abréviations

FT	Facteur de transcription
TFBS	Transcription factor binding site
BS	Binding site
EM	Expectation and maximization
RAR	Retenoic acid receptor
ER	Estrogen receptor
AP-1	Activator protein 1
ARNm	Acide ribonucléique messager
ADN	Acide désoxyribonucléique
DBD	Domaine de liaison à l'ADN
E2	17β-estradiol
ERE	Estrogen response element (Elément de réponse aux estrogènes)
LBD	Domaine de liaison du ligand
MCF7	Lignée cellulaire de carcinome mammaire exprimant ER
pb	Paire de base (nucléotide)
CRM	Cis-regulatory module
PWM	Position weight matrix (matrice de poids de positions)
HMM	Hidden Markov model (chaines de Markov cachées)
IC	Information content (contenu d'information)
ChIP	Chromatin immuno-precipitation
IUPAC	International Union of Pure and Applied Chemesitry
NR	Nuclear receptor (récepteur nucléaire)
PCR	Polymerase Chain Reaction
TSS	Transcription start site (site d'initiation de la transcription)
PFM	Position frequency matrix (matrice de fréquence)

Liste des tableaux

Introduction

Article 1

Article 2

Liste des figures

Introduction

Article 1

Article 2

À la mémoire de mon oncle et de ma grand-mère.

Remerciements

Le long de toutes ces années, j'ai eu le privilège de rencontrer beaucoup de personnes, certaines sont devenues mes amis et m'ont aidé à aller de l'avant dans tous les domaines de ma vie. Ces personnes m'ont appris que la vie est une succession de réussites et d'échecs et ce sont toutes ces choses qui ont fait de moi ce que je suis aujourd'hui. J'ai toujours senti leur amour et j'ai toujours senti la présence du bon dieu à mes cotés pendant mes moments de joies et mes moments de peines.

Sylvie, je m'adresse à la femme exceptionnelle que vous êtes et la directrice de recherche qui m'a guidé pendant ma maitrise et mon doctorat. J'ai appris de mes années de travail avec vous le sens de la rigueur et de l'assiduité. Vous m'avez soutenu dans des moments difficiles et vous m'avez écouté et compris. Aujourd'hui je vous dis merci pour tout.

Dr Major et Dr Lemieux un grand merci pour tous vos conseils et suggestions.

Un grand merci aux membres du laboratoire Mader pour leur amitié pendant toutes ces années, spécialement Martine, Khalid, Slim et David L.

Elaine, merci pour ton amitié et ta patience.

À mes amis, la voix de ma conscience et de ma raison qui ont été toujours là pour moi pour écouter mes délires, Diala, Rahima, Nora, Nour, Fady, Sawsan, Chahra,Cloé, Kamel et Charif

À toute ma famille, mes parents en particulier, pour leur soutien inconditionnel et leur amour.

Introduction

Cette thèse est divisée en quatre parties principales.

La première partie regroupe les chapitres 1 et 2 et introduit le contexte biologique de la problématique traitée dans le cadre de cette thèse : la régulation de l'expression des gènes. En particulier, nous allons nous attarder à l'étude et à l'influence des facteurs de transcription et de leurs sites de fixation à l'ADN. Ensuite, nous présenterons les différentes techniques expérimentales utilisées pour l'identification des interactions protéines-ADN.

Dans la partie 2 et qui regroupe les chapitres 3, 4 et 5, nous présenterons les modèles et les approches algorithmiques pour la représentation et la recherche des éléments cis-régulateurs. En particulier, nous allons détailler les algorithmes utilisés pour la prédiction des sites de fixation des facteurs de transcription à l'ADN. Nous présenterons aussi quelques approches qui sont spécifiques à l'analyse des données de ChIP-chip et de ChIP-séquençage. Nous discuterons des limites et avantages de ces approches et nous présenterons quelques pistes pour améliorer leur spécificité et leur sensibilité.

La troisième partie regroupe les chapitres 6 et 7. Dans le chapitre 6 (premier manuscrit), nous présenterons une nouvelle approche, SAMD-ChIP, pour la découverte des sites de fixation de facteurs de transcription à l'ADN. Cette approche est spécifique à l'analyse des données de ChIP-chip et de ChIP-séquençage et exploite les propriétés de distribution des sites liées dans ces régions. Dans le chapitre 7, nous présenterons une approche pour la prédiction des éléments composites. Cette approche utilise les motifs prédits par les approches de découverte de motifs ou bien les modèles répertoriés dans les bases de données de facteurs de transcription.

Enfin, dans la dernière partie (chapitre 8) nous présenterons une discussion des résultats obtenus par l'application de nos deux approches sur des données de ChIP-chip et de ChIP-séquençage réalisés contre 7 facteurs de transcription dans les cellules MCF7 du cancer du sein. Nous discuterons aussi des améliorations à apporter dans le futur aux approches proposées dans le cadre de cette thèse.

Chapitre 1

1. Régulation de la transcription

1.1 ADN, gène et génome

Le génome humain est composé d'environs 3 milliards de paires de bases (pb), dont 1.5% qui codent pour des gènes [1]. La taille de ces derniers est variable, allant de l'ordre de 1000 bp à un million de pb pour les gènes les plus longs. Les gènes sont composés de segments d'ADN appelés exons, séparés par des régions appelées introns.

L'ADN (acide désoxyribonucléique) est composé d'une chaine longue d'unités appelées nucléotides. Un nucléotide est formé par l'une des quatre bases suivantes : Adénine (A) et Guanine (G) qui forment les purines, la Cytosine (C) et la Thymine (T) qui forment les pyrimidines, à laquelle sont associés un sucre appelé désoxyribose et un groupe phosphate (P) (Figure 1).

Par convention, l'extrémité 5' de l'ADN est du coté du groupe phosphate (P) et l'extrémité 3' est du coté du groupe hydroxyle OH (Figure 1).

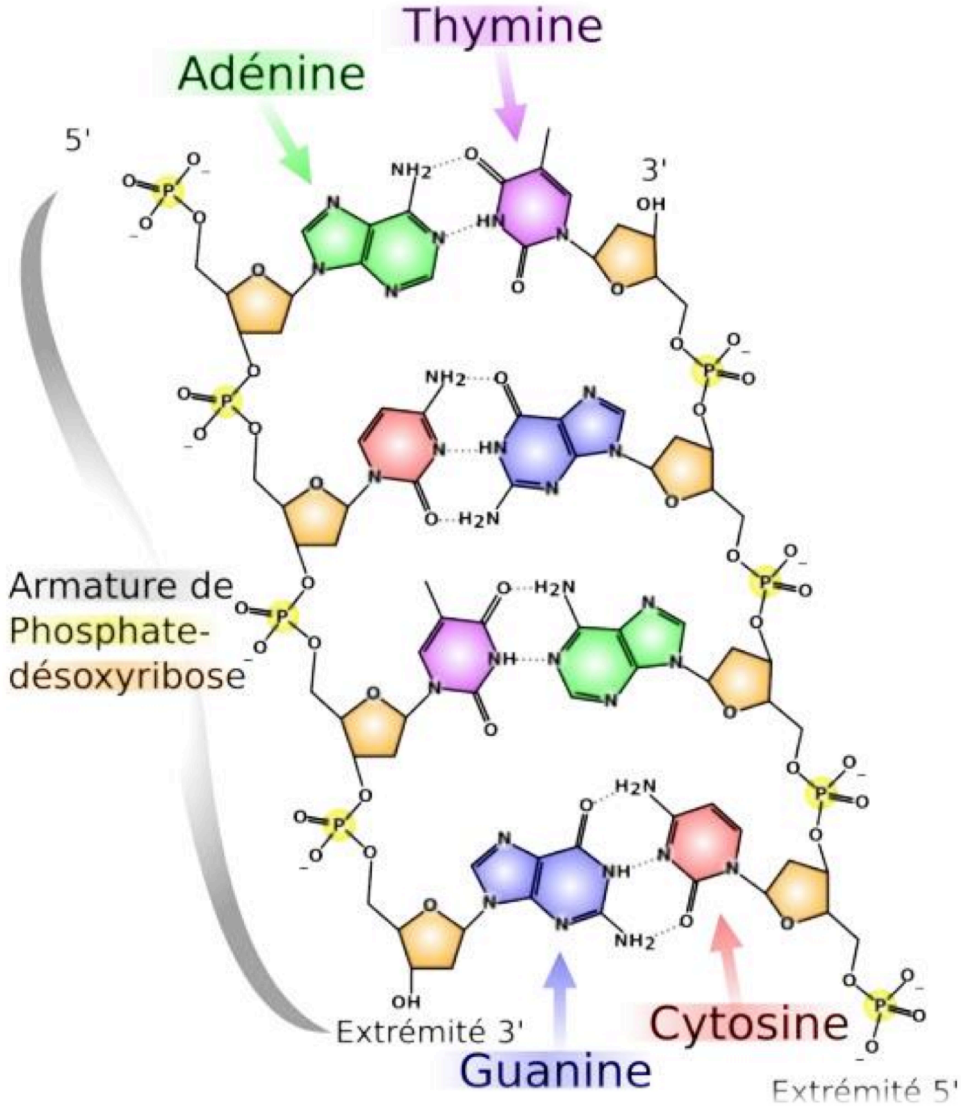


Figure 1. Structure chimique de l'ADN.

Le squelette est formé d'une succession de groupements phosphates et de sucres sur lesquels sont liées les bases azotées. Les tirets représentent les liaisons hydrogènes impliquées dans l'appariement des bases (image Wikimedia Commons).

La molécule d'ADN est associée à des protéines histones pour former une structure plus compacte, appelée le nucléosome (Figure 2).

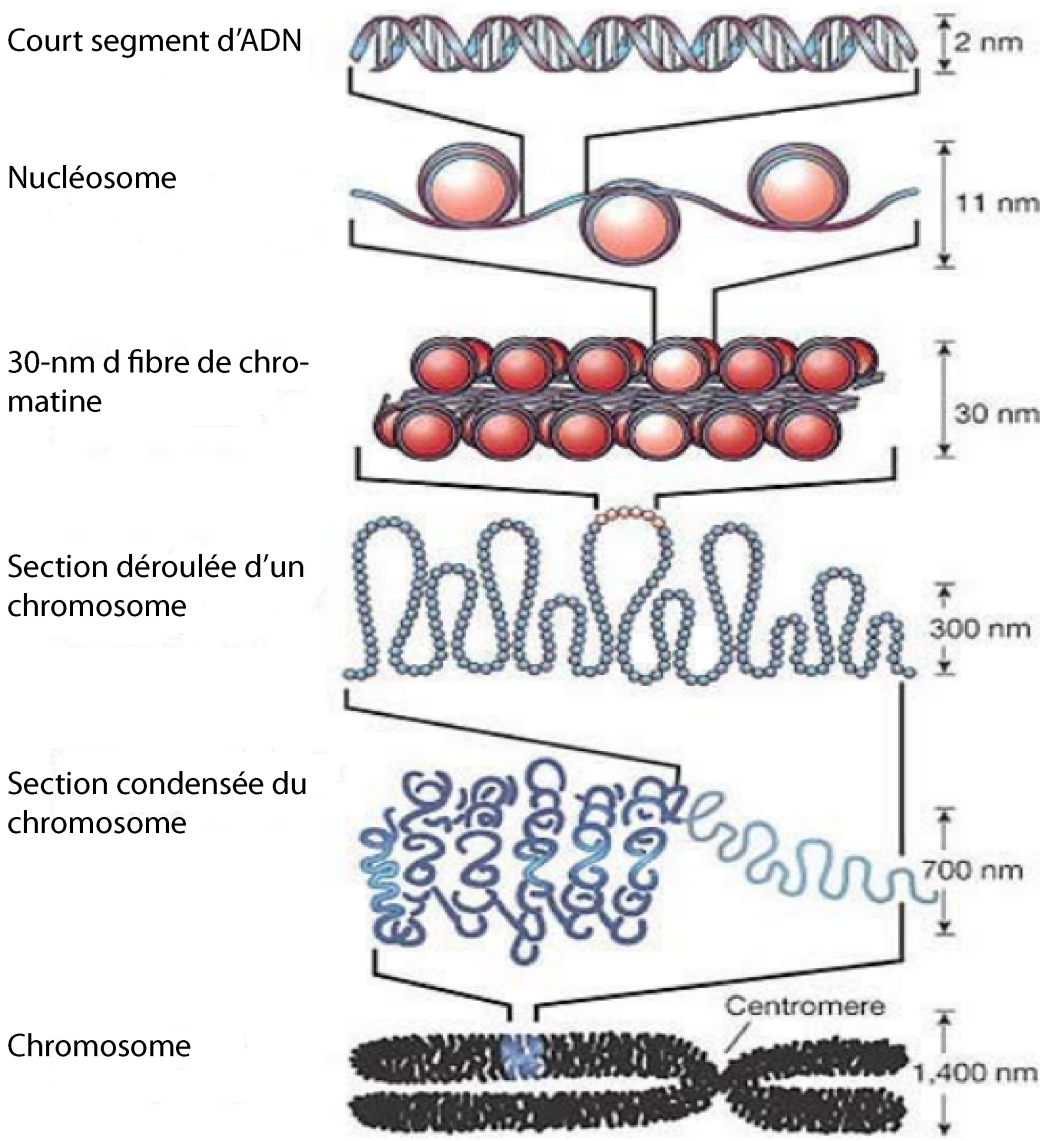


Figure 2. Les différents niveaux de compaction de la chromatine.

La figure est adaptée de http://atlasgeneticsoncology.org/Educ/ChromatinEducFr.html.

1.2 Le nucléosome

La molécule d'ADN est compactée en sous unités, appelées nucléosomes. Chaque nucléosome est composé de 147 bp enroulées autour d'un octamère des protéines histones H2A, H2B, H3 et H4 [2]. Les nucléosomes sont séparés par des régions d'ADN libre (linker en anglais) de 20-90 bp qui sont associées aux histones H1 [3]. Les histones sont de petites protéines extrêmement conservées au cours de l'évolution. Chaque histone comporte un domaine central globulaire et des extrémités moins structurées qui sortent du nucléosome et sont le siège de diverses modifications post-traductionnelles.

1.3 Domaines chromatiniens

1.3.1 L'euchromatine

L'euchromatine représente les domaines de chromatine active. Elle est riche en gènes actifs. Son organisation suggère une structure plus ouverte et un arrangement peu régulier des nucléosomes [4]. Dans ces régions les histones sont acétylées et riches en méthylation sur la lysine 4 de l'histone H3. Cette marque de chromatine caractérise des domaines de chromatine active [5].

1.3.2 L'hétérochromatine

L'hétérochromatine est la chromatine mise en silence par répression épigénétique [4]. L'hétérochromatine est riche en séquences répétées et relativement pauvre en gènes. Les histones de l'hétérochromatine sont hypoacétylées, pauvres en méthylation de la lysine 4 de l'histone H3 et enrichies en méthylation de la lysine 9 de l'histone H3[5]. Des facteurs spécifiques lient la chromatine et entraînent sa condensation et l'inhibition de la transcription.

L'ADN de l'hétérochromatine est organisé en un enchaînement régulier de nucléosomes et est relativement résistant à la digestion par les nucléases [6]. L'hétérochromatine constitutive est définie comme la chromatine qui reste condensée dans tous les types cellulaires. Elle est composée de séquences répétées et est située au niveau des centromères

et des télomères des chromosomes. Il existe également quelques blocs isolés d'hétérochromatine constitutive dispersés dans le génome. Elle est relativement pauvre en gènes, bien qu'un certain nombre de gènes essentiels y soient localisés. L'hétérochromatine facultative est définie comme de l'euchromatine qui est mise en silence à certains stades du développement ou dans certains types cellulaires [7].

Ci-dessous quelques aspects quantitatifs de l'organisation du génome humain [http://www.humans.be/pages/biomolorganisation.htm :2010] :

Nombre de gènes : 30 000 - 50 000
Densité : 1 gène toutes les 40 000 pb, soit en moyenne 130 gènes par bande chromosomique (~3000 par chromosome)
Taille : en moyenne 10 - 15 Kb, avec d'énormes variations (1,5 Kb pour la globine, 2500 Kb pour la dystrophine)
Exons : nombre très variable de 0 (histones) à 79 (dystrophine)
taille moyenne : 200 pb (faibles variations)
Introns : énormes variations : 0,5 à 30 Kb
Distance intergénique : 20 -30 Kb
Acide ribonucléique messager (ARNm) : taille moyenne : 2,5 Kb (grandes variations)

1. 4 Régulation de l'expression des gènes

L'expression des gènes codant pour des protéines consiste en une succession de deux étapes principales qui vont permettre de produire, à partir de la matrice d'ADN, des protéines. Ces deux étapes sont : la transcription et la traduction.

Lors de l'étape de la transcription, une molécule intermédiaire, l'ARNm est synthétisée dans le noyau en utilisant la séquence d'ADN d'un gène comme modèle. Puis l'ARNm subit une phase de maturation et d'épissage afin de produire un ARN mature qui pourrait être traduit en protéine. Lors de cette phase, les régions non codantes de l'ARNm, nommées introns, sont excisées pour ne conserver que les portions codantes, appelées

exons. L'ARNm ainsi obtenu est ensuite transporté à l'extérieur du noyau pour être traduit en protéine. Lors de cette deuxième étape de traduction, les triplets de nucléotides de l'ARNm sont traduits en acides aminés et assemblés pour former une protéine. Ces différentes étapes sont présentées dans la Figure 3.

La transcription est régulée par deux catégories principales de protéines [8, 9]. La première catégorie, appelée les facteurs de transcription de base est formée de l'ARN polymérase II et des facteurs de transcription généraux. La seconde catégorie est composée des facteurs de transcription et les cofacteurs.

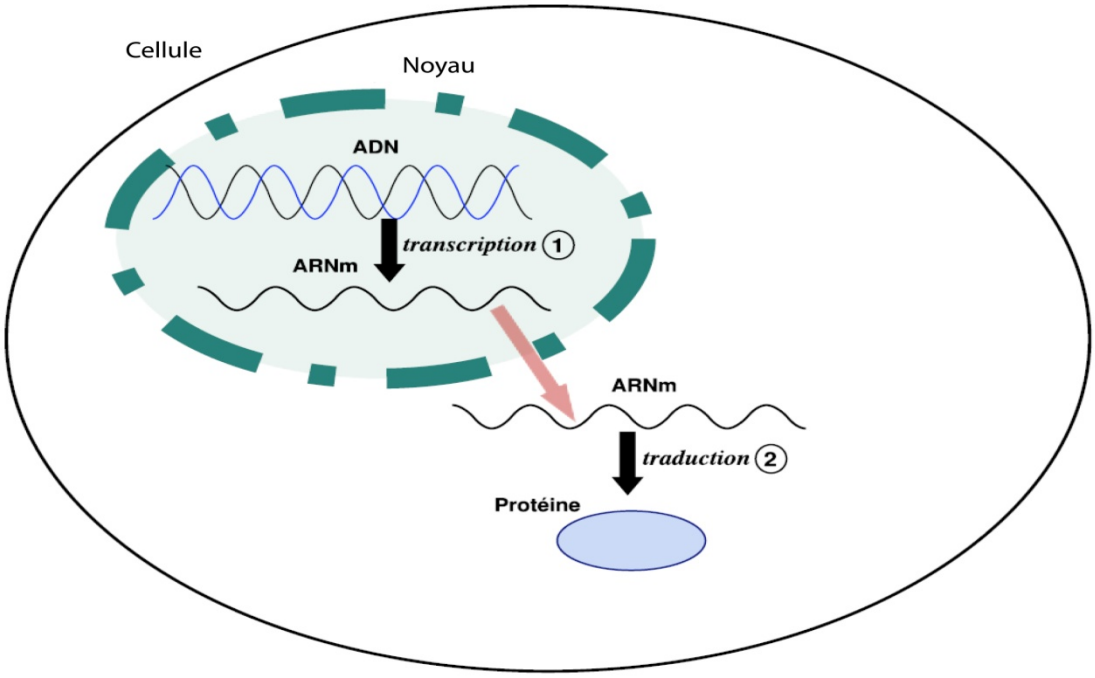


Figure 3. De l'ADN à la protéine.

Dans le cadre de cette thèse, nous allons nous concentré sur l'étape de la transcription. Nous allons nous intéresser en particulier à l'étude de la régulation de la transcription par les facteurs de transcription et leurs mécanismes de liaison à l'ADN. Cependant, il existe

d'autres classes de régulateurs de l'expression des gènes dont les microARNs. Ces derniers des petits ARN non codant de ~ 22 nucléotides de long. A l'instar des gènes codant pour les protéines, la majorité des gènes codant pour les microARNs sont transcrits par la Polymérase II. Les microARNs contrôlent l'expression des gènes au niveau post-transcriptionnel. En effet, Ils ciblent l'ARNm des gènes pour le dégrader ou bloquer sa traduction [10].

1. 5 Régulation de la transcription

Le développement récent des techniques génomiques à large échelle, comme les micro-puces d'ADN et les techniques de séquençage haut débit, a grandement contribué à l'identification des profiles d'expression des gènes dans différents tissus et sous différentes conditions physiologiques [11, 12]. Toutefois, ces techniques ne donnent aucune information sur les mécanismes qui orchestrent et contrôlent cette régulation. Tout dérèglement dans l'un de ces mécanismes peut mener à un fonctionnement anormal de la cellule. Il est donc indispensable de comprendre et de caractériser les éléments impliqués dans ces mécanismes, en particulier l'action de protéines régulatrices – les facteurs de transcription – qui jouent un rôle important dans la régulation de la transcription. Cependant, les mécanismes de régulation des gènes par cette famille de protéine ne sont pas totalement élucidés.

1. 5.1 La synthèse des protéines

Chez les eucaryotes, la transcription nucléaire est assurée par les ARN polymérases I, II et III (principalement) [13, 14]. L'ARN polymérase I est responsable de la transcription du précurseur des ARN ribosomiques (ARNr). L'ARN polymérase II transcrit majoritairement les ARN messagers (ARNm), ainsi que certains ARN non-codants. Enfin, l'ARN polymérase III synthétise les ARN de transfert (ARNt), l'ARN ribosomique 5S ainsi que certains petits ARN non-codants.

Dans un schéma simplifié, L'ARN polymérase II (Pol II) se lie au site d'initiation de la transcription d'un gène (TSS), aidé par d'autres protéines comme la protéine TBP qui lie

l’ADN à son tour pour initier la transcription. La Pol II n'est pas capable de démarrer seule la synthèse d'ARNm au niveau d'un promoteur. L'initiation de la transcription nécessite la présence de facteurs auxiliaires, appelés facteurs généraux de transcription (ou facteur de transcription de base), qui sont au nombre de 6 : TFIIA, TFIIB, TFIID, TFIIE, TFIIF et TFIIH [15]. Les facteurs généraux de transcription interviennent dans la reconnaissance du promoteur, le recrutement de l’ARN polymérase II et l’ouverture de la bulle de transcription (Figure 4). La Pol II et les facteurs généraux de transcription constituent la machinerie transcriptionnelle de base, qui est la cible d'activateurs ou de répresseurs qui modulent le taux d'expression de chaque gène en réponse à divers signaux. La machinerie transcriptionnelle est stabilisée sur l’ADN par d’autres cofacteurs [8, 9]. D’autres protéines de la famille des facteurs de transcription (FTs), exprimées de manière tissus spécifiques, sont nécessaires pour mener à bien la transcription des gènes.

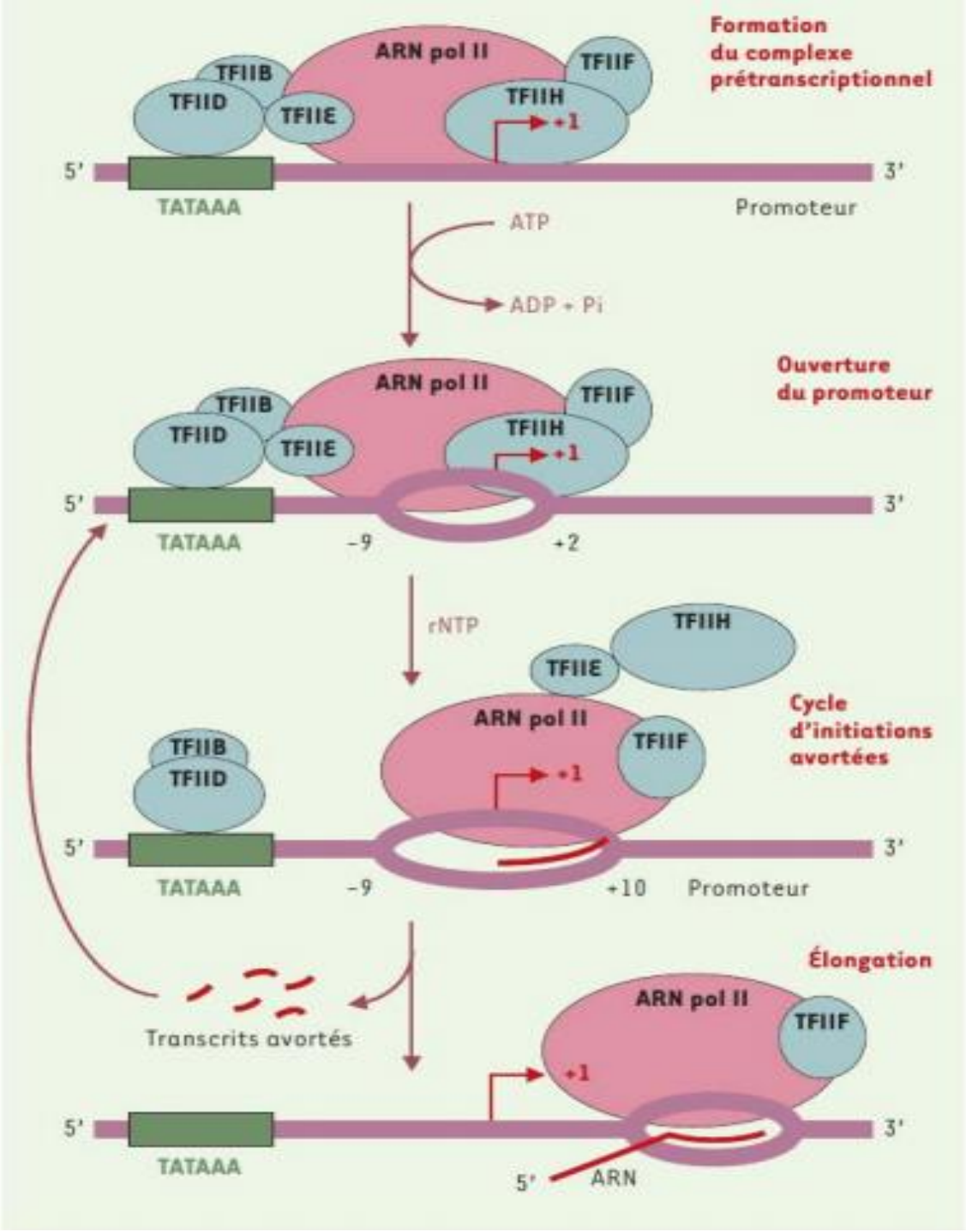


Figure 4. Initiation de la transcription et facteurs généraux de transcription.

La figure est adaptée de http://thesesups.ups-tlse.fr/1052/1/Hennion_Magali.pdf

1. 5.2 Les facteurs de transcription

Les FTs sont des protéines spécialisées dans le contrôle du niveau d'expression des gènes. Ils reconnaissent des séquences d'ADN spécifiques, généralement de 5 à 30 paires de bases (pb) de long, appelées éléments cis–régulateurs [16]. En général, ces séquences sont localisées dans les promoteurs proximaux, proches du site d'initiation de la transcription (TSS) de leurs gènes cibles. Toutefois, on peut les trouver dans des régions plus distales comme les régions activatrices (*enhancers)* et les régions répressives (*silencers)*[17]. La liaison directe à l'ADN d'un FT n'est pas toujours requise. En effet, certains FTs peuvent être recrutés sur l'ADN par d'autres partenaires par des interactions protéines-protéines. Pour assurer leurs fonctions, les FTs nécessitent l'action supplémentaire de protéines spécialisées appelées cofacteurs. Les cofacteurs transcriptionnels sont des protéines qui ne lient pas l'ADN. Ils interagissent avec les FTs pour moduler leurs activités, soit en activant ou en réprimant la transcription de leurs gènes cibles [18].

L'activité des FTs est assurée par un domaine de trans-activation. Trois types de domaines sont partagés par l'ensemble des FTs connus : un domaine riche en glutamine, un domaine riche en proline et un domaine à hélice-α acide, riche en acides aminés aspartique et acide glutamique [19]. Cependant, certains FTs ne possèdent pas un domaine de trans-activation et agissent sous formes d'homodimères ou d'hétérodimers. Il existe aussi une catégorie de FTs qui possèdent un domaine de liaison à des hormones comme c'est le cas par exemple de la famille des récepteurs nucléaires [20].

1. 6 Les familles de facteurs de transcription

Cette section est tirée de la source suivante **:**

(http://atlasgeneticsoncology.org/Educ/TFactorsFr.html)

Les FTs présentent des caractéristiques structurales communes. Ils partagent deux domaines : un domaine de liaison à l'ADN et un domaine d'activation de la transcription. Les FTs ont été regroupés en famille selon leur domaine de liaison à l'ADN (DBD : DNA

binding domain) [21]. Ce dernier détermine le type de séquence liée par chaque famille de FTs. Les FTs sont classés selon leur DBD en 4 catégories :

1. 6.1 Les protéines à homéo-domaine

Chez les eucaryotes, un grand nombre de FTs impliqués dans le développement possèdent un motif de liaison à l'ADN de 60 résidus similaires au domaine Hélice-tour-hélice des répresseurs bactériens. Ces FTs, appelés protéines à homéo-domaine ont d'abord été identifié chez la drosophile.

1. 6.2 Les protéines avec un domaine à doigt de zinc

Les doigts de zinc sont de courts domaines protéiques, d'environ 30 à 50 acides aminés selon leur type, qui fonctionnent au moins par paires. Il existe trois types de doigt de zinc dits C2H2, C4 ou C6 (où C représente un résidu cystéine et H un résidu histidine). Le motif C2H2 est le motif le plus courant codé dans le génome humain. Il possède une séquence consensus de 23 à 26 résidus avec deux résidus cystéines (C) et deux résidus histidines (H) conservés dont les chaines latérales fixent un ion de Zinc Zn2+. La structure de ce domaine est constituée de deux feuillets beta anti parallèles et d'une hélice alpha. Ce motif permet de fixer l'ADN et l'ARN. Le premier doigt de zinc a été identifié dans la protéine TFIIIA. La famille des récepteurs nucléaires contient des motifs, ZnF de type C4, qui reconnaissent une séquence d'ADN organisée en palindrome, en répétition directe ou répétition inversée.

1. 6.3 Les protéines avec un domaine Hélice-Boucle-Hélice

Le motif Hélice-Boucle-Hélice (*bHLH pour basic hélix-boucle-hélix*) a été identifié sur quelques protéines régulatrices du développement et des gènes eucaryotes qui codent pour des protéines qui lient l'ADN. Les protéines contenant ce domaine ont la capacité de lier l'ADN et de former des dimères. Elles ont en commun un motif de 40-50 acides aminés comportant deux hélices alpha séparées par une région de longueur différente en forme de boucle. Ces protéines forment des dimères via l'interaction entre les résidus hydrophobes sur les 2 hélices alpha. La plupart des protéines à domaine HLH contiennent un domaine

basique adjacent au domaine HLH, ce domaine basique (b) est nécessaire pour la liaison à l'ADN. Les protéines contenant ce domaine sont appelées les bHLH, comme les facteurs ubiquitaires E12/E47 ou bien les facteurs spécifiques de tissu comme les bHLH myogéniques.

1. 6.4 Les protéines avec un domaine à glissière à leucine

Ce domaine est un motif structural qui permet la fixation des protéines sur l'ADN. Il fait intervenir la répétition de Leucines tous les 7 résidus. L'interaction des deux hélices forme la glissière. Par exemple la protéine GCN4 possède un tel motif. Les protéines possédant ce domaine se fixent à l'ADN sous forme de dimères. Des analyses de mutagenèse ont montré que la leucine est nécessaire à la dimérisation.

Par exemple, les FTs avec un domaine formé en glissière à leucine lient l'ADN sous forme de dimères. Certains FTs, de la famille des récepteurs nucléaires ligands dépendant, possède un troisième domaine qui permet de lier le ligand.

En outre, dans certains cas, le domaine de liaison à l'ADN des FTs fournit des informations sur leur fonction. Par exemple, les FTs possédant un homéo-domaine sont souvent associées à des processus de développement, et ceux de la famille des facteurs de régulation par interférence sont généralement associés avec le déclenchement des réponses immunitaires contre les infections virales [19].

Cependant, Il est intéressant de noter que chez l'humain, la majorité des FTs connus appartiennent à trois types de famille de DBD (Figure 5). Ces résultats concordent avec des études antérieures conduites chez la souris [22].

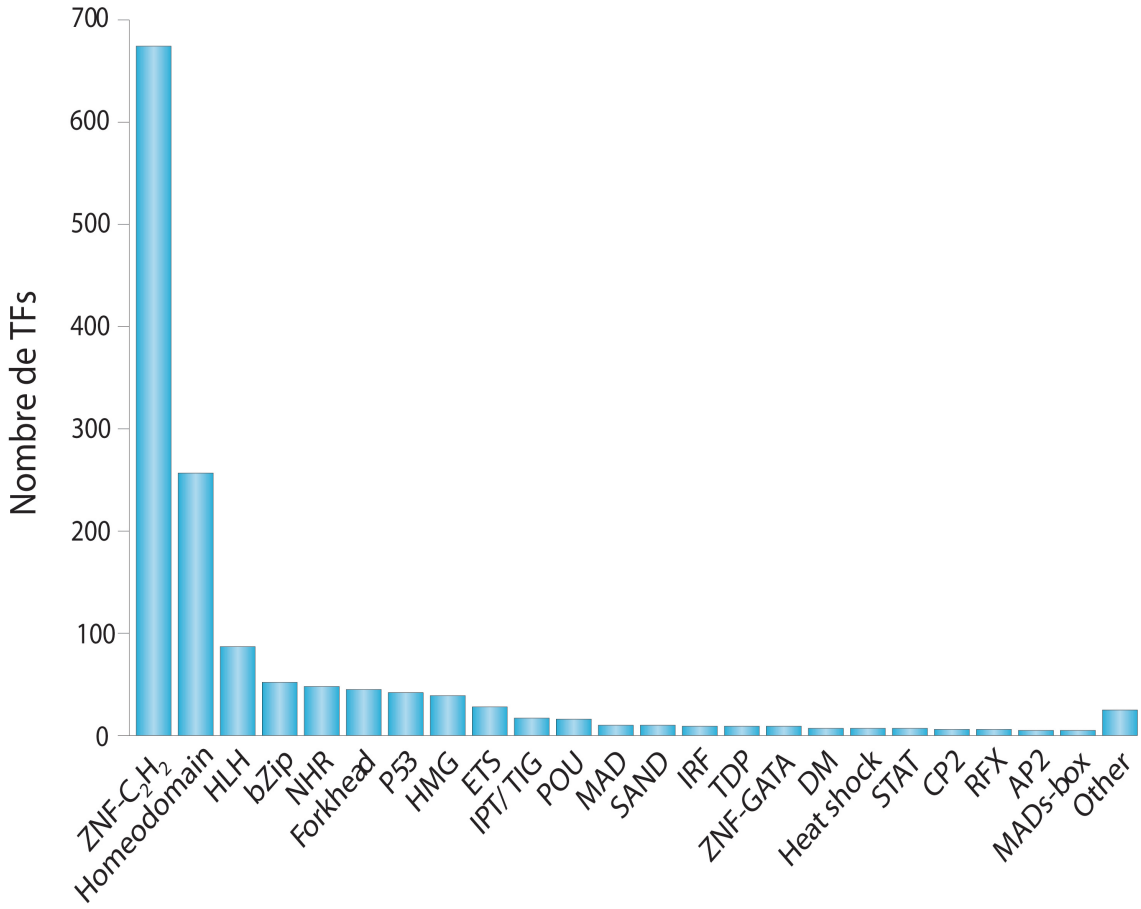


Figure 5. Classification des FTs selon leur domaine de liaison à l'ADN (DBD).

Les FTs sont classés en famille selon la composition de leur domaine de liaison à l'ADN. Les familles avec moins de cinq FTs sont classées dans other. La figure est adaptée de [23].

1. 7 Les récepteurs nucléaires

Les récepteurs nucléaires sont des facteurs de transcription formée de deux classes : les récepteurs nucléaires ligands dépendant et les récepteurs nucléaires orphelins pour lesquels le ligand n'a pas été caractérisé. Le génome humain et celui de la souris codent pour 49 et 48 récepteurs nucléaires respectivement [24]. Ils sont impliqués dans différents processus cellulaire normaux comme le développement ainsi que dans certains processus

pathologiques, comme le cancer. Un des exemples les plus connu est celui du récepteur aux estrogènes, impliqués dans le développement du cancer du sein. Les récepteurs nucléaires possèdent des propriétés structurales communes. Ils possèdent un domaine de liaison à l'ADN (DBD), dont le rôle est d'interagir avec une séquence spécifique d'ADN et le domaine de liaison du ligand (LBD). Ils sont présents sous forme d'homo ou d'hétéro-dimers qui se lient aux demi-sites d'ADN séparés par des espacements de longueur variable. Les deux demi sites peuvent être organisés dans différentes conformations : répétitions directes, palindrome ou palindrome inversé [25, 26]. Par exemple, Le récepteur des œstrogènes (ERα) reconnait des palindromes dont les demi sites (A/G)GGTCA sont espacés de trois nucléotides (IR3) [27]. (Figure 6).

1. 7.1 Le récepteur des œstrogènes ERα

Le récepteur des œstrogènes ERα, est un facteur de transcription ligand dépendant de la famille des récepteurs nucléaires. En absence de son ligand, le 17beta-estradiol (E2), le récepteur se trouve dans un état inactif dans le cytoplasme ou dans le noyau. Cette inactivation est mediée par des protéines chaperonnes HSP70 et HSP90 dont le rôle est de maintenir le récepteur dans cette conformation [16]. La présence d'E2 induit un changement de conformation du récepteur et ce dernier est libéré des chaperonnes. La liaison du ligand E2 au récepteur permet sa dimérisation. Il est connu que ERα peut former des homodimères ou des hétérodimères et peut lier l'ADN proche des sites d'initiation de la transcription (TSS) comme à de longues distances du TSS [17, 28]. Le dimère ERα ainsi formé lie une séquence d'ADN palindromique, appelée communément élément de réponse aux œstrogènes (ERE : RGGTCAnnnTGACCY), et induit la transcription de ses gènes cibles [16].

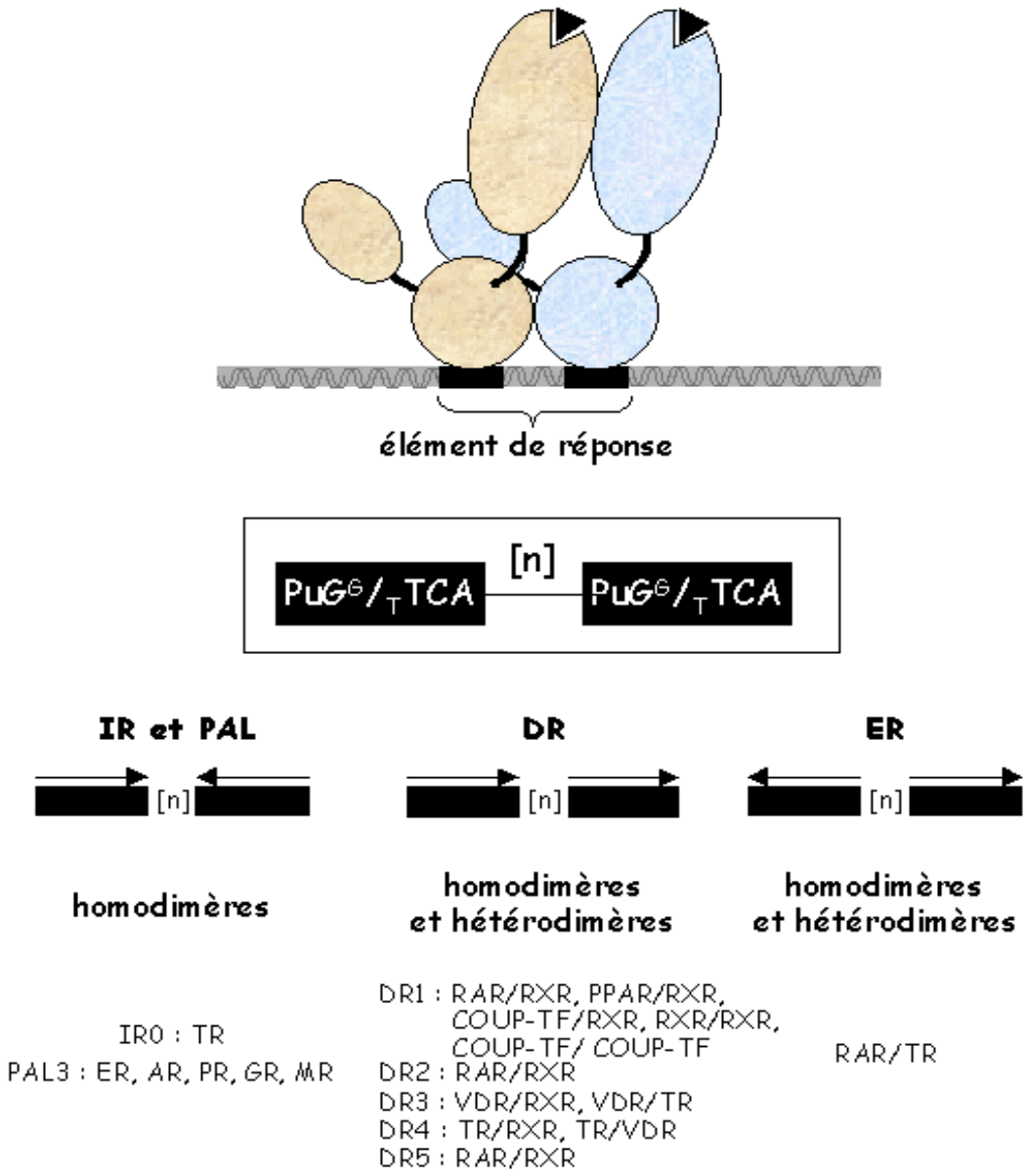


Figure 6. Mécanismes de liaison à l'ADN des récepteurs nucléaires.

Chaque rectangle noir schématise un demi-site de réponse dont la séquence consensus est indiquée (n représente le nombre de nucléotides séparant les deux demi sites. Les IR (Inverted Repeat), PAL (Palindrome Repeat), DR (Direct Repeat : répétitions directes)). La figure est adaptée de (http://scd-theses.u-strasbg.fr/842/01/html/these_body.html)

1.8 Implication des facteurs de transcription dans divers processus biologiques

Les FTs sont impliqués dans une variété de processus biologiques (Figure 7). De par leur rôle clé dans la cellule, lorsque mutés ou altérés par des modifications post-traductionnelles, les FTs sont à l'origine du développement de diverses pathologies comme le cancer. En outre le récepteur des estrogènes ER alpha (ERα), un facteur de transcription de la famille des récepteurs nucléaires, qui est impliqué dans le développement du cancer du sein [29]. Chez l'humain, on estime à 164 le nombre de FTs (12% des FTs connus) qui sont associés directement à 277 maladies ou syndromes [30].

L'étude des FTs et plus particulièrement l'identification systématique de leurs sites de fixation à l'ADN à l'échelle du génome sont des étapes incontournables pour comprendre les mécanismes de régulation de la transcription, qui sont responsables de l'expression temporelle et tissue spécifique des gènes. Cependant l'information sur les FTs n'est disponible que pour une petite fraction de FTs. Une étude récente a recensé le nombre d'articles publiés dans Pubmed traitant des différentes familles de FTs [23]. Elle a montré que certaines familles de FTs ont été beaucoup plus étudiées que d'autres (Figure 8).

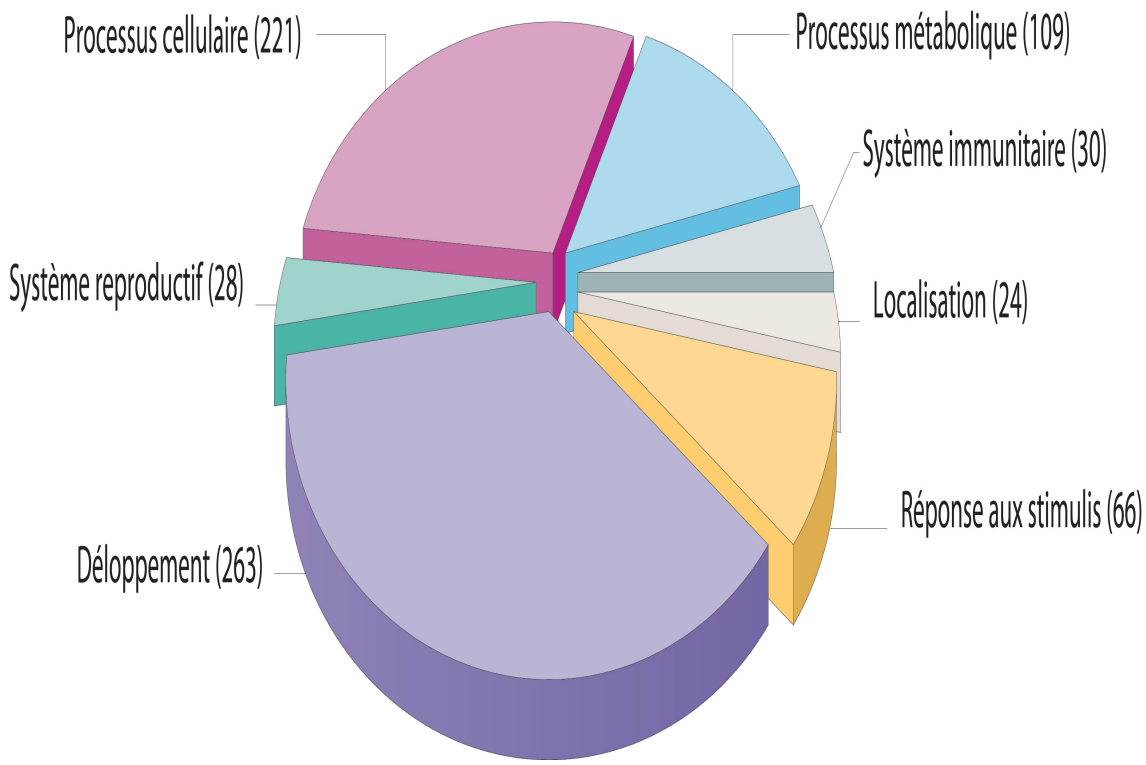


Figure 7. Implication des FTs dans différents processus biologiques.

La figure est adaptée de [23].

1.9 Les facteurs de transcriptions : quelques statistiques

Dans le génome humain, on estime le nombre de gènes codant pour des FTs à 1350 [23], 200 chez la levure et 900 chez les nématodes [31]. Cependant, l'information sur les sites d'ADN liés par ces FTs ou bien leur implication dans des complexes liant l'ADN, n'est disponible que pour une petite fraction de FTs. Par exemple chez l'humain on estime à 62 le nombre de FTs pour lesquels la liaison à l'ADN et la fonction ont été validées expérimentalement [23]. La majorité des FTs chez l'humain n'ont pas été encore caractérisé. L'information disponible sur ces FTs a été inférée à partir des autres organismes en utilisant l'information sur les gènes orthologues (Figure 9).

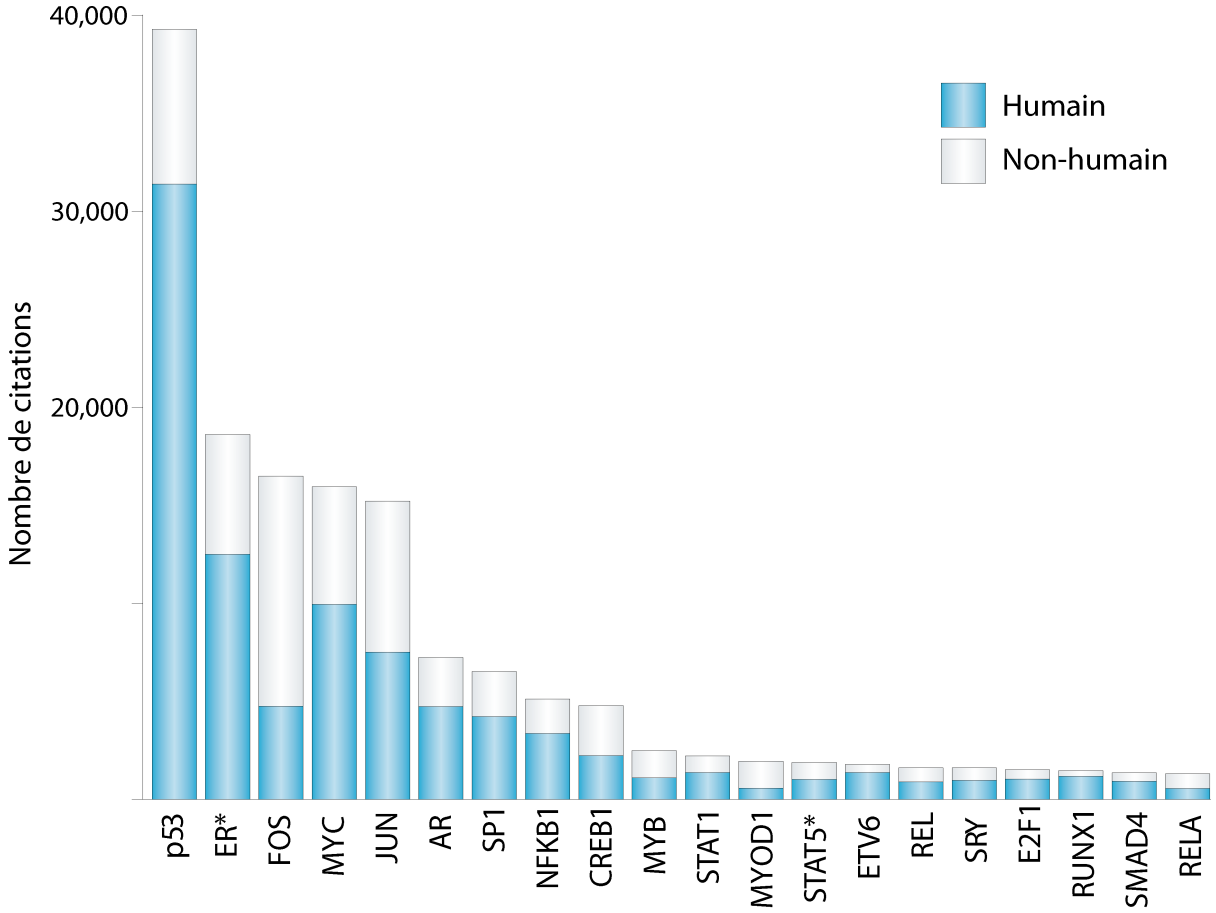


Figure 8. Nombre de citations pour les 20 FTs les plus cités dans Pubmed.

ER* représente les citations combinées pour ESR1 et ESR2. De même pour STAT5* qui combine le nombre de citations pour STAT5A et STAT5B. La figure est adaptée de [23].

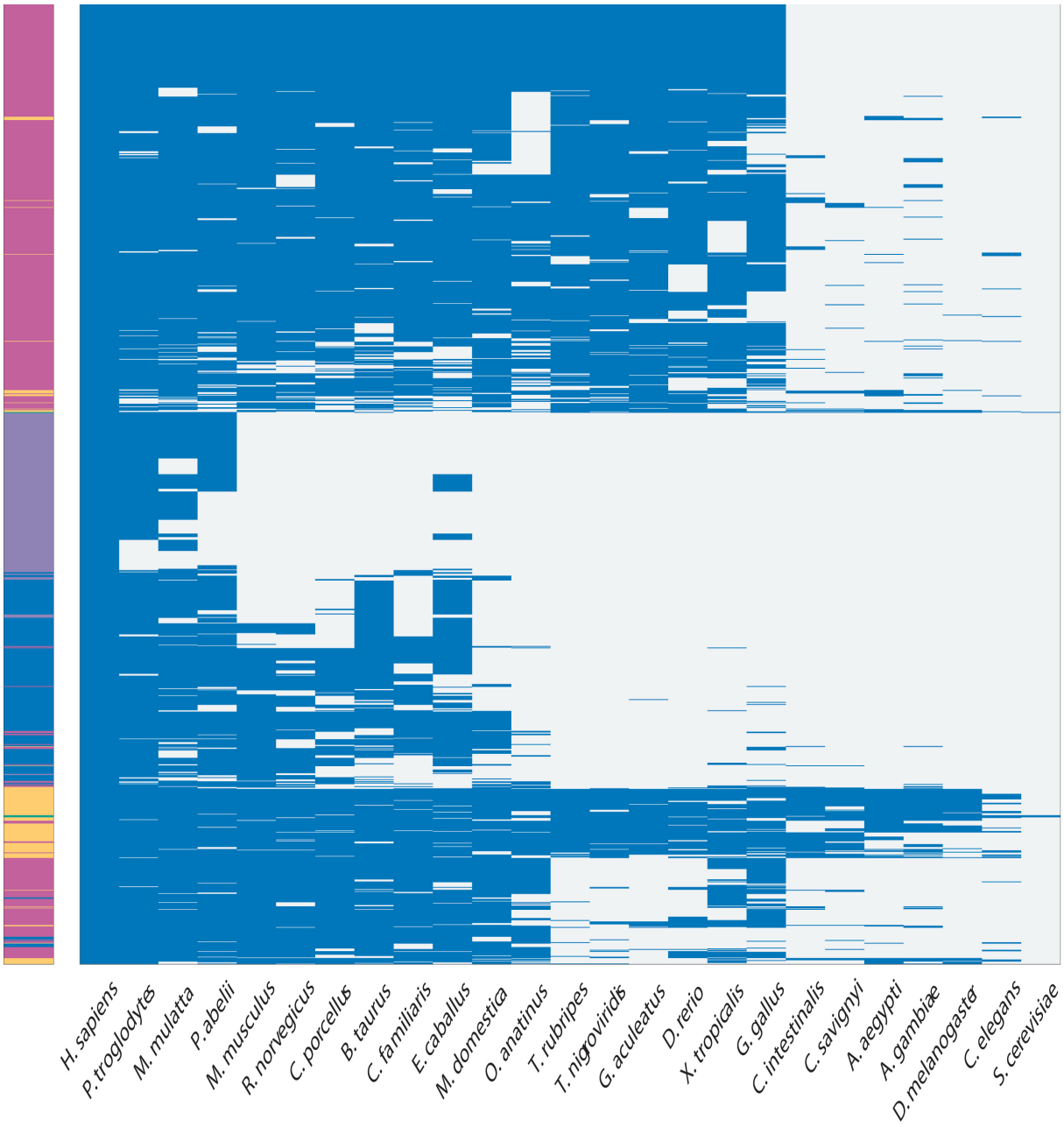


Figure 9. Conservation des FTs humain dans 24 espèces d'eucaryotes.

L'image en couleur (heatmap) représente les FTs (lignes) et les espèces (colonnes) regroupés selon l'absence ou la présence des gènes orthologues. La barre des couleurs sur la gauche indique si les FTs sont spécifiques aux primates (violet), aux mammifères (bleus), aux vertébrés (rose), aux Métazoaires (jaune) ou bien présents chez tous les eucaryotes (vert). La figure est adaptée de [23].

1.10 Expression des facteurs de transcription dans différents tissus

Les FTs lient l'ADN et agissent le plus souvent en combinaison synergétique par l'intermédiaire d'interactions protéine-protéine entre eux et avec la machinerie de transcription de base. Une des autres fonctions principales des FTs, est de recruter des cofacteurs responsables du remodelage de la chromatine qui facilitent ou bloquent l'accès à l'ADN pour d'autres FTs [32].

Dans une étude récente Vaquerizas et collaborateurs ont examiné l'expression des FTs à travers 32 tissus différents [23]. Ils ont identifié 161 FTs qui sont exprimés dans tous les tissus avec des niveaux d'expression similaires et 349 FTs qui sont sélectivement exprimés dans quelques tissus spécifiques (Figure 10). Parmi ces FTs ubiquitaires on retrouve le régulateur cardiaque CLOCK, le facteur de croissance GlI2 (Kruppel family member) et le facteur TBX1 (Tbox). Parmi les 349 FTs tissus spécifiques, 123 de ces facteurs ont un niveau d'expression différent dans un tissu comparé aux autres. Ces FTs pourraient être utilisés comme des marqueurs tissus spécifiques.

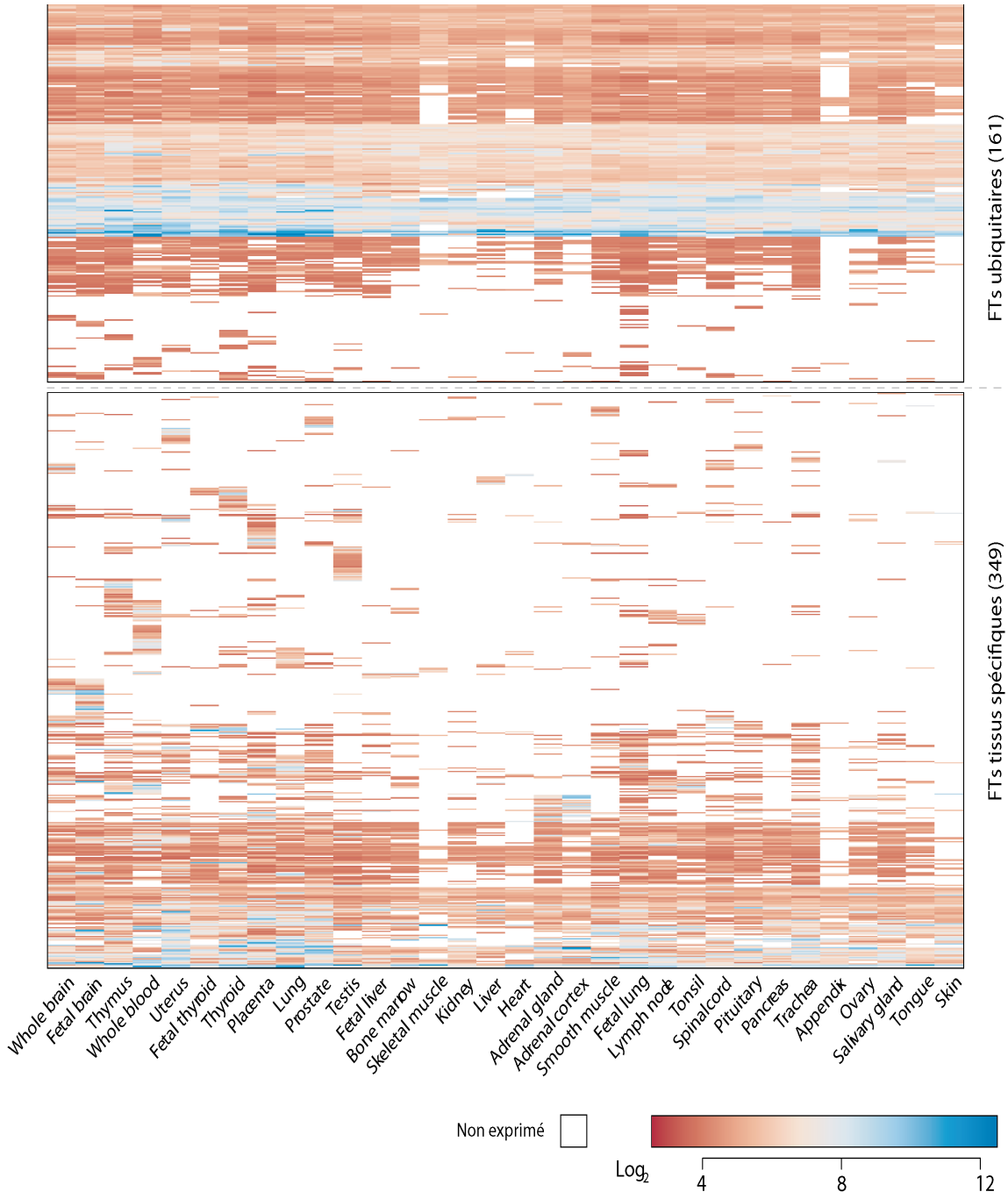


Figure 10. Heatmap représentant l'expression de FTs (RNAmessager) dans 32 organes et tissus humain.

Les lignes représentent les FTs. Les FTs spécifiques et ubiquitaires sont regroupés selon leur niveau d'expression (rouge : faible niveau d'expression, bleu : niveau d'expression élevé, blanc : non exprimé). La figure est adaptée de [23].

1.11 Les éléments cis-régulateurs

L'association entre facteur de transcription et gènes cibles est souvent déterminée par la présence ou l'absence du site lié par ce FT et sa localisation à proximité de ses gènes cibles. Ces éléments cis-régulateurs sont organisés dans différentes régions : les promoteurs, les régions activatrices (enhancers), les régions répressives (silencers) et les régions barrières (insulators) (Figure 11). Les promoteurs sont séparés en deux types de régions : promoteurs proximaux (core promoter) et les promoteurs distaux, dépendamment de type de FTs liés ainsi que leur distance du TSS. Le promoteur proximal, est situé proche du TSS et sert d'ancrage pour les facteurs généraux de transcription comme la Pol II et les facteurs de transcription de la famille TFII [33, 34]. Le promoteur distal est situé en général à moins de 1kb du TSS et contient les sites de fixation à l'ADN de certains FTs dont le rôle est de stabiliser le recrutement sur l'ADN de la machinerie basale de transcription, aidé par des co-facteurs [33, 34].

Les deux séquences les mieux étudiées sont les boîtes CAAT et les motifs riches en GC. Ces séquences d'ADN sont en général reconnues par des protéines qui régulent la transcription par des contacts directs avec la machinerie de transcription de base. C'est le cas de l'activateur transcriptionnel Sp1 qui se fixe sur les motifs riches en GC et y recrute à la fois la machinerie de transcription [35] et la machinerie de remodelage de la chromatine [36] .

Chez les procaryotes et certains eucaryotes comme la levure, les éléments cis-régulateurs sont généralement situés dans la région du promoteur proximal à quelques centaines de paires de bases du TSS. Alors que chez les eucaryotes supérieurs la distribution des éléments cis-régulateurs est beaucoup plus complexe. En effet, ils peuvent être trouvé en

amont ou en avale du TSS du coté 5’ ou 3`des gènes ou bien à des distances de l’ordre de dizaines de milliers de paires de bases du TSS [8].

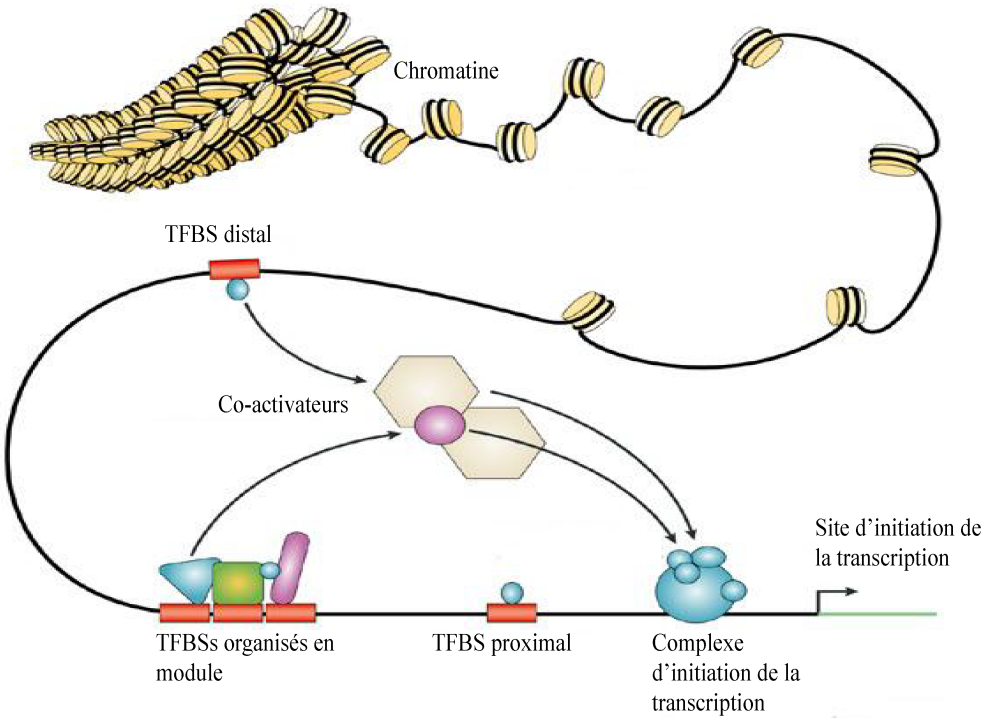


Figure 11. Distribution des éléments cis-régulateurs.

La figure est adaptée de [37].

1. 11.1 Les régions activatrices (enhancers en anglais)

Les régions activatrices sont des séquences spécifiques impliquées dans l’activation de la transcription à distance. Ces séquences peuvent être localisées autant en amont qu’en aval du TSS, et même parfois dans les exons et les introns. Elles sont reconnues par des FTs exprimés de façon ubiquitaire ou spécifique dans certains tissus. Deux modèles ont été proposés pour expliquer le mode d’action des activateurs [38]:

– Le premier modèle propose que la protéine activatrice, initialement recrutée au niveau de l'activateur, se déplace le long de l'ADN jusqu'à trouver sa cible.

– Le second modèle propose la formation d'une boucle de chromatine qui permet un regroupement spatial de l'activateur et du promoteur.

Les activateurs sont aussi impliqués dans le recrutement de certains FTs, spécialisés dans le remodelage de la structure de la chromatine afin de faciliter l'accès à d'autres régions régulatrices [39, 40].

1. 11.2 Les régions répressives (silencers en anglais)

Les régions répressives, quant à elles, sont connues pour leur rôle dans la répression des gènes en recrutant des FTs et des enzymes de remodelage de la chromatine [41]. Récemment il a été montré que les régions répressives sont impliqués dans plusieurs aspects affectant la régulation des gènes, comme la rétention cytoplasmique des FTs, la structure de la chromatine et l'épissage des introns [42]. L'activité des régions répressives est indépendante de leur orientation et de leur localisation. En effet, ces régions ont été identifiées dans les exons , les introns et dans les régions 3' UTR [42].

1. 11.3 Les régions barrières (insulators en anglais)

Les régions barrières (Figure 12) sont des régions d'ADN ayant l'une ou l'autre des deux propriétés suivantes [43, 44] :

– Lorsqu'un gène est situé entre deux insulateurs, il est protégé de l'effet de la structure de la chromatine environnante. Ici l'insulateur joue le rôle de barrière qui empêche la progression de la structure répressive et permet ainsi son expression [45].

– D'autre part, lorsqu'un insulateur est inséré entre un activateur et un promoteur, il bloque la communication entre ces deux éléments et empêche l'activation du gène. Dans ce cas on parle de rôle de bloqueur [45].

Les mécanismes d'action des insulateurs sont encore mal définis. Certaines études leur confèrent un rôle de régions frontières délimitant des territoires transcriptionels afin de

limiter l'effet des régions activatrices ou répressives sur la transcription des gènes cibles [46]. De plus, les insulateurs facilitent la formation de boucles de chromatine [38]. L'exemple le plus étudié est celui de CCCTC-binding factor (CTCF), décrit comme activateur de la transcription, répresseur de la transcription et comme barrière [46, 47].

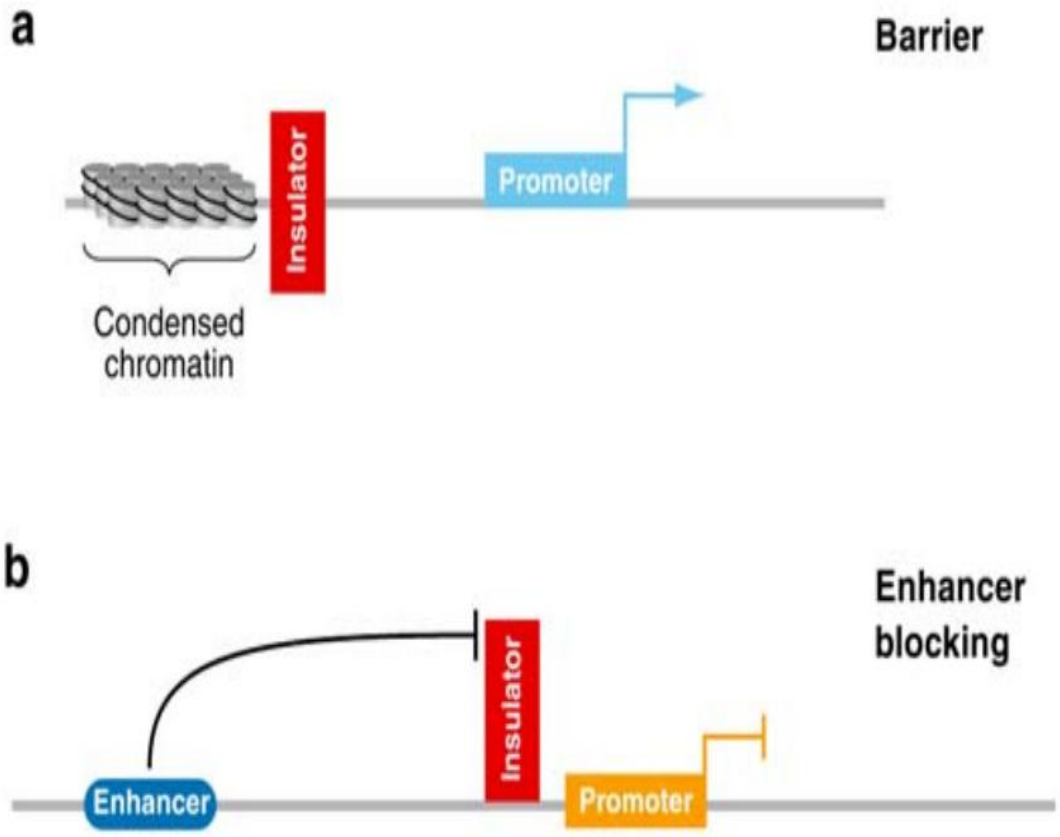


Figure 12. Propriétés des régions barrières.

Les régions barrières sont caractérisées par (a) leur capacité à bloquer la propagation de la structure chromatinienne environnante et (b) à empêcher la communication entre un activateur et un promoteur. La Figure est adaptée de [43].

Dans une étude récente, Witcher et ses collaborateurs ont montré que l'inactivation du suppresseur de tumeurs p16 (INK4a) est reliée à l'absence de la liaison de CTCF à l'ADN

[48]. En effet, dans cette étude les auteurs ont décrit l'existence d'une barrière chromatinienne en amont du gène p16 qui est perdue lorsque ce gène est anormalement inactivé. Ils ont montré que la liaison à l'ADN de CTCF au voisinage de cette barrière a été perdue. La perte de liaison à l'ADN de CTCF coïncide parfaitement avec l'inactivation du gène p16 dans plusieurs types de cellules cancéreuses.

1. 12 Bases de données des FTs

L'information sur les sites de fixation à l'ADN des FTs est répertoriée dans des banques de données à l'instar de TRANSFAC [49] et de JASPAR [50]. Ces dernières contiennent les modèles statistiques représentants les motifs liés par les FTs. Il est estimé à 900 le nombre de modèles répertoriés dans TRANSFAC à partir de sites publiés et validés expérimentalement. Dans JASPAR on a recensé moins de modèles comparé à TRANSFAC (138 modèles). Cependant, JASPAR utilise des critères plus stricts dans la sélection des sites fixés par les FTs

1. 13 Rôle de la chromatine dans la régulation des gènes

Le génome des eucaryotes est organisé en complexe protéine-ADN, appelée chromatine. La structure de la chromatine est régulée par trois mécanismes principaux : les complexes de remodelage de la chromatine qui utilisent l'énergie libérée de l'hydrolyse de l'ATP pour déplacer les nucléosomes [51], les modifications covalentes des histones et de l'ADN, impliquées dans la régulation de la structure de la chromatine [52, 53] et enfin les variantes d'histones, comme l'histone H2A.Z, connue pour son rôle comme délimiteur des régions de chromatine active [54].

1. 13.1 Modifications post-traductionnelle des histones

Les histones sont la cible de nombreuses modifications covalentes comme la phosphorylation, la méthylation, l'ubiquitination et la sumoylation [55]. Ces modifications peuvent prendre place sur les extrémités N-terminal des histones H3 et H4 ou bien dans le domaine globulaire qui forme le nucléosome. Les résidus lysines (K) peuvent être mono-,

di- ou tri-méthylés, alors que les résidus arginines peuvent seulement être mono- ou di-méthylés [53]. En général, la méthylation des résidus H3K9, H3K27, H3K64 et H4K26 sont impliquées dans l'inhibition de la transcription alors que la méthylation des résidus H3K4, H3K36 sont des marques de transcription active [56]. Un autre exemple est la métylation de la lysine 4 de l'histone H3 (H3K4), généralement associée avec l'activation de la transcription ainsi que le recrutement de facteurs de remodelage de la chromatine [57]. Par exemple, les histones H3 et H4 sont acétylées sur les résidus lysine, situés dans leurs extrémités N-terminal par des enzymes acétyl-transférase (HAT). Cette acétylation neutralise la charge positive des lysines et permet ainsi la décondensation de la chromatine [58]. Une autre catégorie d'enzymes, les désacétylases d'histones (HDAC) sont responsables de contrer l'effet des HAT en rétablissant la structure condensée de la chromatine [59].

Les modifications covalentes des histones ne sont pas indépendantes les unes des autres. Par exemple, la phosphorylation de la Ser10 de l'histone H3, favorise l'acétylation de la lysine 14 (H3K14), qui à son tour précède la méthylation de la lysine 9 (H3K9) [60].
Un autre exemple est celui de la méthylation de l'Arg3 de l'histone H4 qui précède et favorise l'acétylation des Lysines 8 et 12 [61].

1. 13.2 Le positionnement du nucléosome

Le positionnement des nucléosomes dans le génome joue un rôle important dans l'accessibilité des FTs à leurs sites de fixations à l'ADN, conférant au nucléosome un rôle d'activateur et de répresseur de la transcription [62]. Par conséquent, pour prédire efficacement les sites de fixation de FTs à l'ADN, il est nécessaire de déterminer à priori où sont localisés les nucléosomes à travers le génome et quels sont les mécanismes qui gouvernent cette distribution.

In vivo, trois mécanismes sont impliqués dans la réduction de l'occupation des nucléosomes à certaines régions :

(a) Des complexes de remodelage de la chromatine sont recrutés par des protéines activatrices pour générer des régions pauvres ou dépletées de nucléosomes et faciliter la transcription [63].

(b) Lors du cycle d'élongation par la polymérase II, les nucléosomes sont désassemblage/réassemblage dans les régions codantes, ce qui génère pour les régions hautement transcrites une diminution de la densité nucléosomique [64].

(c) Enfin, il a été montré que les nucléosomes ont une préférence pour certaines séquences d'ADN et qu'à l'inverse, d'autres types de séquences leurs sont défavorables [65, 66]. C'est le cas des séquences Poly(dA :dT), généralement de longueur 10 à 20 nucléotides, qui sont connues pour résister aux distorsions nécessaires à la formation des nucléosomes. Par conséquent, ces séquences contribuent à créer des régions déplétées de nucléosomes, ce qui pourrait favoriser la fixation des FTs dans ces régions. Le profile de distribution des nucléosomes est tissus spécifiques. Il diffère d'une espèce à une autre dépendamment du stage du développement et du type cellulaire.

1. 13.3 La méthylation de l'ADN

La méthylation de l'ADN sur les résidus Cytosine (C) des di-nucléotides CpG, joue un rôle dans la régulation de l'expression des gènes. Il s'agit de l'addition d'un groupement méthyle en position 5 des cytosines ou en position 6 des adénines. Cette méthylation influence la flexibilité de l'ADN et par conséquent son affinité aux nucléosomes et les interactions FT-ADN [67]. Les CpG méthylés sont généralement localisés dans les régions répétées du génome et sont responsables de l'inactivation de la transcription ainsi que l'immobilité des éléments transposables [68]. Chez les mammifères, trois enzymes, des méthyl-transférase, sont responsables de la méthylation de l'ADN : Dnmt1, Dnmt3A et Dnmt3B. L'importance du mécanisme de méthylation de l'ADN a été montrée chez la souris, où il a été observé que la perte de l'une de ces enzymes est létale [69]. Ces modifications agissent sur l'affinité de l'ADN à l'octamère d'histones et par conséquent module la structure de la chromatine qui à son tour affecte la transcription. La nature

dynamique et réversible de ces modifications permet une restructuration continuelle de la chromatine pour activer ou inhiber la transcription.

1. 14 Comment les facteurs de transcription reconnaissent leurs cibles sur l'ADN ?

Les FTs peuvent reconnaitre plusieurs patrons de séquences qui divergent d'un motif consensus. La liaison des FTs à ces différentes séquences se fait avec une affinité suffisante pour assurer la régulation des gènes cibles. Cependant, la présence de cette séquence à elle seule n'implique pas la liaison du FT à l'ADN et ne définit pas la fonctionnalité de ce site [70-72]. Chez les eucaryotes, la structure de la chromatine est intimement impliquée dans les interactions FT-ADN et les interactions protéines-protéines. Dans sa forme compacte, le nucléosome masque complètement l'ADN et empêche l'accessibilité des FTs à leur site de fixation. Mais cette structure est dynamique et change sous l'action des facteurs de remodelage de chromatine pour décompacter cette dernière et permettre ainsi l'accès à l'ADN [73, 74].

Les interactions FT-FT peuvent aussi faciliter l'accès aux régions d'ADN liées [75]. Elles peuvent intervenir soit entre des FTs qui lient des sites proches ou bien à travers la formation de boucles de chromatine [76]. Par exemple, Lupien et collaborateurs [77] ont montré que FOXA1 est impliqué dans l'ouverture de la chromatine afin de faciliter le recrutement du récepteur des œstrogènes ERα à ses sites d'ADN et activer la régulation des gènes cibles. Ces interactions peuvent dans certaines situations, bloquer la fixation à l'ADN d'un FT au lieu de la favoriser. Ce phénomène peut avoir lieu par exemple dans le cas de deux FTs dont les sites de fixation à l'ADN se chevauchent [78]. Un autre exemple est celui mis en évidence par Holmes et ses collaborateurs. Ils ont montré dans la lignée de cellules du cancer du sein MCF7 traitées aux œstrogènes, que LEF-1 et Nkx3-1 sont recrutés sur l'ADN et empêchent la liaison du récepteur des œstrogènes ERα à certains sites [79]. Ils ont émis l'hypothèse que ce blocage est médié par l'association de ces deux FTs avec des désacétylases d'histones (HDAC1) et augmente la condensation de la chromatine.

Chapitre 2

2. Approches expérimentales pour la caractérisation des interactions ADN-protéines

Les interactions protéines-ADN constituent l'un des mécanismes les plus importants qui gouvernent la régulation de la transcription. Des altérations dans la fonction de certains FTs sont souvent la cause de maladies comme le cancer, à l'instar du récepteur des estrogènes qui est surexprimé dans le cancer du sein et le récepteur des androgènes dans le cancer de la prostate [80]. Par conséquent, plusieurs efforts ont été entrepris pour caractériser les profile de liaison à l'ADN de tous les FTs, c'est le cas par exemple du projet ENCODE [81].

Plusieurs approches expérimentales ont été développées pour identifier les sites liés par un FT donné. Dans ce qui suit, nous allons passer en revue quelques unes des approches les plus utilisées :

2.1 La technique du retard sur gel ou EMSA (Electrophoretic Mobility Shift Assay)

La technique EMSA repose sur le fait qu'un complexe ADN-protéine ou ARN-protéine migrera moins vite dans un gel non-dénaturant que l'ADN ou l'ARN nu. La migration de ce complexe nous permet de juger si la séquence d'ADN ou d'ARN a été reconnue par une protéine [82]. Afin de vérifier la spécificité de l'interaction ADN-protéine, ce complexe est incubé au préalable avec un anticorps qui reconnait la protéine d'intérêt. Le complexe formé par ADN-protéine-anticorps migrera moins vite que le complexe ADN-protéine. .

Cependant, ces approches permettent uniquement la caractérisation de quelques séquences liées par le FT ciblé. En effet, Ces techniques n'offrent pas une représentation réelle et

globale du profile de toutes les variations de la séquence d'ADN pouvant être liée par ce FT.

2.2 SELEX (Systemtic Evolution of Ligands by Exponentiel enrichment)

Cette technique permet la sélection progressive des séquences d'ADN interagissant avec une ou plusieurs protéines. Son principe de base consiste à amplifier un ensemble d'oligo-nucléotides de séquences aléatoires. Dans cet ensemble de séquences, on y trouve celles reconnues par la protéine d'intérêt et d'autres qui ne sont pas liées. Ensuite une expérience EMSA est réalisée pour choisir les séquences reconnues par la protéine d'intérêt. Les séquences ainsi obtenues sont à nouveau amplifiées et les bandes retardées sur le gel sont purifiées. Ce cycle peut être exécuté plusieurs fois. SELEX est une expérience in vitro qui nécessite de scanner un large éventail de clones pour assurer une bonne efficacité.

2.3 La technique PBM (Protein Binding Microarray)

Une autre technique qui fait appel à la technologie des micropuces à ADN a été proposée par Badis et collaborateurs [83]. Cette technique, désignée sous le nom de PBM (protein binding micro-array) est une approche *in vitro* pour l'identification de toutes les séquences liées par une protéine d'intérêt. Elle consiste à générer par une approche computationnelle un lot de séquences d'ADN double brin d'une longueur spécifique. Ces séquences sont hybridées par la suite sur une micropuce. Une protéine purifiée, ayant été préalablement marquée ou étant révélée par un anticorps marqué, est mise en contact avec la micropuce. Parmi les inconvénients de cette technique est qu'elle nécessite des quantités élevées de la protéine purifiée et elle est limitée aux oligomères dont la taille ne dépasse pas les 10 bp.

2.4 In vivo versus in vitro

Les techniques *in vivo* pour l'identification des interactions ADN-protéines représentent l'affinité réelle du FT à sa séquence d'ADN. En général, les séquences identifiées *in vitro*

présentent un biais pour celle liées *in vivo* [64]. Les différences observées entre l'affinité de liaison *in vivo* versus *in vitro* pourraient être expliquées par :

(a) Le FT peut interagir avec d'autres protéines *in vivo,* qui pourraient induire un changement dans sa conformation et par conséquent une affinité pour un autre site de liaison à l'ADN différent de celui lié *in vitro.*

(b) L'effet de la chromatine sur l'accessibilité aux sites de liaison à l'ADN *in vivo*. En effet, certaines régions d'ADN accessibles *in vitro* peuvent être masquées *in vivo*.

Le développement des techniques génomiques à large échelle comme la technique du ChIP-chip et du ChIP-séquençage a contribué largement à l'identification de nouveaux motifs d'ADN fixés par les FTs [84, 85]. Ces techniques permettent de localiser à l'échelle du génome l'ensemble des régions liées *in vivo* par un FT d'intérêt.

2.5 La technique de ChIP

La compréhension des mécanismes de régulation de la transcription nécessite une identification à l'échelle du génome des interactions ADN-protéines ainsi que les marques d'histones. La technique par excellence pour investiguer ces mécanismes est la technique d'immuno-précipitation de la chromatine (ChIP), une technique *in vivo,* utilisée pour localiser les régions génomiques liées par une protéine d'intérêt ou bien des modifications post-traductionnelles des histones [86]. La première étape consiste à lier de manière covalente (pontage) les protéines et l'ADN en utilisant un agent de fixation, comme le formaldéhyde. L'ADN est ensuite fragmenté en utilisant soit une méthode mécanique, qui est la sonication, ou bien une digestion en utilisant des enzymes spécifiques comme nucléase micrococcale (MNase). Cette étape génère des fragments courts dont la taille varie entre 200 bp à 1kb en moyenne. Ensuite, un anticorps spécifique à la protéine d'intérêt est ajouté aux complexes afin d'immuno-précipiter tous les fragments d'ADN liées. Après une étape de lavage pour éliminer tous les fragments d'ADN non fixés (non liés par la protéine d'intérêt ainsi que les liaisons faibles ou non spécifiques) le pontage est renversé par la chaleur pour séparer les complexes protéine-ADN. Les fragments d'ADN ainsi obtenus

sont amplifiés par PCR en vu de leur identification [87]. Dans le cas où l'expérience de ChIP cible des nucléosomes ou bien les modifications d'histones, les fragments sont digérés à la nucléase micrococcal (MNase) plutôt que d'être fragmentées par sonication [86].

2.5.1 Limitations de la technique de ChIP

Bien que la technique de ChIP soit considérée comme la technique *in vivo* par excellence pour identifier les régions liées par un FT d'intérêt, elle présente plusieurs limitations. Parmi lesquelles on peut citer :

- Souvent, le signal de ChIP n'est pas assez élevé pour le distinguer du contrôle utilisé.
- La position exacte du site lié n'est pas connue à cause de la résolution de la technique.
- Elle n'offre aucune information sur la fonctionnalité du site/région identifié(e).
- Limitée par la qualité des anticorps utilisés.

Le développement des techniques de micro-puces d'ADN et celles du séquençage à haut débit ont permis de coupler la technique de ChIP à ces deux nouvelles approches dans le but d'identifier toutes les régions génomiques liées par la protéine d'intérêt. De plus, ces techniques offrent la possibilité de générer de nouveaux modèles pour le positionnement des nucléosomes et une cartographie à large échelle des modifications post-traductionnelles des histones [87].

2.6 La technique de ChIP-chip

La technique de ChIP-chip [88] combine la technique d'immuno-précipitation de la chromatine et les puces à ADN [12]. Elle offre un grand avantage comparée à la technique de ChIP. En effet, au lieu de cibler un nombre limité de régions génomiques, dont la sélection est souvent biaisée, des milliers voir des millions de régions pourraient être explorées dans une même expérience. Pour certaines espèces, il est possible d'utiliser des micro-puces qui couvrent la totalité du génome (utilisation de sondes à chaque 10-100bp).

Pour les génomes de taille plus large, seulement une région pourrait être ciblée (les promoteurs des gènes, des chromosomes spécifiques). Elle permet ainsi de découvrir de nouvelles régions génomiques liées par la protéine d'intérêt. Dans une expérience de ChIP-on-chip, après une étape de ChIP, les fragments liés sont d'abord amplifiés par PCR avant d'être marqués par un fluorochrome (Cy5). De la même façon, l'ADN contrôle (qui n'est pas enrichi par la protéine d'intérêt) est amplifié ensuite marqué par un fluorochrome différent (Cy3). Les deux lots de fragments (liés et non liés) sont ensuite hybridés sur une puce à ADN (tilling array). La puce est lue par un laser et l'intensité des deux signaux est lue afin de calculer pour chaque région le ratio Cy5/Cy3 et d'identifier les régions génomiques pour lesquelles le signal est enrichi significativement par rapport au contrôle (Figure 13).

Cependant, la résolution de la puce utilisée, est un paramètre très important dans l'identification de toutes les régions liées.

2.7 La technique de ChIP-séquençage (ChIP-seq)

L'avènement de la technologie de séquençage (next-generation sequencing NGS) a permis le séquençage de millions de fragments d'ADN et offre une résolution beaucoup plus élevée comparée à la technique de ChIP-chip. En l'occurrence, la technique de séquençage à été utilisée dans plusieurs applications comme le séquençage du génome, séquençage des ARN messagers pour les études de l'expression de gènes (RNA-Seq) et l'identification des interactions ADN-protéines (ChIP-Séquençage).

Dans la technique de ChIP-séquençage (ChIP-Seq), les fragments d'ADN liés sont séquencés au lieu d'être hybridés sur une puce [87].

Après une expérience de ChIP, les fragments obtenus sont ensuite utilisés pour construire une librairie pour le séquençage. Une présélection des fragments est réalisée pour en choisir ceux dont la taille varie entre 150-300 bp. Des petits fragments de 30-70 nucléotides sont séquencés de chaque extrémité et sont ensuite séquencés. Les fragments sont distribués de manière comparable sur les deux brins d'ADN. Plusieurs type de plateformes de

séquençage sont disponibles : le séquenceur Illumina, Applied Biosystem'SOLiD et la plate forme Helicos [89]. Ces plateformes peuvent générer jusqu'à 100-400 millions de fragments séquencés (reads) en un seul tour et 60-80% de ces fragments peuvent être alignés à des régions unique dans le génome de référence [87].

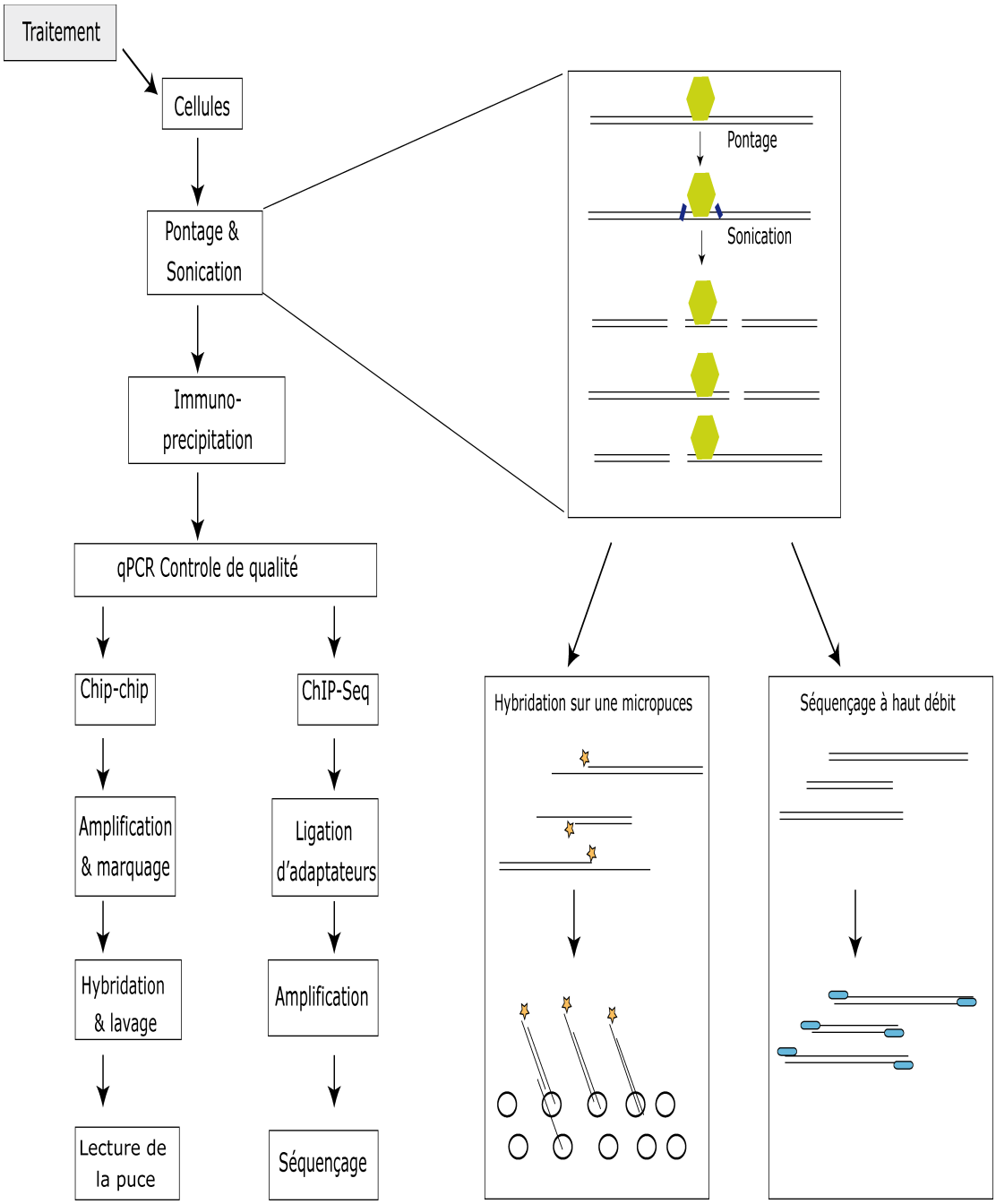


Figure 13. Schéma de la technique de ChIP-chip et de ChIP-séquençage.

2.7.1 Avantages de la technique de ChIP-seq comparée à la technique de ChIP-chip

La résolution de la technique de ChIP-seq, qui est de l'ordre d'une centaine de paires de bases, est l'une de ses caractéristiques les plus importantes (Figure 14). Dans le cas du ChIP-chip, on aurait besoin d'un très grand nombre de sondes pour avoir une résolution comparable. Ceci est quasiment impossible pour les grands génomes, comme ceux des mammifères, à cause du coup très élevé de la technique. Par ChIP-seq, on peut obtenir une couverture de tout le génome. Tandis que par ChIP-chip cela dépendra du nombre, de la qualité et le type des sondes utilisées. Ceci permet une identification plus précise des sites liés par les FTs, des modifications post-traductionnelle de la chromatine ainsi que les positions des nucléosomes.

La technique de ChIP-chip souffre aussi du problème d'hybridation des sondes et pourraient générer un taux élevé de faux positifs. En effet, le processus d'hybridation est affecté par plusieurs facteurs comme la composition en CG et la séquence des probes. Plusieurs analyses ont montré qu'une fraction des régions observées par ChIP-seq, n'ont pas été identifiées par ChIP-chip [87, 90].

La technique de ChIP-seq offre une meilleure résolution, moins d'artéfacts et une plus vaste couverture de la région d'intérêt, comparée à la technique de ChIP-chip [87]. En effet, avec la technique de ChIP-seq le génome entier peut être séquencé. De plus, la technique de ChIP-seq est plus appropriée pour l'analyse des régions répétées dans le génome, qui sont masquées dans le cas du ChIP-chip [90].

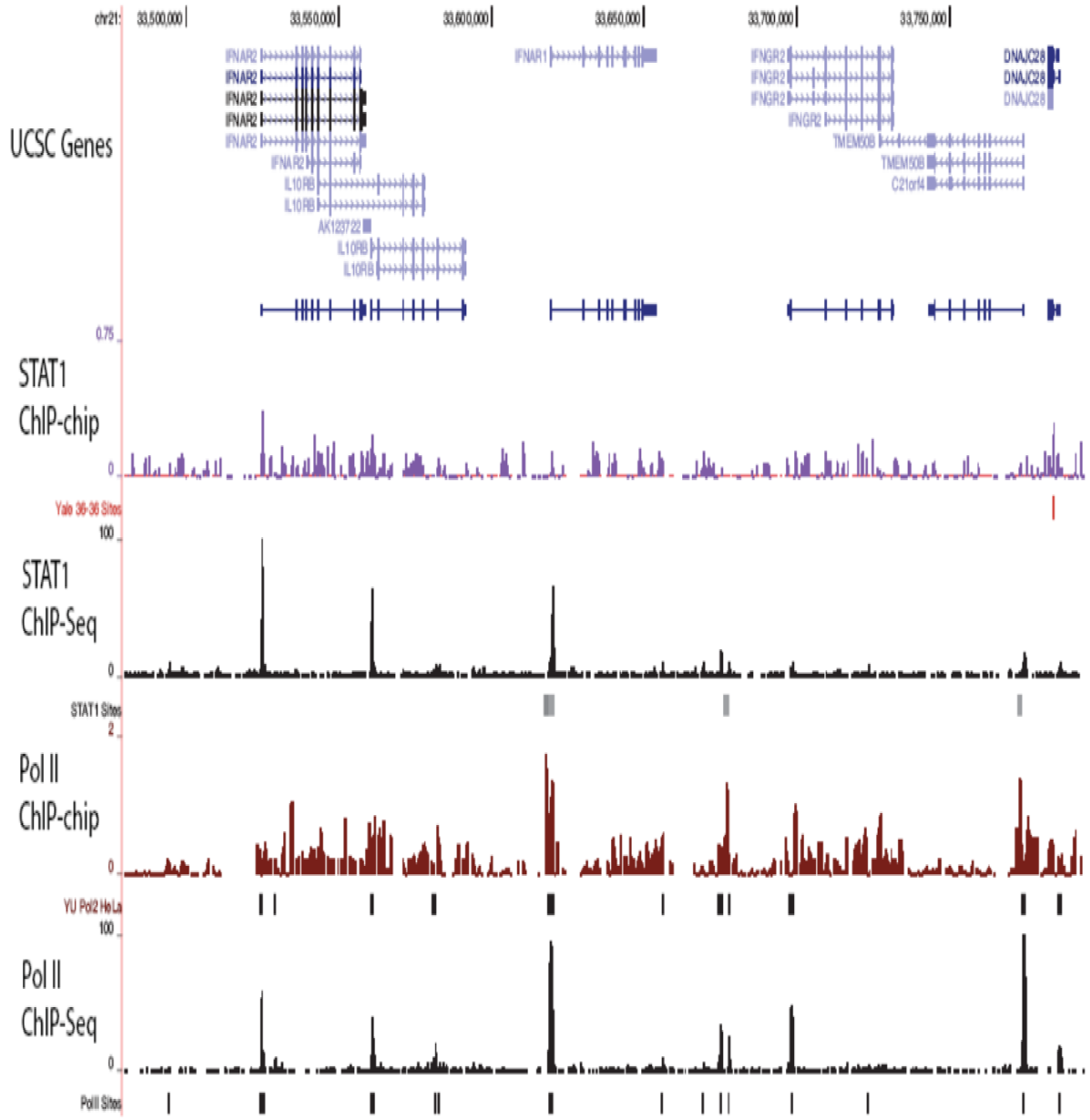


Figure 14. Comparaison du profile de liaison à l'ADN d'un FT et de la Pol II par ChIP-chip et ChIP-seq.

2.7.2 Limites de la technique de ChIP-seq

Quoiqu'elle offre de meilleurs résultats comparés à la technique de ChIP-chip, la technique de ChIP-seq souffre de plusieurs biais qui affectent le résultat final. A titre d'exemple, on peut citer le biais au contenu en CG dans la sélection des fragments à séquencer ainsi que pendant l'amplification avant et après les avoir séquencés. Les régions riches en CG ne sont pas bien amplifiées [87]. Cette technique est aussi très sensible à la quantité de matériel utilisé.

2.8 Limites et avantages des techniques *in vivo*

Les techniques *in vivo* à large échelle comme la technique de ChIP-chip et de ChIP-seq, ont permis de générer le profile de liaison à l'ADN de plusieurs FTs et par conséquent contribuer à mieux comprendre les mécanismes d'interaction FT-ADN et leur rôles dans la régulation des gènes cibles. Ces techniques ciblent un FT donné dans des conditions particulières de la cellule. Ceci génère un profile de liaison à l'ADN qui est spécifique aux conditions de l'expérience et ne pourrait être généralisé de manière systématique pour ce même FT.

Dans le cas des techniques *in vitro,* la question est de pouvoir associer les sites identifiés à une ou plusieurs conditions *in vivo.* En effet, même si le FT lie le site avec une très haute affinité *in vitro*, on ignore quand et ou ce site est lié *in vivo* par ce même FT.

2.8.1 Qualité des anticorps utilisés

Un anticorps spécifique à la protéine d'intérêt permet d'obtenir des niveaux d'enrichissement significatifs comparé au contrôle utilisé. Ceci rend facile l'identification des régions liées par la protéine ciblée par l'anticorps. Quand il s'agit d'identifier des marques de chromatines, le choix de l'anticorps est encore plus sensible. En effet, certains anticorps ne sont pas spécifiques et peuvent identifier différentes marques d'histones [91].

2.8.2 Choix du contrôle

Afin de permettre une identification efficace des régions liées, il est important de choisir le bon contrôle. Plusieurs types de contrôles sont utilisés : l'ADN total (input DNA sans immuno-précipitation) ; de l'ADN obtenu par immuno-precipitation mais sans anticorps ou bien de l'ADN non spécifique, obtenu en utilisant un anticorps contre une protéine connue pour ne pas lier l'ADN [92]. Dépendamment du contrôle utilisé, les régions enrichies sont identifiés comme étant celles pour lesquelles le ratio entre signal de ChIP et celui observé dans le contrôle est significatif. Dans une expérience de ChIP-seq par exemple, le nombre de fragments qui couvrent une région liée peut être de l'ordre de milliers. Cependant, il arrive que des régions de forte activité transcriptionnelle ne soient pas identifiées à cause de la faible intensité du signal mesuré dans ces régions. [91].

2.9 Utilisation de la technique de ChIP-seq pour étudier les mécanismes épigénétiques

La chromatine est formée de l'association des histones avec l'ADN. Elle contient l'information sur les modifications covalentes des histones, importantes pour comprendre la fonction et l'activité de ces régions et leur implication dans la régulation des gènes [93]. Les modifications post-traductionnelles des histones nous renseignent sur l'accessibilité des régions d'ADN aux FTs. Par exemple Mikkelsen et collaborateurs [94] ont observé des changements dans certaines marques de chromatine pendant la différentiation des cellules ES chez la souris. Ces changements se traduisent par des patrons d'expression différents des gènes. De même, Heintzman et collaborateurs [93], à leur tour ont montré que la méthylation de la lysine 4 de l'histone H3 (H3K4me1) est enrichie dans les régions activatrices (enhancers) alors que la tri-méthylation de la lysine 27 (H3K27me3) caractérise les régions inactives de chromatine. Chez la levure, il a été montré que la tri-méthylation de la lysine 36 de l'histone H3 (H3K36me3) est associée aux gènes actifs et souvent distribuée le long de la séquence de ces gènes [95].

Plusieurs marques de chromatines ont été analysées par ChIP-chip et ChIP-seq en utilisant des anticorps qui ciblent des peptides portant des modifications spécifiques comme la méthylation, l'acétylation et la phosphorylation. Ces informations, associées aux profiles de liaison à l'ADN des FTs permettent d'améliorer considérablement la prédiction des sites liées par ces derniers [96, 97].

Chapitre 3

3. Identification et analyse des régions liées par les facteurs de transcription

3.1 Alignement des fragments séquencés au génome de référence

La plupart des plateformes de séquençage offrent un logiciel spécifique pour l'alignement des fragments séquencés au génome de référence. Le choix de la méthode d'alignement est très important étant donné le nombre de fragments générés par cette technique. En effet, ces logiciels doivent tenir compte de plusieurs paramètres comme la qualité de l'alignement et le temps nécessaire pour aligner tous les fragments. Parmi ces logiciels on peut citer le logiciel Eland et Stampy de la plateforme Illumina [98]. D'autres logiciels comme MAQ (Mapping and Assembly with qualities) [99] et Bowtie [100] sont indépendants de la plateforme utilisée. Afin d'assurer une bonne qualité d'alignement, la plupart des logiciels d'analyse des données de ChIP-seq considèrent uniquement les fragments qui sont alignés à une région unique du génome. Cette étape nécessite des logiciels très performants étant donné le nombre de fragments générés qui sont de l'ordre de 10^7 à 10^8.

3.2 Analyse des régions liées : outils d'identification des pics

Le passage des données brutes, obtenues par ChIP-chip ou ChIP-seq, à un type de données qui pourraient être exploité à fin de répondre aux questions biologiques posées nécessite deux étapes :

- Aligner les régions obtenues au génome de référence.
- Identifier les bordures génomiques des régions liées où l'intensité du signal (ChIP-chip) ou bien le nombre de fragments séquencés (ChIP-Seq) est significativement plus enrichit comparé au contrôle.

Par exemple, la technique de ChIP-chip permet de mesurer pour chaque sonde de la puce, sa valeur d'enrichissement, sous forme d'un ratio (valeur d'enrichissement) du signal de ChIP versus le contrôle, généralement calculé dans une échelle logarithmique. Dans la technique de ChIP-seq on évalue l'enrichissement d'une région selon le nombre de fragments séquencés qui sont alignés à cette région. Le résultat pourrait être représenté dans différent formats (i.e .bed, .wig). Afin de visualiser graphiquement ces régions, on peut utiliser différents outils à l'instar du Génome browser de UCSC (http://genome.ucsc.edu/) ou signalMap de NimblGen (http://www.nimblegen.com/products/software/signalmap).

Après avoir aligné les fragments séquencés au génome de référence, l'étape suivante consiste à identifier les régions génomiques, enrichies de ces fragments, comparés au contrôle. En effet, ces fragments sont distribués à travers tous le génome. L'intensité du signal de couverture varie d'une région génomique à une autre dépendamment du nombre de fragments qui couvrent chacune de ces régions. L'identification des régions enrichies se fait à l'aide de logiciels d'identification de pics. Plusieurs approches ont été développées à cette fin [101]. En résumé, la procédure d'identification de pics consiste à utiliser une fenêtre de longueur fixe pour scanner le génome et compter le nombre de fragments qui couvrent chacune des régions génomique identifiées par ChIP. La significativité de la distribution du nombre de fragments dans chaque région est ensuite évaluée selon des critères spécifiques à chaque outil, comparé au contrôle utilisé. La table 1 résume quelques outils utilisés pour l'identification des pics.

Certaines approches d'identification de pics utilisent la bidirectionnalité des fragments séquencés pour estimer deux distributions différentes mais proches l'une de l'autre, l'une sur le brin positif et l'autre sur le brin négatif [102]. Une distribution combinée est estimée à partir des deux précédentes en les déplaçant vers le centre ou en faisant une extension des fragments dans les deux sens (Figure 15). Plusieurs approches sont utilisées pour mesurer l'enrichissement des régions identifiées comme des pics. L'approche la plus simple consiste à calculer le niveau d'enrichissement comme le ratio entre le nombre de fragments qui couvrent la région liée comparée au contrôle. Cependant, un niveau d'enrichissement de 10

obtenu à partir d'un ratio de 1000/100 est statistiquement différent de celui obtenu d'un ratio 100/10. L'estimation de la distribution par une loi de Poisson [94] ou une loi Binomial [92] donne une meilleure représentation de l'enrichissement observé car elles ne tient pas compte uniquement du ratio mais aussi du nombre de fragments alignés dans ces régions.

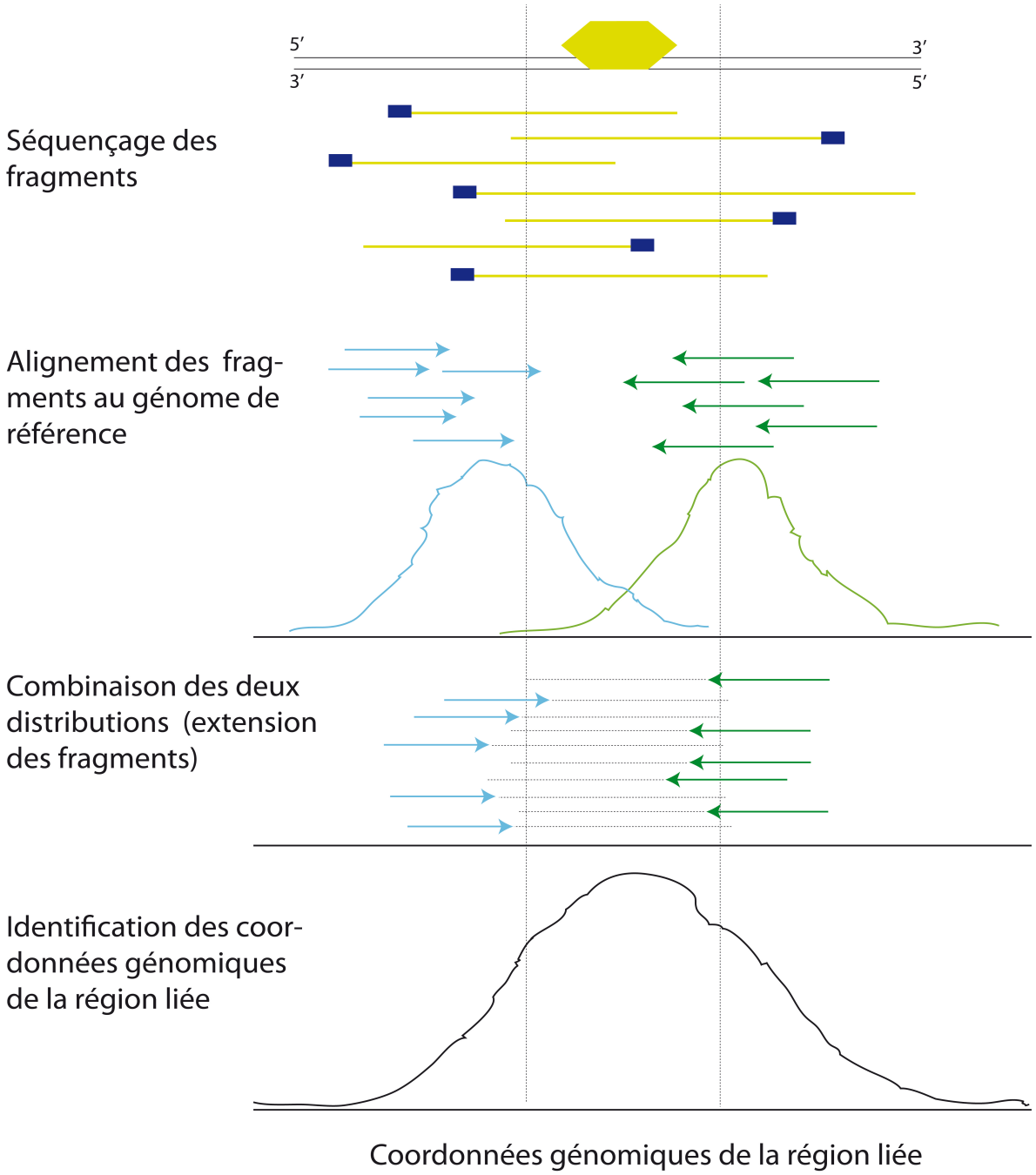


Figure 15. Procédure d'identification de pic à partir des données obtenues par ChIP-seq.

Les fragments sont d'abord alignés au génome de référence, ensuite une distribution est estimée en utilisant une loi de Poisson ou une loi Binomiale pour trouver la région du pic.

Il est important de noter que la distribution des régions liées présente différents profiles de pics : localisés étroits, localisés mixtes et enfin des pics larges distribués sur de grandes distances (Figure 16). Les pics étroits localisés, sont généralement associé aux régions liées par les FTs. Les pics mixtes quand à eux, ont été observés dans le cas de certaines protéines comme l'ARN polymérase II [103] et quelques marques d'histones qui caractérisent des régions de chromatine active comme la tri-méthylation de la lysine 4 de l'histone H3 (H3K4me3). Les pics larges sont associés aux modifications d'histones dans les régions de chromatine inactive, à l'instar de la tri-méthylation de la lysine 27 de l'histone H3 (H3K27me3). Cette dernière peut s'étaler sur de larges régions génomiques [94, 95].

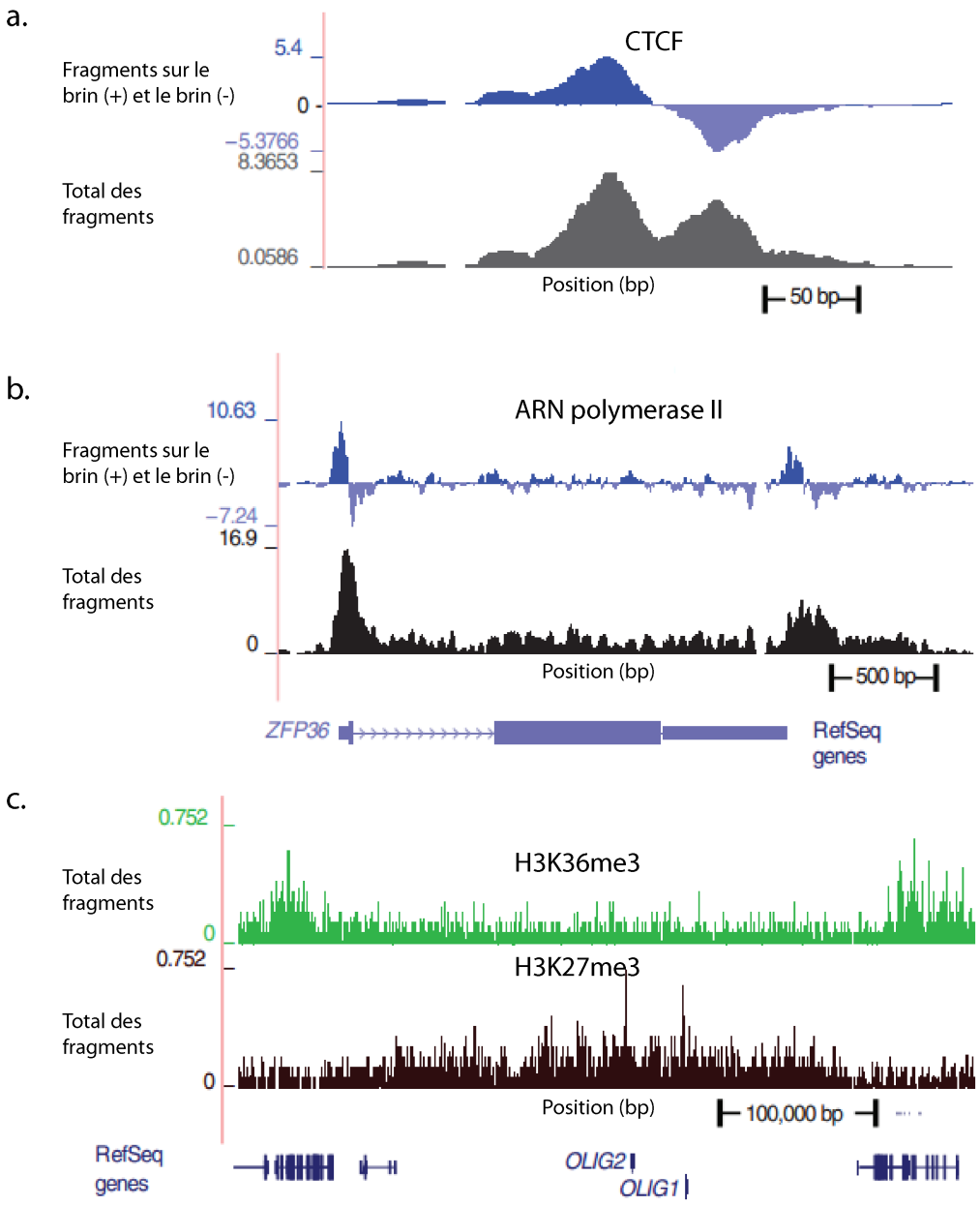


Figure 16. Figure présentant les différents profiles de pics observés.

(a) pics localisés associés en général au FTs (b) pics mixtes associés au profile de Polymérase II et quelques modifications d'histones actives (c) pics larges associés à des marques d'histones caractérisant les régions de chromatine inactive.

Après avoir identifié les régions enrichies, une étape importante consiste à valider une partie des régions identifiées afin de tester la qualité des prédictions. Cette validation se fait généralement par PCR quantitative (qPCR).

Les fragments ainsi obtenues sont enrichies non seulement en sites de fixation du facteur de transcription immuno-précipité (interaction directe avec l'ADN) mais aussi en sites de fixation d'autres facteurs de transcription qui médient sa liaison à l'ADN (interaction indirecte ou coopérative). Des outils de découverte et de recherche de motifs sont ensuite utilisés pour prédire les sites liés par les FTs dans ces régions.

Table 1. Exemple d'outils d'identification de pics

Non de l'outil	Site web
PeakFinder 2.0.1	http://woldlab.caltech.edu/html/chipseq_peak_finder
GeneTrack 1.0.1	http://code.google.com/p/genetrack/
FindPeaks 3.1.9.2	http://www.bcgsc.ca/platform/bioinfo/software/findpeaks/
SISSRs v1.4	http://sissrs.rajajothi.com/
QuEST 1.0	http://mendel.stanford.edu/sidowlab/downloads/quest/
MACS 1.3	http://liulab.dfci.harvard.edu/MACS/
CisGenome v1	http://www.biostat.jhsph.edu/~hji/cisgenome/
PeakSeq v1.01	http://www.gersteinlab.org/proj/PeakSeq/
Hpeak 1.1	http://www.sph.umich.edu/csg/qin/HPeak/

3.3 Recherche des sites d'ADN liés par le FT immuno-précipité et les sites liés par ses partenaires

Plusieurs analyses peuvent être effectuées sur les régions liées. Dans le cas où la séquence d'ADN liée par le FT d'intérêt est connue (répertoriée dans les bases de données de facteurs de transcription), l'analyse consiste à scanner les régions obtenues par ChIP-chip

ou ChIP-seq en utilisant le motif lié par ce FT afin d'identifier toutes ses occurrences. La même procédure peut être utilisée pour vérifier si des sites liés par d'autres FTs sont enrichis dans ces régions dans le but d'identifier d'éventuels partenaires du FT immuno-précipité. Afin de pouvoir mesurer l'enrichissement de ces sites dans les régions liées, divers types de séquences contrôles peuvent être utilisées. La significativité statistique de cet enrichissement est évaluée en utilisant différentes fonctions de score [104, 105] . Il est important d'insister ici sur le choix de l'ensemble de contrôle à utiliser car c'est lui qui détermine la valeur de l'enrichissement observé. En effet, plusieurs biais peuvent être introduits soit dans les régions liées ou bien dans l'ensemble contrôle. Parmi ces biais en peut citer : le contenu en GC et la présence des régions répétées. Par exemple, il n'est pas approprié d'utiliser un ensemble contrôle riche en AT pour identifier des sites qui sont eux même riches en AT. De plus, un ensemble contrôle riche en AT va être biaisé pour les motifs GC riches [104, 106].

3.4 Annotation des régions liées

Une des questions importante associée à la prédiction des sites liés par les FTs concerne la fonctionnalité de ces sites. En effet, la présence d'un site similaire au profile lié par un FT donné ne détermine pas sa fonctionnalité. Cependant, les régions liées par les FTs sont souvent annotées par rapport aux gènes afin d'établir une relation entre l'événement de liaison à l'ADN et la régulation des gènes cibles. Par exemple, si une corrélation est observée entre la liaison à l'ADN d'un FT donné et l'expression d'un ou de plusieurs gènes, ceci pourrait suggérer une régulation de ces gènes par ce FT. Aussi, ces données peuvent être utilisées pour identifier d'éventuels complexes de FTs à partir de l'information sur leurs sites de fixation à l'ADN. Plusieurs outils ont été développés pour répondre à ces questions, parmi lesquels on peut citer UCSC genome browser [107] , Ensembl Biomart [108],ChIPpeakAnno[109] et SeqMonk (http://www.bioinformatics.bbsrc.ac.uk/projects/seqmonk/).

3.5 Effet des séquences répétées

Plusieurs études ont montré la présence des éléments cis-régulateurs dans les régions répétées [110-112]. Bourque et collaborateurs ont analysé le profile de liaison à l'ADN de plusieurs FTs et ont montré que des fractions importantes des régions liées par ces FTs sont localisées dans différentes familles de régions répétées [113]. Ils ont montré que 19% des régions d'ADN liées par le récepteur des estrogène ERα, sont localisées dans des régions MIR et 33% des régions liées par CTCF sont localisées dans des régions répétées de type B2. Cependant les techniques de ChIP-chip et de ChIP-seq présentent des limitations quand à l'analyse des régions répétées du génome. En effet, dans le cas d'une expérience de ChIP-chip les régions répétées sont exclus de la puce (aucune probe de la puce n'est située dans une région répétée). Par conséquent, les régions répétées sont masquées lors des analyses subséquentes ce qui pourrait sous estimer le profile de liaison de la protéine d'intérêt à l'échelle du génome. La technique de ChIP-Seq quand à elle utilise le génome tout entier pour identifier les fragments séquencés. Cependant, les fragments alignés à plus d'une région du génome sont éliminés. Par conséquent, certains sites localisés dans des régions répétées seront perdus. Les logiciels qui autorisent l'alignement à plus d'une région génèrent un taux plus élevé de faux positifs. Le choix d'un algorithme d'alignement dépend donc des objectifs de l'analyse, à savoir augmenter la spécificité (fragments alignés à des régions uniques) ou bien avoir une bonne sensibilité (accepter les fragments localisés dans les régions répétées) [90].

3.6 Modélisation des sites de fixation à l'ADN des FTs

Afin de pouvoir exploiter les données sur les sites de fixation des facteurs de transcription, il est important en premier lieu de définir un modèle pour décrire et modéliser l'ensemble des sites liés par un FT donné. La suite des analyses dépend en plus grande partie du modèle choisi. Dans la littérature, deux modèles ont été largement utilisées : la représentation par consensus et la représentation par un profile statistique basé sur

l'alignement des séquences liées et appelé communément matrice de fréquence (PFM) ou matrice de poids de positions (PWM) [16].

3.6.1 Représentation par consensus

La représentation par consensus est utilisée pour représenter les propriétés statiques de certains sites comme ceux des enzymes de restriction [114]. L'approche consiste à aligner l'ensemble des séquences liées ensuite de définir le nucléotide le plus fréquent à chacune des positions du site. Cependant, nous avons besoin de définir comment les autres sites diffèrent du site consensus afin de décider s'ils peuvent être considérés comme des sites de liaison valide pour le FT d'intérêt. De plus, certaines positions n'ont pas une préférence pour un nucléotide donné, et acceptent plus d'un nucléotide. Afin de représenter cette préférence la solution a été d'utiliser le code IUPAC (Table 1) pour modéliser l'ensemble des sites liés par le FT. Cette représentation, bien que simple à utiliser, présente le désavantage de ne pas tenir compte des préférences en nucléotides à chacune des positions. Par conséquent, ce modèle n'est pas approprié pour représenter la variabilité des sites fixés par les FTs.

Table 2. Code IUPAC

IUPAC	Nucléotides	Mnémoniques
A		Adénine
C		Cytosine
G		Guanine
T		Thymine
R	A or G	puRines
Y	C or T	pYrimidines
W	A or T	Liaison hydrogène faible
S	G or C S	Liaison hydrogène forte
M	A or C	aMino groupe
K	G or T	Keto groupe

H	A,C,T	Différent de G
B	C,G,T	Différent de A
V	A,C,G	Différent de T
D	A,G,T	Différent de C
N	A,C,G,T	N'importe quels nucléotides

3.6.2 Représentation en utilisant un profile statistique

Une représentation plus flexible de l'ensemble des sites liés par un FT est possible grâce à la modélisation par matrice. Cette dernière est obtenue en alignant (sans espacements) l'ensemble des sites liés par un FT donné, ensuite mesurer la fréquence des quatre nucléotides à chacune des positions du site. Il en résulte une matrice de dimension 4 x m, m étant la longueur du site lié. La somme de chaque colonne étant égale au nombre de sites alignés [16, 115, 116].

L'ensemble des sites liés par un FT sont compilés dans un modèle quantitatif appelé matrice de fréquence (PFM). Cette dernière est construite par comptage du nombre d'occurrences de chaque nucléotide à chacune des positions de la séquence. Cette matrice de comptage est ensuite convertie en une matrice de probabilités (PWM) en introduisant un nouveau paramètre qui est la fréquence de chaque nucléotide dans le génome de référence (Figure 17). La matrice de probabilité peut être visualisée graphiquement en utilisant la représentation par logo [117].

L'évaluation d'une nouvelle séquence par rapport à la matrice se fait par sommation des valeurs de probabilités correspondant à chacune des positions du site [37] et le score obtenu reflète l'énergie de l'interaction entre le FT et cette nouvelle séquence. Afin de distinguer les sites potentiels (forte énergie de liaison) et les sites de faible énergie (sites de faibles affinités), il est important de fixer un seuil à partir duquel on peut considérer que le site est dérivé à partir du modèle (de la matrice) ou bien du bruit. Il existe différentes approches

pour le choix du seuil. Ce dernier peut être soit un score statique ou bien sous forme d'une fraction ou d'une probabilité [37]. Dans le cas où l'on est intéressé par représenter l'affinité de la séquence, le choix d'un score statique est plus approprié. Par contre le seuil basé sur la probabilité de liaison, représente la probabilité d'observer un certain score.

La représentation par matrice présente une limitation, à savoir l'assomption de l'indépendance entre les différentes positions de la séquence (i.e la distribution d'un nucléotide à une position donnée n'affecte pas la distribution à une autre position). Cette assomption est valide pour un certain nombre d'exemples [37] mais ne peut être généralisée pour l'ensemble des FTs.

À partir d'expérience *in vivo* et *in vitro*, des sites liés par les FTs sont compilés dans des PWMs et répertoriés dans des banques de matrices comme TRANSFAC [49] et JASPAR [50].

Étape 1: Collection des sites liés par un TF, tous les sites devraient avoir la même longueur

Étape 2: Compatge de la fréquence de chaque nucléotide A,C,G,T à chacune des positions du motif. On obtient une matrice de fréquence

Étape 3: La matrice de fréquence est convertit en une matrice de poids en utilisant les formules suivantes:

$$p(b,i) = \frac{f_{b,i} + s(b)}{N}$$

s(b): est un pseudo compte
N : nombre de sites

$$W_{b,i} = \log_2 \frac{p(b,i)}{p(b)}$$

p(b): probabilité génomique de la base b

Étape 4: Score d'un nouveau site en utlisant la matrice de poids

Étape 5: Representation graphique de la matrice

a.

Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Ensemble de sites liés par un TF

b. Matrice de fréquence (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

c. Matrice de poids de positions (PWM)

A	-1.93	0.79	0.79	-1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	-1.93	-1.9	0.79
C	0.45	-1.93	0.79	1.68	-1.93	-1.93	-1.93	0.45	-1.93	-1.93	-1.93	-1.93	0.00	0.79
G	0.00	0.45	-1.93	-1.93	-1.93	-1.93	-1.93	-1.93	0.66	-1.93	1.30	1.68	1.07	-1.93
T	0.15	0.66	-1.93	-1.93	1.07	0.66	0.79	0.00	0.00	0.79	-1.93	-1.93	-0.66	-1.93

d. Score d'une nouvelle séquence $\Sigma = 8.53$ (59% du score maximum)

0.45	0.66	0.79	1.68	0.45	0.66	0.79	0.45	0	0.79	0.00	1.68	-0.66	0.79
C	T	A	C	A	T	T	A	T	A	A	G	T	C

e. Représentation graphique du motif

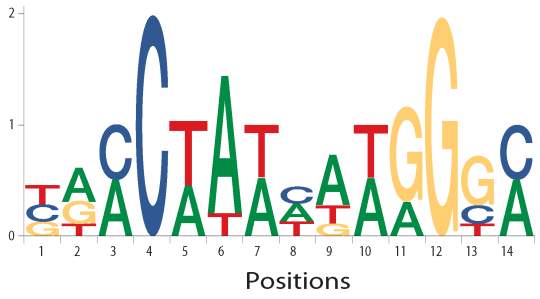


Figure 17. Modélisation par matrice des sites liés par un FT.

La figure est adaptée de [37].

3.6.3 Représentation par des modèles de Markov cachés

Badis et collaborateurs [83] ont montré par des analyses *in vitro*, l'importance de l'interdépendance des positions dans les sites liés par les FTs. Ils ont aussi montré que cette dépendance, ne fait pas intervenir uniquement les positions adjacentes du site lié mais peut survenir aussi entre les positions séparées de plus d'un nucléotide.

Une autre limitation de la représentation par matrice, est l'impossibilité de représenter l'information sur les délétions et les insertions. En effet, l'espacement entre deux demi sites, dans le cas d'un FT liant un site avec espacement, ne peut être incorporé au model. Dans ce cas un modèle plus flexible est nécessaire. Les modèles de Markov cachés (HMM) [20, 106] ont été proposé pour palier à cette limitation. Les HMMs peuvent être perçu comme une extension de la représentation par matrices, qui permettent une description plus détaillée des caractéristiques de la séquence ainsi que la modélisation des interactions, utilisant différents ordres, entre les positions de la séquence. Par définition, un HMM consiste en une série d'états, où chaque état peut décrire l'état suivant basé sur une probabilité d'émission. Les états sont connectés entre eux sous forme d'une chaine ayant un état initial et un état final. La probabilité de passage d'un état à un autre est appelée probabilité de transition. Plusieurs chemins (donc séquences) peuvent être créés en suivant des passages différents dans la chaine avec différentes probabilités.

3.7 Problèmes liés à la prédiction des sites de fixations de facteurs de transcription à l'ADN en utilisant les modèles de TFBS répertoriés dans les bases de données de facteurs de transcription

Les modèles décrits ci-haut sont utilisés pour prédire de nouveaux sites qui pourraient être lié par le FT pour lequel le modèle a été créé. Cependant, deux observations très importantes sont à considérer :

a) Un modèle pour un FT donné, génère en moyenne une prédiction pour chaque 500-1500 bp [118]. Le nombre élevé de prédictions est biologiquement irréaliste. Par conséquent une fraction importante des motifs prédits en utilisant ces modèles représente des faux positifs.

b) Certaines de ces prédictions reflètent l'affinité *in vitro* de la liaison du FT à l'ADN, mais pas nécessairement son affinité *in vivo* [119].

Ces observations démontrent que ces modèles, bien que très utiles, manquent d'informations importantes et nécessaires pour distinguer les sites fonctionnels (sites liés *in vivo* par les FTs) des sites non fonctionnels. Ce qui nous amène à nous poser la question suivante : où trouver l'information manquante pour créer une représentation proche du profile lié par le FT *in vivo* ?

Nos connaissances actuelles des mécanismes de liaison à l'ADN des FTs sont loin d'être complètes. Plusieurs aspects pouvant affectés cette interaction ne sont pas incorporés dans les modèles actuels, à savoir :

3.7.1 Accessibilité des régions génomiques analysées :

Quand une région est scannée en utilisant un de ces modèles, on assume que cette région est accessible au FT en tout temps. Cette assomption est fausse car l'ADN est organisé en une structure chromatinienne dynamique et complexe [4, 83, 97] et les régions du génome ne sont pas toutes accessibles au même temps et sous les mêmes conditions.

3.7.2 Organisation des FTs (modularité)

En général les FTs n'agissent pas seuls. Ils s'organisent en modules pour réguler leurs gènes cibles [83, 106, 120-122]. Le rôle de cette modularité est soit de stabiliser la liaison du FT à l'ADN ou bien d'augmenter son affinité à celui-ci. Il est donc important d'incorporer l'information sur l'organisation modulaire des FTs dans la modélisation de leur profile de liaison à L'ADN. L'information sur les motifs composites, liés par deux FTs ou plus est inexistantes, à l'exception de quelques complexes pour lesquels l'information est répertoriée dans les banques de FTs.

3.7.3 La dégénérescence des sites liés par les FTs

Les FT peuvent présenter une affinité pour des séquences d'ADN qui dévient de quelques positions de leurs sites consensus. Cette variabilité dans le profile du motif lié rend difficile leur identification car on ne dispose pas de modèles prédictifs qui pourraient nous renseigner sur le degré de dégénérescence à autoriser. De plus, la variabilité observées dans les sites d'ADN liés diffère d'un FT à un autre [37].

Chapitre 4

4. Statistiques pour évaluer l'exceptionnalité de la distribution des sites fixés par les facteurs de transcriptions

La compréhension des propriétés statistiques des sites d'ADN liés par les FTs permettrait de mieux comprendre leurs contraintes fonctionnelles afin de développer des approches computationnelles efficaces pour leur prédiction. En effet, il n'y a pas de consensus quant à la localisation des sites liés par les FTs (promoteur proximal ou distal) ni à la variabilité observée dans leurs séquences. De plus, l'incorporation des propriétés biologiques des FTs ainsi que celles de leurs sites de fixation à l'ADN dans le modèle statistique permettrait d'améliorer significativement la qualité de ces approches.

Avant d'implémenter un algorithme de découverte de motifs, il est important de bien définir deux paramètres. Le premier est l'ensemble de référence, appelé aussi contrôle négatif, contre lequel l'exceptionnalité ou la surreprésentation des motifs prédits seront comparées. Le but étant de pouvoir distinguer les motifs qui présentent un potentiel pour être lier par les FTs dans un jeu de données test et ceux prédits pour cause de biais de l'approche expérimentale et/ou computationnelle utilisée. Le deuxième paramètre représente les fonctions de scores utilisées pour évaluer l'exceptionnalité des motifs prédits dans le jeu de données test comparés à ce qui est observé dans le jeu de données contrôle.

En général, on considère que les motifs observés dans le jeu de données test devraient avoir des propriétés statistiques spécifiques à ce jeu de données et ne pas présenter ces mêmes propriétés dans le jeu de données contrôle. Le but étant donc de modéliser les

propriétés des motifs prédits par une fonction de score qui a le pouvoir de différencier un vrai motif (site lié par un FT) d'un faux motif (bruit).

4.1 Choix de l'ensemble de référence

Les variations dans la composition en nucléotides des jeux de données analysés affectent considérablement les résultats des outils de découverte de motifs d'ADN liés par les FTs. En effet souvent les motifs liés par les FTs sont soit riches en di-nucléotides CG ou bien en di-nucléotides AT [16]. De plus, il existe dans le génome des régions de faible complexité comme les régions Poly Adénylées (Poly A) et les ilots CpG. Par conséquent, il est important de choisir un ensemble de référence qui ne présente pas un biais de composition pour ces séquences. Pour qu'un motif soit considéré comme un candidat potentiel, il faut montrer que sa distribution dans le jeu de données test est différente de celle attendu dans un jeu de données contrôle [123]. Souvent on utilise la propriété de surreprésentation comme mesure pour identifier les motifs enrichis. Le niveau de cette surreprésentation dépend donc directement de l'ensemble de référence utilisé. En effet si par exemple on cherche à détecter la surreprésentation d'un motif riche en GC, les ensembles de références riches en GC peuvent masquer des enrichissements potentiels. Par contre, ce biais de composition va favoriser l'enrichissement apparent des motifs pauvres en GC ou les motifs riches en AT [124, 125]. Il est donc indispensable de choisir un ensemble de référence qui permet d'observer des enrichissements réels et d'éliminer ceux qui sont dus à des biais de composition. Par exemple, pour prédire les motifs enrichis dans les promoteurs d'un groupe de gènes co-régulés, on pourrait utiliser tous les promoteurs des gènes du génome comme contrôle[104]. En général, trois types d'ensembles contrôles sont utilisés :

(a) Ensemble contrôle créé en utilisant un modèle de Markov pour un certain ordre

(b) Ensemble contrôle échantillonné aléatoirement à partir du génome de référence.

(c) Ensemble contrôle créé à partir du jeu de données à analyser en randomisant aléatoirement les positions des nucléotides.

4.2 Choix de la fonction de score

Une des propriétés clés utilisée pour évaluer l'exceptionnalité des motifs prédits est la surreprésentation. En effet, supposons qu'on dispose d'un ensemble de séquences représentant par exemple un groupe de promoteurs de gènes co-régulés. Nous avons donc identifié dans ce groupe de promoteurs une liste de sites dont la séquence est similaire. L'idée est de montrer que cette même liste de séquences, qu'on pourrait aligner ensemble pour former un motif, ne présente pas les mêmes caractéristiques dans un jeu de données sélectionné aléatoirement. Dans ce cas, on pourrait émettre l'hypothèse que la présence de ce motif dans le jeu de données analysé, est peut être associé à un rôle biologique (comme par exemple être liés par un FT) et par conséquent pouvoir expliquer la co-régulation des gènes dont les promoteurs contiennent une occurrence de ce motif.

4.2.1 Le score Z (Z-score)

La première assomption est de considérer que les positions dans chacune des séquences analysées sont indépendantes les unes des autres. La probabilité d'observer un nucléotide $P_{i\ [\ A,C,G,T\ ]}$ à la position j d'une séquence S dépend uniquement de sa distribution dans l'occurrence dans laquelle il apparait indépendamment des nucléotides adjacents. Par conséquent, la probabilité d'observer une occurrence $p=p_1p_2p_3.....p_m$ à chacune des positions de la séquence analysée, est égale au produit des probabilités d'observer chacun des nucléotides qui la compose et elle est donnée par :

$$\Pr(p) = \prod_{i=1}^{m} Pr\ (p_i)$$

Où $Pr(p_i)$ représente la probabilité d'observer le nucléotide p_i dans la séquence et m est la longueur du motif. Une mesure d'exceptionnalité des motifs prédits et qui pourrait découler de cette probabilité est le Z-score [125, 126], définit par l'équation suivante :

$$Z(p) = \frac{\boldsymbol{Obs(p) - Exp(p)}}{\sqrt{\boldsymbol{Var(p)}}}$$

Où Obs(p) représente la fréquence du motif p dans les séquences analysées, Exp(p) and Var(p) représentent la fréquence attendue du motif p et sa variance dans l'ensemble contrôle. La valeur de la variance est souvent estimée à partir de Pr(p). Le Z-score permet de comparer la fréquence d'un motif entre deux jeux de données, mais ne fourni aucune information sur sa composition et le degré de conservation des nucléotides à chacune des positions.

4.2.2 Le contenu en information d'un motif (IC pour information content)

Une autre mesure, le contenu d'information du motif (IC pour Information content) [127] , permet de représenter cette information. Pour un motif avec un profile M, son IC est définit par :

$$\text{IC(M)} = \sum_{i=1}^{4} \sum_{j=1}^{m} m_{ij}\, log\, \frac{m_{ij}}{b_i}$$

m_{ij} étant la fréquence du nucléotide i à la position j et b_i est la fréquence génomique du nucléotide i.

À partir de la mesure IC on peut estimer le degré de conservation de chacune des positions du motif. Sa valeur maximale est atteinte lorsqu'un seul nucléotide apparait à chacune des colonnes. Alors que la valeur du IC est nulle, quand la fréquence des nucléotides est similaire à celle observée dans l'ensemble contrôle utilisé ou bien si tous les nucléotides ont la même fréquence d'apparition à cette position.

4.2.3 Le score MAP (maximum of a posteriori)

Une autre mesure dérivée du IC est le maximum de vraisemblance MAP (maximum of a posteriori) [128, 129]. La valeur MAP d'un profile *M* est définit par :

$$MAP = \frac{log(x_m)}{w} \left\{ \sum_{i=1}^{w} \sum_{j=1}^{4} p_{ij}\, log\ p_{ij} - \frac{1}{x_m} \sum_{alls} log(p_0(s)) \right\}$$

w étant la longueur du motif, x_m le nombre d'occurrences de la matrice *M*, p_{ij} est la probabilité du nucléotide *j* à la position *i* de la matrice *M* et $p_0(s)$ représente la probabilité de générer le modèle à partir de l'ensemble contrôle utilisé. La première partie de l'équation représente l'entropie du motif *M*. Plus les sites composant le motif sont proches l'un de l'autre, plus cette valeur est élevée. Pour un vrai TFBS, la valeur de son score MAP est d'autant plus grande que les sites qui le composent sont similaires.

4.2.4 Le score LR (likelihood ratio)

Un autre moyen d'évaluer l'exceptionnalité d'un motif est d'estimer pour chacune de ces occurrences, la probabilité qu'elle soit générée soit à partir du motif ou bien d'une distribution aléatoire. Cette probabilité, appelée le score de vrai semblance (likelihood ratio) [130, 131], est calculée par :

$$LR(M, W, B) = \sum_{k=1}^{n} log \frac{P(w_k\ /M)}{P(w_k/B)}$$

Avec *M* et *B* qui représentent la matrice et le contrôle respectivement et *n* la longueur du motif. $P\ (w_k/M)$ et $P\ (w_k/B)$ sont les probabilités de générer l'occurrence w_k à partir de la matrice et du contrôle respectivement. Alors que le IC dépend uniquement de la matrice, le LLR quand à lui dépend des sites qui sont utilisés pour construire la matrice.

4.2.5 Autres fonctions de score

Un autre type de fonctions de score teste la distribution spatiale des motifs prédits ou bien leur fréquence dans le jeu de données test comparé à ce qui est attendu aléatoirement. La première fonction détermine la spécificité de la distribution du motif dans le jeu de données analysé (Groupe specificity score). Ce score évalue si le type de distribution du motif prédit ainsi que sa fréquence sont spécifiques aux séquences analysées. Par exemple, dans le cas d'un FT qui régule un groupe de gènes, on s'attendrait à ce que le motif d'ADN lié par ce FT, soit plus enrichit dans les promoteurs de ces gènes comparés à tout les gènes du génome [106, 132]. La deuxième fonction de score a pour but de tester si la distribution des motifs prédits présente un biais positionnel dans le jeu de données test, qui n'est pas observé dans le jeu de données contrôle. Par exemple, dans le cas des données de ChIP-chip et de ChIP-séquençage, les sites liés sont distribués à proximité du centre des fragments immuno-précipités [133]. Par conséquent, on s'attendrait à observer un enrichissement significatif des motifs liés dans ces régions comparé aux séquences utilisées comme contrôles.

Chapitre 5

5. Approches computationnelles pour l'identification et la découverte des éléments cis-régulateurs

La compréhension des mécanismes complexes qui gouvernent les processus biologiques dans la cellule, requière une caractérisation des éléments cis-régulateurs qui modulent l'expression des gènes au niveau transcriptionnel et post-transcriptionnel. La caractérisation de ces éléments nécessite des approches efficaces tant expérimentales que computationnelles. Les premières approches computationnelles consistaient à analyser les régions de promoteurs de gènes ayant certaines propriétés en communs (profile d'expression, fonction biologique, localisation cellulaire etc..) dans le but d'identifier les motifs (signaux) qui sont enrichis dans ces régions. Ces motifs pourraient en effet correspondre à des sites liés par les FTs qui régulent ce groupe de gènes. Cependant, les sites liés par un FT sont généralement courts (6-20 bp) et dont la séquence dévie du consensus (séquences ayant des variabilités à certaines positions). Ceci rend difficile leur identification d'autant plus que les régions de promoteurs où sont recherchés ces sites sont de taille variable et en générale de l'ordre de 1k pour les organismes de petit génome à quelques milliers de paires de bases pour les organismes supérieurs [37].

Plusieurs approches computationnelles ont été développées pour l'identification et la découverte des éléments cis-régulateurs. Elles sont classées comme suit :

5.1 Approche pour la recherche et la découverte des sites liés par les FTs

5.1.1 Approches supervisées

Cette catégorie d'approches utilise les modèles de TFBS répertoriés dans les banques de FTs [49, 50, 134]. Leur principe consiste à scanner une région d'ADN en utilisant des librairies de TFBS précompilés afin d'identifier ceux qui sont significativement enrichis dans cette séquence comparé à une séquence contrôle. Cependant, ces approches sont confrontées à certaines contraintes à savoir :

(a) Peu de TFBSs sont validés *in vivo.*

(b) Les modèles de TFBSs connus sont construits dans la plupart des cas à partir d'un nombre limité de sites validés. Par conséquent, ils ne représentent pas toute la variabilité observée dans un motif lié par un FT.

(c) Ces TFBSs ont été identifiés dans des conditions expérimentales spécifiques (conditions, lignée cellulaire). Par conséquent, on ignore si le FT peut lier des variantes du modèle connu ou bien si sous d'autres conditions, ce même FT lierait un motif complètement différent.

5.1.2 Approches Non-supervisés

Ces approches, appelés communément approches de découvertes de motif d'ADN, ont l'avantage de ne pas se limiter aux modèles de TFBSs existants. En effet, leur principe de base repose sur l'idée de prédire tous les motifs ayant certaines caractéristiques statistiques dans le jeu de données analysé comparé à ce qui est attendu dans des jeux de données contrôles. Les tables 3 et 4 représentent quelques outils de découvertes motifs, classés selon l'approche utilisée.

5.1.2.1 Définition du problème

Étant donné un groupe de n séquences de longueur l, et un motif de longueur k. il existe $(l-k+1)^n$ combinaisons de sites à évaluer. D'un point de vu algorithmique, ce problème est

NP-complet [106]. Par conséquent, il n'existe pas de solution à ce problème dans des temps de calcul polynomiales. Pour cette raison, la solution a été de développer des heuristiques, qui au lieu d'énumérer toutes les possibilités existantes, adoptent des stratégies pour échantillonner seulement une sous fraction des données analysées. Ainsi, les solutions qui seront trouvées ne sont pas forcément des solutions optimales au problème, mais elles représentent les meilleures solutions selon la stratégie adoptée [121]. Pour une longueur de motif connue, le but de ces stratégies est de prédire à partir du jeu de données analysé, l'ensemble des groupes de séquences similaires, où chaque groupe de séquences représente un profile de motif pouvant correspondre à un site lié par un FT. Elles différent principalement en :

(a) La méthode utilisée pour sélectionner (extraire) les sous séquences formant le motif
(b) Le modèle statistique utilisé pour représenter le motif (consensus, matrice, HMM).
(c) Les fonctions de scores et l'ensemble de référence (contrôle) permettant d'évaluer l'exceptionnalité ou la surreprésentation du motif.

Ces heuristiques peuvent être classées en deux catégories principales :

5.1.2.2 Approches basées sur une représentation par consensus

Le motif consensus d'un site lié par un FT, représente le site de plus haute affinité. Cependant ce même FT, peut lier des formes qui dévient du consensus de quelques positions. En général, le nombre de positions variables autorisées dépend de la taille du site lié par le FT. Les outils de découverte de motifs basés sur cette approche [20, 135-137] procèdent comme suit :

a. Définir la longueur *k* des motifs à prédire
b. Énumérer toutes les instances de longueur *k* à partir des données analysées.
c. Calculer la fréquence de chacune des instances trouvées en (b) en autorisant un nombre *e* de substitutions entre les instances similaires.

d. Définir un seuil *s* pour filtrer les motifs qui apparaissent dans un nombre non significatif de séquences (par exemple garder les motifs présents dans au moins 50% des séquences).
e. Ordonner les motifs selon leur fréquence ou bien d'autres mesures statistiques.

Une limitation majeure de ces approches est le temps de calculs nécessaires pour traiter tous les motifs d'une certaine longueur. Une solution à ce problème a été d'utiliser les arbres de suffixes, une structure de données beaucoup plus flexible et rapide pour indexer le jeu de données à analyser. Néanmoins, cette approche est efficace si l'on considère seulement les motifs courts et sans substitutions. SMILE [138], WEEDER [139],RSAT [140] et YMF [141] sont des exemples d'outils appartenant à cette catégorie d'approches.

5.1.2.3 Approches basés sur une représentation par un modèle statistique

Le manque de flexibilité dans la représentation par consensus a motivé le développement d'une nouvelle génération d'approches basées sur une représentation quantitative de l'information contenu dans le motif, appelée représentation par matrice de poids ou matrice de fréquence. L'ensemble des séquences de même longueur, liées par un FT donné, sont compilées dans une matrice de fréquence. Une séquence représentant la « séquence consensus » est définie en associant à chaque colonne le nucléotide ayant la meilleure fréquence (le meilleur score). Tous les sites générés à partir de cette matrice et ayant un certain degré de similarité (pourcentage de similarité) avec la séquence de score maximal, sont considérés comme des cibles potentielles pour être liés par le FT pour lequel la matrice a été créée. Deux catégories principales d'outils de découverte de motif ont implémenté ce mode de représentation :

5.1.2.3.1 Les approches énumératives

Les approches appartenant à cette catégorie [142-145] sont basées sur une énumération exhaustive de tous les segments d'ADN pour une longueur spécifique, à partir d'un ensemble de séquences de départ (promoteurs de gènes, fragments de ChIP, etc..). Un critère de similarité est ensuite utilisé pour regrouper les segments d'ADN similaires. Ces derniers sont compilés dans des matrices de fréquence qui représentent les profiles des motifs créés. Les motifs sont en général classés selon leur contenu d'information (IC).

Partant de l'hypothèse que le signal biologique (TFBS) devrait être plus enrichit dans l'ensemble test comparé au contrôle, l'évaluation statistique des motifs ainsi prédits, est réalisée en mesurant leur niveau de surreprésentation dans l'ensemble des séquences où ils ont été trouvées comparé à ce qui est attendu aléatoirement ou dans un ensemble contrôle défini par l'utilisateur. Une étape très importante lors de l'utilisation de ces approches est donc de définir à l'avance l'ensemble contrôle (négatif) à utiliser ainsi que la ou les statistique(s) adéquate (s) pour évaluer la surreprésentation des motifs prédits.

Ces approches bien que très efficaces grâce à leur stratégie de recherche exhaustive, elles sont cependant moins utilisées à cause des ressources exigées en temps de calcul. Pour échapper à cette contrainte, certains algorithmes énumératifs, à l'instar de MDscan [128] réduisent le nombre de séquences dans lesquelles les motifs sont recherchés en sélectionnant un sous ensemble de séquences de départ. Les séquences restantes sont ensuite utilisées pour raffiner les modèles initiaux. Ces derniers ont été créés par groupement de sous séquences similaires de même longueur. En général, la similarité est définie comme étant la distance entre chaque paire de sous séquences de même longueur (nombre de positions contenant un nucléotide différent). Ce paramètre, qui dépend en général de la longueur du motif recherché, est soit prédéfini dans l'outil lui-même ou bien fixé par l'utilisateur. Il est important de noter ici que la longueur des motifs à prédire contribue directement à augmenter la complexité de ces algorithmes. En effet, plus le motif est long, plus l'algorithme nécessite du temps pour évaluer la similarité entre toutes les

paires de sous séquences de même taille. Par conséquent, ces algorithmes se limitent généralement à la prédiction de motifs de longueur courte ou moyenne entre 6 et 10 bp [37].

5.1.2.3.2 Les approches probabilistes

Un autre type d'heuristiques a été développé pour résoudre le problème de la découverte de motifs [16]. Au lieu de procéder par une recherche exhaustive comme c'est le cas des approches énumératives, Les algorithmes dits probabilistes, procèdent en trois étapes :

1. Echantillonnage aléatoire d'un ensemble de positions de départ à partir d'un sous groupe de séquences parmi l'ensemble de séquences à analyser. La plupart des algorithmes appartenant à cette catégorie sont flexibles quand au nombre de positions à échantillonner par séquence (zéro, une ou plusieurs positions).
2. Création d'un profile du motif (matrice) en utilisant les sites trouvés aux positions échantillonnées dans (a). Une longueur de motif devrait être spécifiée.
3. Le profile créé en (b) est ensuite raffiné jusqu'à ce que le modèle converge vers une solution optimale en utilisant l'un des deux algorithmes suivants : Expectation et Maximisation (EM) et Gibbs sampling. Ces deux algorithmes ont été largement utilisés pour développer plusieurs approches de découvertes de motifs d'ADN. Dans la section suivante, nous allons décrire un peu plus en détail ces deux algorithmes :

(a) Expectation et Maximisation EM [146, 147]

Soit un profile de motif *M* lié par un FT. *M* est la matrice créé à partir de l'ensemble des occurrences liés par ce FT, trouvées sur un sous ensemble de séquences $S_1 \ldots\ldots S_m$. Le principe de l'algorithme EM est de trouver l'occurrence qui maximise la probabilité du profile sur chacune des séquences S_i $i=1..$ m. L'algorithme estime la probabilité $p_{ij,}$ de trouver une occurrence du profile à chaque position j de la séquence S_i. L'occurrence qui maximise la probabilité du profile est sélectionnée, ensuite un nouvel alignement est créé. Ces deux étapes sont répétées pour chacune des séquences S_i $i=1..$ m. Afin d'assurer la convergence du profile vers la solution optimale, le processus EM est répété jusqu'à ce que

la probabilité du profile atteigne sa valeur maximale et plus aucune amélioration n'est possible. Le but étant de prédire plus d'un motif, l'algorithme procède d'abord par masquer toutes les occurrences du motif trouvé avant de réinitialiser le processus pour rechercher d'autres motifs.

(b) Gibbs Sampling [148, 149]

Les algorithmes probabilistes ont apporté un gain considérable en temps de calculs comparé aux approches énumératives. Cependant, un des problèmes majeur de l'algorithme EM est la convergence prématurée des solutions trouvées vers des maximums locaux. Une des solutions proposées pour contrer cette limitation a été de relancer l'algorithme un certain nombre de fois et de choisir la solution qui maximise la probabilité de prédiction. Néanmoins, ceci ne garantit toujours pas la convergence de l'algorithme vers un maximum global. L'algorithme de Gibbs a été proposé comme alternative à EM car sa stratégie de découverte, bien qu'elle soit aussi basée sur un échantillonnage aléatoire des positions de départ, ne souffre pas des mêmes limitations que EM. Les étapes principales de l'algorithme de Gibbs sont décrites ci-dessous :

Soit un jeu de données de n séquences $S_1 \ldots\ldots S_n$ et une longueur de motif w.

1. Sélectionner aléatoirement une occurrence de longueur w sur chaque séquence Si (l'algorithme utilise toutes ou une sous fraction de séquences).
2. Choisir aléatoirement une séquence S_k.
3. Créer une matrice M à partir de toutes les occurrences trouvées sur les $k-1$ séquences (sans S_k)
4. Pour chaque position i de S_k, estimer en utilisant le modèle M, le score de l'occurrence de longueur w qui commence à la position i. Ce score représente la probabilité que cette occurrence soit générée à partir du modèle M plutôt que d'un modèle aléatoire.
5. Dans l'algorithme EM, l'occurrence ayant le score le plus élevé est sélectionnée systématiquement sur chaque séquence. Dans l'algorithme de Gibbs, le choix de l'occurrence se fait de manière stochastique. Au départ de l'algorithme, toutes les

occurrences ont la même probabilité de faire partie du modèle. Après chaque itération, l'occurrence ayant le meilleur score n'est pas forcément celle qui sera choisie. Afin de maximiser les chances de convergence vers une solution optimale, plusieurs itérations de l'algorithme sont nécessaires.

Dans le cas des algorithmes basés sur les stratégies EM ou de Gibbs Sampling, il est conseillé de relancer l'algorithme plusieurs fois sur le même jeu de données afin d'augmenter la probabilité de trouver toutes les solutions potentielles. Une étape de comparaison est toutefois nécessaire pour filtrer le résultat final (Choisir les meilleurs motifs, éliminer la redondance, etc…). Grâce à leur nature probabiliste, les algorithmes basés sur les stratégies EM et Gibbs Sampling sont généralement plus rapides que les approches énumératives. Cependant, l'étape d'initialisation des modèles affecte considérablement le résultat final [104, 106].

Table 3. Algorithmes probabilistes et énumératifs pour la découverte de motifs d'ADN liés par les FTs

Algorithme	Approche	Classification
EM	Expectation maximization	PWM
MEME	Expectation maximization	PWM
LOGOS	Expectation maximization	PWM
Improbizer	Expectation maximization	PWM
PhyME	Expectation maximization	PWM
EC	Algorithme Génétique	Autres
FMGA	Algorithme Génétique	Autres
GAME	Algorithme Génétique	Autres
Gibbs sampler	Algorithme Génétique	PWM
MACAW	Algorithme Génétique	PWM
AlignACE	Algorithme Génétique	PWM
ANN-Spec	Algorithme Génétique	PWM
Bioprospector	Algorithme Génétique	PWM
GLAM	Algorithme Génétique	PWM
SeSiMCMC	Algorithme Génétique	PWM
PhyloGibbs	Algorithme Génétique	PWM
GibbsST	Algorithme Génétique	PWM
by Staden	Énumération	Consensus
WordUP	Énumération	Consensus
Oligo-Analysis	Énumération	Consensus
Dyad-Analysis	Énumération	Consensus
YMF	Énumération	Consensus
ITB	Énumération	Consensus
Weeder	Énumération	Consensus
MOPAC	Énumération	Consensus
DMotif	Énumération	Consensus
MaMF	Enumeration	Consensus

PWM : Matrice de poids de positions (position weigth matrix)

Table 4. Algorithmes appartenant à d'autres catégories d'approches de découverte de motifs d'ADN liés par les FTs

Algorithme	Approche	Classification
MUSA	Biclustering	Autres
EMD	Clustering-based ensemble	Autres
PhyloCon	Consensus	
QuickScore	Consensus	PWM
MotifSeeker	Data fusion and ranking	Autres
MobyDick	Dictionary	Consensus
WordSpy	Dictionary	Consensus
Footprinter	Dynamic programming	Autres
WINNOWER	Graph	Autres
MDScan	Greedy algorithm	Autres
Projection	Hashing	Autres
MITRA	Prefix tree/Graph	Consensus
PhyloScan	Scanning	
PHYLONET	Sequence alignment	
SMILE	Suffix tree	Consensus
Verbumculus	Suffix tree	Consensus
Consensus	Weight matrix	PWM

5.2 Autres approches pour la découverte de motifs d'ADN liés par les FTs

5.2.1 Approches basées sur l'utilisation de la conservation des éléments cis-régulateurs entre les espèces proches

L'organisation et la fonction génomique des gènes sont intimement liées à la façon dont ils ont évolué. La figure 9 montre l'histoire évolutive des gènes qui codent pour 1391 FTs et leurs orthologues pour lesquels des données sont disponibles à travers 24 génomes eucaryotes allant de la levure au chimpanzé. Il y a cinq groupes de facteurs de transcription

avec des profiles distincts de conservation: ceux présents uniquement chez les primates; principalement chez les mammifères, les vertébrés, les métazoaires, et dans la plupart des eucaryotes, y compris la levure [23] . L'apparition de nouveaux gènes codant pour des FTs coïncide avec l'émergence de la complexité des organismes et leur a permis de développer de nouvelles fonctionnalités [150].

Plusieurs études ont montré que les gènes codant pour les FTs ont subi une pression sélective par rapport à d'autres gènes [151, 152]. Un des exemples les plus connus est celui de FT forkhead box 2 (FOXP2). Ce FT est impliqué dans le développement du langage chez l'humain [153]. Bien que le gène codant pour ce FT soit conservé à travers les mammifères, il contient deux changements d'acides aminés qui sont présents dans le gène humain, mais pas chez les autres primates, ce qui suggère fortement qu'il a été ciblé par la sélection lors de l'évolution du génome humain [154].

En parallèle, il existe des preuves pour une sélection positive dans les séquences promotrices du génome humain, régions connues pour être riche en sites de liaison à l'ADN des FTs. Ceci a été observé principalement pour les gènes impliqués dans la fonction neuronale et la nutrition [155]. À leur tour, ces différences ont eu un effet direct sur l'activité des FTs. Ceci a été démontré dans une étude de ChIP-chip qui a ciblé quatre régulateurs tissus spécifiques exprimés dans le foie et hautement conservée entre l'humain et la souris. Cette étude a montré que 40% à 90% des sites de liaison à l'ADN des FTs diffèrent entre les deux organismes [156]. Au niveau de l'expression des gènes, des comparaisons de transcriptomes de primates ont montré qu'un petit sous-ensemble de gènes - en particulier des FTs – ont des niveaux d'expression différents chez l'humain comparé aux autres primates. Alors que la plupart des gènes ont maintenu des profiles d'expression similaires [157, 158].

Plusieurs études récentes ont montré qu'une fraction significative des régions cis-régulatrices, sont conservées entre plusieurs espèces [156, 159-164]. Partant de ces observations, une nouvelle catégorie d'approche pour la prédiction des sites de fixations de

FTs a été développée. Ces approches assument que les régions régulatrices sont sujettes à une pression sélective à travers l'évolution et par conséquent elles sont en générale conservées parmi les espèces proches. En effet, plusieurs régions non codantes sont conservées entre des génomes de vertébrés [23, 33]. Par exemple, il a été montré que 98% des éléments régulateurs connus dans le tissu musculaire, sont localisés dans la région la plus conservée entre l'humain et la souris [163].

Deux stratégies ont été implémentées pour utiliser la conservation dans le processus de découverte de sites liés par les FTs. Dans la première méthode on identifie d'abord les régions conservées entre deux ou plusieurs espèces, ensuite ces régions sont analysées par les algorithmes de découverte de motifs pour trouver des sous séquences enrichies et qui pourraient correspondre à des sites liés par les FTs [165, 166]. La deuxième stratégie consiste à utiliser un ou plusieurs algorithmes de découverte de motifs sur le jeu de données d'intérêt, ensuite la conservation des motifs prédits est examinée [156, 167]. Un autre type d'approches utilise l'assomption que les gènes orthologues partagent les mêmes mécanismes de régulation et par conséquent sont sous le contrôle d'un même ensemble de FTs. Plusieurs études ont soutenu l'idée que l'utilisation de la conservation des éléments cis-régulateurs contribue à réduire le nombre de faux positifs tout en garantissant une sensibilité acceptable de l'approche utilisée [168, 169]. Cependant ce paramètre devrait être utilisé avec précaution si l'on considère les études qui ont montré par exemple que le taux de conservation des sites liés par les FTs entre l'humain et le rat est au dessous de 30% [150]. Dans une autre étude, Duncan et collaborateurs [156] ont analysé le profile de liaison à l'ADN de quatre (04) FTs (FOXA2, HNFA1, HNFA4 et HNF6) chez l'humain et la souris par la technique de ChIP-chip. Leur résultats montrent que dépendamment des FTs, 41%-89% des promoteurs de gènes orthologues liés par un des FT chez une espèce ne l'est pas chez l'autre par ce même FT. Un inconvénient majeur de ces approches, est l'assomption que le taux de mutation dans les régions régulatrices a évolué de la même façon entre toutes les espèces proches [170].

5.2.2 Approches pour la découverte des modules cis-régulateurs

La régulation des gènes est un processus complexe, contrôlée par une action combinée de plusieurs FTs [171]. Les sites liés par ces derniers, sont généralement organisés en modules dans les régions régulatrices des gènes cibles [172]. Par définition un module cis-régulateur (CRM) est une région d'ADN de quelques centaines de paires de bases composée de plusieurs sites [105]. Par exemple, chez la levure les CRMs constituent des régions d'environ ~300bp qui contiennent en moyenne 10 sites de liaison de FTs [173].

Plusieurs approches destinées à la découverte des modules cis-régulateurs ont été développées [174, 175]. Certaines d'entre elles utilisent l'information sur les interactions protéines-protéines connues ainsi que l'arrangement spatial de leurs sites de fixation à l'ADN [176, 177]. D'autres utilisent les algorithmes de découverte de motifs classiques pour prédire des motifs uniques. Ensuite toutes les paires de motifs dans une fenêtre de longueur fixe, sont analysées à la recherche de groupes de motifs enrichis dans cette fenêtre comparée à ce qui est attendu aléatoirement [134]. Deux types de stratégies sont adoptés par ces approches :

(a) Les méthodes basées sur la prédiction des sites enrichis dans une fenêtre de taille fixe, comparé à ce qui est attendu aléatoirement. Le principe de ces approches est de scanner un ensemble de régions génomiques avec des modèles de TFBSs connus ou prédits par les outils de découverte de motifs. Parmi les outils appartenant à cette catégorie on cite MSCAN [178], MCAST [179] et CisPlusFinder [180].

(b) Les méthodes probabilistes qui implémentent un modèle de Markov caché (HMM) pour représenter l'ensemble des TFBSs formant le CRM. Parmi ces outils, ClusterBuster [181], Stubb [182] et CisModule [183].

A l'instar de TRANSFAC et de JASPAR, des bases de données qui recensent l'information sur les interactions connues entre FTs et l'organisation des modules cis-régulateurs ont été crée. C'est le cas par exemple de la base de données TransCompel [184]. Chaque entrée dans la base de données représente un motif composite de deux sites liés par les FTs, la

méthode expérimentale utilisée pour vérifier si les deux FTs interagissent ensemble ainsi que l'information sur le gène associé à cet élément régulateur.

5.2.3 Approches basées sur l'utilisation des propriétés structurales des domaines de liaison à l'ADN des familles de FTs.

Les séquences d'ADN reconnues par les FTs sont essentiellement définies par la structure de leur domaine de liaison à l'ADN (DBD) [21]. Souvent, les séquences prédites par les outils de découverte de motifs ne reflètent pas la spécificité de la liaison à l'ADN du FT d'intérêt. Certains FTs lient l'ADN en monomère, d'autres en homo ou en hétéro-dimères.

En général, les FTs appartenant à une même famille, par exemple, ont tendance à lier le même profile de motif [185]. Par exemple, plusieurs des membres de la famille bZIP, dont fait partie le FT AP1, lient des variantes du motif TGANTCA [186]. De la même façon, la famille des protéines HLH lient souvent un motif composé de la boite E-boxe (CANNTG) [187]. Toutefois, leur spécificité à ce motif varie selon les deux nucléotides qui composent l'espacement du milieu ainsi que les nucléotides à proximité. Un autre exemple est celui de la famille des récepteurs nucléaires, pour lesquels le domaine de liaison à l'ADN est conservé au cours de l'évolution [188].

Certaines approches de découverte de motifs ont tenté de modéliser l'information sur la structure des FTs et leur mode de liaison à l'ADN pour mieux prédire les séquences liées par les membres de la même famille [185, 189, 190]. MacIsaac et collaborateurs ont développé l'outil THEME [191] qui, à partir des données de ChIP-chip, prédit les sites qui pourraient être liés par un ou plusieurs FTs en utilisant l'information sur les sites liés par les différentes familles de FTs. Leur hypothèse de départ est que les FTs appartenant à une même famille ont tendance en général à lier le même profile du motif d'ADN. Toutefois, ces conclusions ne pourraient pas être généralisées à cause de :

(a) La difficulté d'obtenir l'information sur la structure des DBD de tous les FTs.
(b) Certains FTs appartenant à une même famille peuvent lier des profiles de motifs différents.

5.3 Les approches intégratives

Il est clair que la présence d'une séquence similaire au profile lié par un FT n'est pas suffisante à elle seule pour assurer la liaison *in vivo* de ce FT à l'ADN ainsi que la fonctionnalité de cette interaction. L'environnement cellulaire, la position et l'orientation de la séquence d'ADN ciblée, la méthylation de l'ADN, le positionnement du nucléosome, et l'état de la chromatine sont des paramètres importants dans la détermination de la liaison et de la fonctionnalité de ce site [192, 193]. Il a été montré que la liaison d'un FT à son site est dépendante de cet environnement cellulaire [97]. Par exemple, des expériences sur le facteur de transcription p53 ont montré que la régulation par ce FT varie selon le type de stimuli cellulaire, sa durée ainsi que le type cellulaire dans lequel l'expérience a été réalisée [194]. Chez les eucaryotes, plusieurs études ont montré que les régions liées *in vivo* par les FTs sont déplétés de nucléosomes comparés aux sites non liés [66, 195, 196].

Harbison et collaborateurs ont analysé le profile de liaison à l'ADN de 203 FTs chez la levure par la technique de ChIP-chip [197]. Ils ont montré que 87% des sites liés sont localisés conjointement dans les segments d'ADN qui séparent les nucléosomes (régions de liaison) et dans les régions déplétées de nucléosomes.

Afin d'améliorer la qualité des outils de découverte de motifs, il a été proposé d'intégrer l'information sur la liaison à l'ADN d'un FT (à partir des données de ChIP-chip et de ChIP-seq) avec l'information sur les marques de chromatine, comme la méthylation et l'acétylation des histones ainsi que le positionnement des nucléosomes [198, 199]. L'idée étant d'utiliser un outil de découverte de motifs pour prédire les sites potentiels, suivi d'un filtre qui utilise une information supplémentaire, comme par exemple des marques de chromatine active, pour éliminer les prédictions qui ne répondent pas aux critères fixés. Parmi ces méthodes on peut citer FIMO [200] et CENTIPEDE [96]. Dans CENTIPEDE par exemple, les auteurs proposent une mixture basée sur le modèle de Bayes qui intègre l'information sur la séquence, la conservation (une séquence conservée à plus de chance d'être liée), la distance qui sépare la séquence du TSS et les modifications covalentes des

histones (marques d'histone active ou inactive). Les séquences ainsi prédites sont classées dans deux catégories : liées et non liées. Dans FIMO, les auteurs ont montré que l'utilisation de certaines marques d'histones (H3K4me1, H3K4me3, H3K27ac et H3K9ac), connues pour être associé avec une transcription active, améliore significativement les performances de leur outil.

Ces analyses démontrent l'importance et la nécessité d'utiliser les méthodes intégratives pour la prédiction des TFBSs. Les outils de découverte de motifs actuels abordent la problématique d'un seul angle, qui est celui de prédire la séquence liée par le FT indépendamment de son environnement cellulaire. Cette prédiction peut être améliorée en associant l'information sur la séquence avec l'accessibilité aux régions liées (chromatine ouverte ou condensée), les modifications covalentes des histones et l'information sur la méthylation de l'ADN. Cependant, il est important de noter que ces approches sont tissus spécifiques. Toutes les expériences devraient être effectuées dans le même tissu cellulaire et sous les mêmes conditions expérimentales.

Un autre niveau de complexité, auquel est confrontée la découverte des TFBSs, est celui des outils utilisés pour localiser et extraire le signal significatif. Les limitations observées avec chacun des outils existants suggèrent qu'il est plus judicieux d'utiliser plusieurs d'entre eux afin de maximiser les chances de trouver tous les motifs potentiels. Harbison et collaborateurs [197], ont testé six outils de découverte de motifs en utilisant des données de ChIP-chip contre 172 FTs chez la levure. Ils ont montré que chacun des outils testés a pu prédire au moins un motif non trouvé par les autres. Ils ont montré aussi qu'aucun outil ne performe complètement comparé aux autres.

Comme mentionné ci-dessus, les approches de découverte de TFBSs doivent trouver un compromis entre la complexité algorithmique et la significativité biologique des motifs prédits. En effet, les approches exhaustives, même si elles permettent une meilleure couverture des données analysées, ne font pas l'unanimité à cause de leur temps de calculs qui peut durer des semaines. Les approches probabilistes, quand à elles sont plus rapides mais traitent une sous fraction des données seulement et par conséquent leurs résultats ne

sont pas exhaustifs. Une autre alternative propose d'analyser le même jeu de données en utilisant plusieurs outils [201-203]. Toutefois, ces approches nécessitent des niveaux de traitement supplémentaires pour formater les résultats, filtrer les redondances et standardiser les statistiques utilisées par les différents outils combinés. Il n'est pas toujours facile de réaliser ces différentes tâches car les outils utilisent des techniques différentes pour extraire les motifs et le plus souvent des statistiques différentes pour évaluer la significativité des motifs prédits. Ce dernier point en particulier est très critique car les scores affectés aux motifs prédits ne reflètent pas nécessairement l'affinité du FT à son site de liaison à l'ADN [20, 106].

5.4 Identification des motifs prédits

Une autre étape aussi importante que nécessaire pour compléter le processus de la découverte des sites d'ADN liés par les FTs , consiste à utiliser les bases de données de facteurs de transcription [49, 137] ainsi que des outils de comparaison de motifs [204, 205] afin d'identifier parmi les motifs prédits, ceux qui pourraient être associés à des TFBSs connus. Les motifs qui n'ont pas été associés à aucun des sites connus seront considérés comme de nouvelles cibles potentielles pour des FTs non encore caractérisés ou bien de nouvelles variantes de sites liés par des FTs dont le site de liaison à l'ADN a été déjà caractérisé.

5.5 Approches de découverte de motifs spécifiques aux données issues des expériences de ChIP-chip et de ChIP-séquençage

Les expériences de ChIP-chip et de ChIP-seq ont pour but de cartographier le profile de liaison à l'ADN des FTs. Elles génèrent des milliers de fragments génomiques liés par le FT immuno-précipité, dont la taille varie en générale entre 300 bp à 1kb. Dans la majorité des cas, le motif d'ADN lié par le FT ciblé est connu. Toutefois, il arrive qu'on s'intéresse à un FT pour lequel le site de liaison à l'ADN n'a pas été caractérisé. Bien que ces approches confirment la présence du FT dans le complexe immuno-précipité, cependant elles n'offrent aucune information sur le mécanisme par lequel le FT interagit avec l'ADN

(liaison directe ou recrutement du FT par un autre partenaire transcriptionnel), ni sur la position exacte du site lié. Le défit des algorithmes de découverte de motifs appliqués à ces données, est d'une part de prédire le site liés par le FT immuno-précipité et d'autre part de prédire les sites des FTs qui médient sa liaison à l'ADN par des interactions protéine-protéine.

Toutefois, il est important de s'attarder sur l'analyse de quelques paramètres qui influencent la découverte de motifs dans les données de ChIP-chip et de ChIP-seq :

a) Une des contraintes majeures rencontrées par les outils de découverte de motif dans les données de ChIP-chip et de ChIP-seq, est le nombre élevé de fragments générés par ces techniques, qui est en générale de l'ordre de 10000 fragments ou plus et dont la taille varie entre 300bp à 1kb. Certains algorithmes, à l'instar de MDscan [128], utilisent les fragments les plus enrichis (en se basant sur leur score d'enrichissement) pour créer les motifs initiaux ensuite, les fragments restants sont utilisés pour chercher des occurrences additionnelles des motifs déjà trouvés. Cette stratégie permet de prédire les motifs présents dans la sous-fraction de fragments utilisés. En général, le motif dit primaire, celui qui est lié par le FT immuno-précipité est facilement prédit par ces approches. Alors que les motifs secondaires, liés par les partenaires de ce FT, en particulier ceux qui sont moins fréquents et ceux présents dans la fraction de fragments les moins enrichis seront perdus. D'autres approches, à l'instar de MEME [206] et Gibbs sampler [189] se limitent à faire la découverte de motifs dans un sous-ensemble de fragments sélectionnés aléatoirement. Bien qu'elles permettent un gain considérable en temps de calcul, ces stratégies sont limitées uniquement aux motifs présents dans les fragments sélectionnés. Une solution à cette limitation serait de relancer l'algorithme plusieurs fois pour s'assurer de couvrir le maximum de fragments étant donné qu'à chaque exécution un nouveaul sous ensemble de fragments sera échantillonné. Cependant, une étape supplémentaire est nécessaire afin de comparer les résultats de chaque exécution, d'éliminer les redondances et de sélectionner les candidats potentiels. La taille et le nombre de séquences analysées sont deux paramètres qui influencent de près le résultat de la découverte de motifs. Dans le cas où l'analyse est réalisée sur peu de séquences, on risque de ne pas avoir assez d'information pour trouver tous les motifs qui

s'y trouvent. Hu et collaborateurs [124] ont montré que les performances des outils de découverte de TFBSs augmentent significativement après un certain nombre de séquences. Il est important de noter qu'un nombre élevé de séquences peut augmenter le bruit et obstruer le signal positif. Les motifs dégénérés (de faibles contenu d'information), ou bien ceux dont la distribution est proche de celle observée aléatoirement sont d'autant plus difficiles à trouver.

b) L'affinité du FT à chacune des régions de ChIP-chip ou de ChIP-seq identifiées par les algorithmes d'identifications de pics [207], est mesurée soit par l'enrichissement de la sonde (ChIP-chip) ou bien du fragment séquencé (ChIP-seq). Cette information fournie une indication sur la position approximative des sites liés dans les régions de ChIP-chip et de ChIP-seq. En effet, les sites liés sont enrichis dans la région autour du centre des fragments de ChIP-chip et de ChIP-seq (100-200bp) [208]. Ce paramètre a été exploité par certains outils de découverte de motifs [206, 209] afin de réduire la taille des séquences à analyser et par conséquent limiter l'analyse aux régions ayant plus de chance de contenir les motifs liés.

c) Souvent, les fonctions de score utilisées pour évaluer la significativité des motifs prédits ne reflètent pas l'affinité *in vivo* des interactions FT-ADN. Il est important de développer de meilleures statistiques (fonctions de scores) capables de modéliser l'exceptionnalité du signal positif et de le distinguer du bruit. Les statistiques utilisées par les outils existants se limitent à évaluer la significativité statistique en comparant la distribution du motif dans les régions liées à celle observée dans des régions contrôles.

d) Une des étapes importantes du processus de découverte de motifs est l'interprétation des résultats. En effet, seulement une sous-fraction des motifs prédits pourrait être associée à des sites liés par les FTs. Ceci, soulève le point de comment attesté de la fonctionnalité des motifs prédits. La réponse la plus intuitive est de tester expérimentalement tous ces motifs. Dans la pratique, cette tache est difficile à réaliser, voir même impossible, à cause

du cout associé aux expériences et le temps de réalisation. Une des alternatives proposées a été d'utiliser des informations supplémentaires pour donner plus de poids aux motifs qui ont plus de chance d'être liés *in vivo* par les FTs. Par exemple l'état de la chromatine, le positionnement des nucléosomes, l'information sur les partenaires transcriptionnels du FT d'intérêt et enfin les données d'expression des gènes cibles. Plusieurs études ont montré que la qualité des prédictions a été améliorée en combinant plusieurs types d'informations [20, 200]. Toutefois, toutes ces données devraient être obtenues à partir d'expériences réalisées dans le même tissu cellulaire et sous les mêmes conditions expérimentales.

e) Amélioration de la résolution des techniques expérimentales d'identification des régions liées par les FTs: une des limitations des techniques *in vivo* d'identification des régions liées par les FTs est la taille des fragments générés, qui varie de 300bp à 1 kb. Alors qu'en général les sites liés par les FTs sont de taille inferieure à 30bp. A l'exception des positions qui composent les sites liés par les FTs, le reste des positions sur les fragments de ChIP-chip et de ChIP-seq représentent du bruit. Ce dernier, affecte considérablement les performances des algorithmes de découverte de motifs qui génèrent des taux considérables de faux positifs. Récemment, Rhee et ses collaborateurs ont proposé une nouvelle technique appelé ChIP-exo [210]. ChIP-exo est une technique génomique *in vivo* pour l'identification des régions liées par les FTs. Elle utilise une exo nucléase (lambda) qui dégrade les fragments d'ADN liés par le FT du côté 5'-3'. Dans la technique de ChIP-exo, après immuno-précipitation des fragments liés par le FT d'intérêt en présence d'un anticorps spécifique, des adaptateurs sont ajoutés aux extrémités de ces fragments. Ensuite l'exonucléase lambda est utilisée pour digérer les fragments d'ADN double brins à partir du côté 5` jusqu'à ce que la digestion soit bloquée par la région liée par le FT. Le pontage est alors renversé et après amplification des fragments liés, ces derniers sont séquencés.

Alors que la technique de ChIP-seq offre une résolution de l'ordre de 100-300 bp, ChIP-exo permet d'obtenir les fragments dont la résolution est de l'ordre de quelques nucléotides. En examinant le profile de liaison à l'ADN de cinq (05) FTs par les trois techniques, à savoir la technique de ChIP-chip, ChIP-seq et de ChIP-exo, on a estimé environ 50% des

fragments obtenus par ChIP-chip et 30% des fragments obtenus par ChIP-seq n'ont pas été trouvés par ChIP-exo [210].

DEUXIÈME PARTIE : Méthodes

Chapitre 6

Une nouvelle approche pour la découverte des motifs d'ADN liés par les facteurs de transcription, adaptée aux données de ChIP-chip et de ChIP-séquençage

La caractérisation des mécanismes de régulation des gènes nécessite à priori une compréhension complète du fonctionnement des facteurs de transcription, en particulier la caractérisation de leurs sites de fixation à l'ADN. Le développement des techniques génomique à large échelle comme la technique de ChIP-chip et de ChIP-séquençage, a permis de générer des quantités importantes d'informations sur le profile de liaison à l'ADN de plusieurs FTs. Cette grande avancée dans le domaine expérimentale a motivé le développement d'une panoplie d'approches computationnelles pour analyser et exploiter ces données. Cependant, nos connaissances des FTs et de leurs modes d'action et de liaison à l'ADN sont loin d'être exhaustives et complètes. Il est important de développer une nouvelle génération d'approches computationnelles qui soient capables de modéliser correctement l'information biologique disponible et de créer des modèles prédictifs performants.

Dans ce chapitre, nous avons tenté de répondre à certaines questions relatives à la découverte des sites de fixation de FTs à l'ADN en utilisant des données de ChIP-chip et de ChIP-séquençage. Nous nous sommes posés quelques questions auxquelles nous avons tenté de répondre en développant une nouvelle approche pour la découverte de sites de fixation à l'ADN des FTs. Nous proposons SAMD-ChIP (Statistical Analysis For DNA Motif Discovery in ChIP-chip and ChIP-seq data), une nouvelle approche pour la découverte des sites de fixation à l'ADN de FTs. Elle est composée des modules suivants :

extraction des motifs enrichis autour des régions liées, optimisation des motifs, regroupement des motifs similaires, évaluation des propriétés statistiques des motifs et enfin identification des motifs prédits en utilisant les banques de FTs.

Les analyses conduites dans le cadre de ce chapitre ont tenté de répondre aux questions suivantes :

1. *Peut-on développer un nouvel algorithme de découverte de motifs d'ADN liés par les FT dans les données de ChIP-chip et de ChIP-seq, qui soit plus performant que ceux qui existent déjà ?*

2. *Quelles statistiques utiliser pour mesurer l'exceptionnalité des sites enrichis à proximité du centre des fragments de ChIP-chip et de ChIP-seq ? et à quel point ces statistiques reflètent-elles les propriétés de distribution des sites liés par les FTs ?*

3. *Quel ensemble de référence (contrôle) utiliser pour évaluer si la distribution des motifs prédits est différente d'une distribution aléatoire (uniforme) ?*

4. *Comment éliminer la redondance et regrouper les motifs similaires afin de faciliter l'interprétation des résultats ?*

SAMD-ChIP est présenté sous forme d'un serveur web qui permet à l'utilisateur de soumettre les données à analyser ainsi qu'un ensemble de paramètres via une page web. Une fois l'analyse terminée, l'utilisateur sera notifié par l'envoi d'un message électronique avec l'adresse de la page web où sont enregistrés les résultats de l'analyse. Les utilisateurs peuvent aussi télécharger une version locale de SAMD-ChIP. Le code source est aussi disponible pour les utilisateurs sous certaines conditions.

SAMD-ChIP: A new hybrid algorithm for DNA motif discovery adapted to ChIP-chip and ChIP-sequencing data

M. AID[1,2], S. LEMIEUX[2], S. MADER[1,2]

[1]Biochemistry Department, University of Montreal, Quebec, Canada

[2]IRIC, Institute of Research in Immunology and Cancer, University of Montreal, Quebec, Canada

ABSTRACT

Accurate prediction of transcription factor binding sites (TFBSs) is a crucial step towards understanding gene regulation mechanisms. In this paper, we describe SAMD-ChIP, a new *do novo* DNA motif discovery tool adapted to ChIP-chip and ChIP-sequencing data. SAMD-ChIP combines enumerative and stochastic strategies to predict enriched motifs in the vicinity of ChIP peak summits. Our approach is an automated pipeline that includes motif discovery, motif clustering, motif optimization and finally motif identification using transcription factor (TF) databases. We compared tool performances by evaluating their ability to predict the motif bound by the immuno-precipitated TF (primary motif), to detect sites of other known TFs that can tether the immuno-precipitated TF to DNA, and to predict new motifs. We demonstrate the superior performance of SAMD-ChIP on synthetic datasets as well as on a collection of ChIP-chip and ChIP-sequencing data performed against 7 TFs in MCF7 breast cancer cells. SAMD-ChIP outperforms state-of-the-art motif discovery tools in terms of the number of predicted motifs and the prediction of rare and degenerate motifs. In particular, SAMD-ChIP efficiently identifies gapped motifs such as inverted or direct repeats bound by nuclear receptors and composite motifs resulting from the association of different single TF binding sites.

INTRODUCTION

Regulation of gene expression is a complex process that implicates several actors including transcription factors (TFs). TFs regulate gene expression by binding to specific and short DNA sequences of 6 to 30 bp length (6), commonly called cis-regulatory elements or transcription factor binding sites (TFBSs). TFs can recognize a range of DNA sequences that may differ at several positions according to a pattern imposed by the TF DNA binding domain (DBD) (11).

Different statistical models are used to represent the set of DNA sequences recognized by a specific TF (6). Position weight matrices (PWM) represent the most widely used model (6,7). In this model, the motif is represented as a matrix of nucleotide weights, which represent nucleotide preferences at each position in the motif. While PWMs do not capture nucleotide interdependencies within the motif, they provide a reasonable approximation of TF-DNA binding affinity and offer easily interpretable models (8). More complex models, such as neural networks and Markov chains provide more complete information compared to PWM, but they require a larger number of validated TFBSs for accurate training (9,28). In addition, these models are available only for some TFs, while TF databases like TRANSFAC (13) and JASPAR (12) have compiled PWMs for all known TFBSs.

TFs act rarely alone (55), and their DNA binding sites tend to cluster together to ensure maximal transcription efficiency. Several methods have been developed for cis-regulatory module (CRM) discovery (56,57). They are based on searching for closely located motifs with variable spacers. Because of the complexity of CRM composition, in general these methods are computationally intensive.

In the last decade, developments in functional genomics approaches, in particular Chromatin immuno-precipitation coupled to DNA microarrays ChIP-chip (22,23,24) or DNA sequencing (ChIP-seq) (20,21,25), have provided high resolution genome-wide maps of the *in vivo* TF-DNA interactions. With the rapid advances in next-generation sequencing (NGS) technologies, ChIP-seq is becoming the standard approach for *in vivo* mapping of

TFBSs because it allows whole-genome coverage and offers greater sensitivity compared to ChIP-chip (25). In ChIP-seq assays, proteins are first crosslinked with DNA using formaldehyde then fragmented, usually by sonication, into random size-selected fragments (200-800 bp). Protein-DNA complexes are immuno-precipitated in the presence of a specific antibody against the TF of interest. After reversing the crosslink, size-selected DNA bound fragments (100-200 bp) are sequenced, generating millions of short reads. The next step consists of aligning these reads to a reference genome to identify their genomic locations. These regions are then analysed by peak detection methods (26,27) to find regions of higher sequencing density compared to controls. The output of these methods is a set of genomic fragments bound directly by the immuno-precipitated TF (primary TF) or indirectly via protein-protein interactions.

A rising question is how to exploit these large datasets in order to discover potential short DNA segments bound by the primary TF and TFs which mediate its DNA binding? This requires characterizing over-represented DNA motifs (6-30 bp) in a set of ChIP peak fragments of 200-500 bp without *a priori* information about motif length, diversity and frequency. At this time, a small number of motif discovery approaches have been developed specifically for ChIP-chip or ChIP-seq data analysis (30,31,32,33). As input, these methods need two data sets. The first one represents the identified ChIP peaks (positive sequences) and the second is a background model (negative sequences) against which motif exceptionality will be evaluated. Usually the background model is represented by a Markov chain (3) or randomly selected regions from the same genome (5). It is expected that motifs bound by the primary TF or by its tethering partners will be significantly enriched in bound regions compared to background. ChIP-chip and ChIP-seq experiments generate fragments ranging in length from 200 bp to 1 kb. In addition, bound motifs cluster around the ChIP peak summit (36,37,56). This positional preference property was integrated by some motif discovery tools. For example MEME-ChIP (32) trims automatically the submitted ChIP fragments and performs motif discovery in the 100 bp

central regions. In HMS tool (33), a 200 bp region around the ChIP peak summit is used. However, the ideal window length is not fixed and may vary with the TFs.

Another challenging problem complicating *do novo* motif discovery process is the large number of fragments generated by ChIP-chip and ChIP-seq experiments, which are in the order of tens of thousands. Performing motif discovery using all these fragments is time consuming. In addition, because of the different biases introduced by both experimental and computational experiments, some of these fragments might be false positives. Hence their presence will affect the performance of motif discovery tools. Different strategies are used to select a subset of ChIP peak fragments as positive sequences to perform motif discovery. For example, the MDscan tool (14) proceeds by selecting the most highly enriched ChIP-chip fragments based on their binding scores. In doing so, only motifs present in the highly scored ChIP peaks will be targeted and may be predicted, biasing against low affinity motifs. Other tools, such as MEME-ChIP (32), select randomly a subset of ChIP fragments to perform motif discovery. Approaches based on probabilistic models, which use local search techniques such as expectation-maximization (EM) (2) or Gibbs sampling (15,16) tend to first initialize the model by choosing random starting positions within the set of randomly selected fragments, then proceed through several iterations to increase the probability of finding a globally optimum motif. Thus, these tools have to be run several times to ensure a maximal coverage of the data set. However, we need additional steps to compare results from different runs and eliminate redundant motifs.

During the motif initialization step, some of *de novo* motif discovery algorithms (15,16,2) proceed by running the algorithm several times, masking the found motif instances at each algorithm run. Consequently, the order in which motifs are found will affect considerably the final results as such masking could remove informative motif instances.

Herein, we describe SAMD-ChIP (Statistical Analysis for DNA Motif Discovery in ChIP peak fragments), a new *de novo* motif discovery tool adapted to ChIP-chip and ChIP-seq data, which implements a hybrid approach by combining (a) a stochastic strategy to select

ChIP fragments and (b) an enumerative algorithm to find motif instances. Our approach incorporates information about motifs enrichment and motif positional preferences in the vicinity of ChIP peak summit to predict potential TFBSs. Analysis of simulated data as well as ChIP-chip/seq data showed that our approach performs better compared to the state-of-the-art motif discovery tools in term of the number of predicted motifs and their accuracy. Interestingly, our approach outperforms other algorithms in predicting weak motifs, rare motifs and motifs containing gaps. SAMD-ChIP offers a user-friendly interface and reports the list of optimized and enriched motifs, ranked by their fold enrichment in the vicinity of ChIP peak summits compared to a selected background.

MATERIAL AND METHODS

Motif initialization

SAMD-ChIP aims to find all motifs of length *k,* ranging from a minimum to a maximum motif length (for example from 6 to 20 bp) that are significantly enriched in a user-defined central window of *N* input sequences (positive sequences) compared to what is expected in a set of background sequences (negative sequences).

The motif initialization step aims at creating a collection of initial PWMs by compiling a set of similar instances for each PWM. This set of PWMs is selected based on their enrichment in the central window. In this step, we define seed motifs in the central window of a randomly selected fraction *fraction_1* (size *f1*; $f1 \leq$ N) of the *N* input sequences. Each DNA segment of length *k (*from 6 to 20 bp*)* in the central window of each sequence from *fraction_1,* represents a potential binding site and thus a seed motif. First, for each seed, a new fraction of *f2* sequences, *fraction_2*, are randomly selected from the remaining input sequences. Next, for each seed, we search for its most similar instances in *fraction_2*. Sequences in fraction_1 and fraction_2 are selected randomly, thus, we ensure that each input sequence has the same probability to be selected. The number of sequences *f1* and *f2* are defined by the user as input parameters.

While seeds are exclusively selected from the central window of each sequence in *fraction_1*, their best matches may be located at any position on *fraction_2* sequences. The similarity between the selected seed and its current instance is measured using the Hamming distance (1), which is defined as the number of mismatches *d* between the seed and its instance. On each sequence in *fraction_2*, the best match of the current seed is selected as the instance with the lowest number of mismatches *d* over all seed instances. Only instances which are distant from the seed by less than *k/2* ($d \leq k/2$) are compiled into a position weight matrix *T,* whose fold enrichment is then measured as a ratio between the number of seed matches located in all positive sequences and those located in all negative sequences. By default, fold enrichment threshold is set to 1.5 (Figure S1). Algorithm 1 describes the SAMD-ChIP initialization step.

Algorithm 1 SAMD-ChIP initialization step

Given an input data set of *N* sequences

Given a motif length *k*, a fraction_1 of *f1* sequences and fraction_2 of *f2* sequences

Given a central window of length *c_w*

Repeat for each motif length *k* (6 to 20 bp)

Select randomly a sequence S_i from *N* (current sequence in fraction_1)

For each position *pos* in S_i central window

Read current seed on S_i ,located at position *pos*

Select randomly a sequence S_j ($j<>i$) from *N* (current sequence in fraction_2)

Find the best match for the current seed on S_j *; soit seed_match*

Measure the Hamming distance *d*(*seed,seed_match*)

If *d*(*seed,seed_match*) <= *k*/2

Return *seed_match;* S_j*; seed_match position* on S_j

If *seed_match* position in S_j central window, increase *count_c_w*

else increase *count_background*

Select next S_j

Measure *seed_fold_enrichment = count_c_w/count_background*

If *seed_fold_enrichment >=1.5* compile *seed* matches and create *seed_PWM*

Select next *seed* on S_i

Select next S_i

Next motif length *k*

Motif clustering

The clustering step assembles related matrices to create a set of unique matrices for each motif of length *k*. To create clusters of related motifs, we assess the similarity between two matrices α and β with *n* and *m* occurrences respectively by calculating the average similarity over all their occurrences:

$$Similarity\ (\alpha, \beta) = \frac{\sum_{i=1}^{n} \left[\frac{\sum_{j=1}^{m} best\_offset(\alpha_i, \beta_j)}{m} \right]}{n}$$

Where α_i and β_j are the current occurrences in α and β matrices and *best offset* (α_i, β_j) is the best alignment score between α_i *and* β_j.

To estimate matrix similarity threshold, we compared the similarity between random sequences to that of sequences bound by a given TF. To assess the similarity between random DNA sequences, we generated a collection of 1000 random sequences for each of these motif lengths: 9, 15 and 25 bp. The similarity between two matrices is measured as the average similarity over all their occurrences.

Next, within each group of similar length sequences, we measured the similarity between each sequence pair. Sequence similarity was measured by trying all possible offsets to find the alignment, representing the highest similarity between each sequence pair (a column by column comparison was performed for each sequence pair). We observed that the similarity between random sequences did not exceed a fraction of 40% for the different motif lengths (Figure S2 A).

In the second simulation and in order to estimate the similarity between sequences bound by a specific TF, first we selected a set of matrices for different motif lengths from the JASPAR database [50, 184] . Each matrix represents a collection of sequences bound by a specific TF. Then, we measured the similarity between sequences within each matrix. This analysis reveals that the similarity within each matrix exceeds 50% for all motif lengths (Figure S2 B), reflecting the similitude between sequences bound by the same TF.

To test the above observations using independent data sets, we selected two groups of matrices predicted by SAMD-ChIP initialization step. In the first group, we collected a set of distinct matrices (matrices corresponding to different TFBSs). In the second group we selected matrices representing the same TFBS. This analysis showed that the similarity

within a set of matrices representing the same TF exceeds 50% for different motif lengths (Figure S2 E), whereas, this similarity is less than 40% between distinct matrices (Figure S2 D).

According to these results, we set the similarity threshold between matrices related to the same motif to 50%. Matrices verifying this condition will be grouped together to form a unique cluster. Within each cluster, we measured the information content (IC) of all matrices and the matrix with the highest IC is selected. The matrix IC is given by:

$$IC = \sum_{i=1}^{4} \sum_{j=1}^{k} f_{ij} \log \frac{f_{ij}}{b_i}$$

Where, f_{ij} is the probability of nucleotide i at position j and b_i is the probability of the same nucleotide in the background model.

The matrix IC reflects how each column is conserved and how much the nucleotide frequencies at each column differ from what would have been observed by random. The background nucleotide frequencies are obtained by accounting for nucleotide frequencies in the reference genome.

Motif optimization

The objectives of the motif optimization step are (i) to refine initial matrices by selecting their best instances on each sequence, and (ii) to update refined matrices by scanning the fraction of input sequences that are not used during the initialization step in order to find additional instances. The Refine-Update process is repeated until the matrix remains unchanged (Figure 2).

Motif refinement

We implemented a refinement procedure based on EM strategy (2).

First, for a given PWM with n occurrences ($n \leq N$) found in n sequences (*used_sequences_set)*, select randomly a sequence S_i (*i:1 to n*) and remove from the PWM, its occurrence found on S_i . A new PWM is created from the set of *used sequences_set* without S_i. This new PWM is then used to scan the sequence S_i and select the best PWM occurrence (occurrence with the best score according to the PWM), which is then added to the PWM. This process is repeated for each sequence in the *used_sequences_set* until the PWM converges to the same final refined matrix, defined as PWM_r

Motif update

During the initialization step, matrices are created using n sequences ($n \leq N$). The motif update procedure is called in case of all N input sequences are not used during the initialization step. It consists of scanning each sequence in the N-n remaining sequences, (*remaining_sequences_set*) using the refined PWM_r and selects the occurrence with the best score according to the PWM_r. This new occurrence is added to the PWM_r if its score, measured as the sum of nucleotide weights present at each occurrence position, is higher or equal to the PWM_r mean score. The score of a given matrix occurrence *mat_inst,* of length k is the sum of the PWM values for each nucleotide in the occurrence and given by:

$$Score\ (mat\_inst) = \sum_{i=1}^{k} w_{ij}$$

Where w_{ij} represents the weight of the nucleotide j located at position i in the occurrence. The PWM mean score is measured as the mean score over all matrix occurrences. The selection of matrix mean score as a minimum threshold allows keeping PWM variability without loss of information content at the conserved positions.

After the optimization process, some matrices could lose their enrichment around the ChIP peak summits. The fold enrichment of the optimized matrices is revaluated. Only those matrices that kept their enrichment (matrix fold enrichment around the ChIP peak summits ≥ 1.5) are selected for the next analysis.

Evaluation of motif group specificity

The motif enrichment fold allows retaining only PWMs enriched in a specified window close to the ChIP peak summits compared to what is expected in the background, in accordance to the positional bias property of TFBS in the vicinity of the ChIP peak summits. However, this test does not indicate whether the observed fold enrichment could be reproduced randomly or not. To make this assessment for each matrix, we compare its distribution vectors in ChIP peak fragments (*Hist_ChIP)* and in background sequences (*Hist_bg)*. The background matrix distribution vector *Hist_bg* is defined as the mean of the matrix' distribution vectors *Hist_bg* over 1000 simulated background data sets generated by shuffling the ChIP peak fragments sequences using the Ushuffle tool (http://digital.cs.usu.edu/ mjiang/ushuffle/). Like in the ChIP peak fragments, the instance with the best matrix score is selected for each sequence in the background. The matrix distribution vectors *Hist_bg* and *Hist_ChIP* are compared using a Kolmogorv-Smirnov test (KS). We set the KS *P-value* threshold to 0.05. This probability represents the matrix group specificity *P-value*, which reflects the probability to reproduce the motif distribution by random. Motifs with a *P-value* ≥ 0.05 are discarded.

Motif enrichment evaluation

Matrix enrichment in the central window around the ChIP peak summits is evaluated for different matrix cut-offs using a base score cut-off of 40% and 10% increments. Negative sequences (background) are created by shuffling 1000 times the set of positive sequences using Ushuffle tool (http://digital.cs.usu.edu/mjiang/ushuffle/). Enriched matrices are first used to scan the two datasets and all matrix instances having a score equal or higher than the current cut-off are selected irrespective of the number of instances per sequence. To

assess the statistical significance of the observed enrichment in positive versus negatives sequences, we compare both the numbers of matrix instances found in positive and in negative sequences, and the numbers of unique sequences in the two data sets containing at least one matrix occurrence. We evaluate the significance of these comparisons using a Fisher exact test using R language. We expect that matrices representing potential TFBSs, should maintain their enrichment for different cut-offs.

Motif Identification

This step consists of comparing SAMD-ChIP predictions to TF databases in order to identify those matrices representing known TFBSs. To this end, we used STAMP tool (10), which aligns input PWMs against public or user-provided databases of known motifs, and returns lists of the highest-scoring matches. STAMP implements various comparison metrics, alignment methods (local or global, gapped or ungapped), multiple alignment strategies and tree-building methods.

The SAMD-ChIP motif identification module loads each matrix in the STAMP web server to identify the highest-scoring matches among STAMP pre-compiled matrices. In addition, we also implemented a local STAMP version to allow comparison of discovered matrices against user-provided matrices.

SAMD-ChIP implementation

SAMD-ChIP initialization, clustering, optimization and identification steps were implemented using Perl programming language. All statistics were done using R. Final results output was implemented using HTML and Javascripts.

Estimation of the accuracy of motif prediction by motif discovery algorithms.

To test the accuracy of the predicted motifs, we measured the average Hamming distance between the inserted motif instances and the predicted ones (33). The average Hamming distance between the inserted motif α and the predicted motif β is given by:

$$H(\alpha, \beta) = \frac{1}{k} \sum_{j=\{A,C,G,T\}} \sum_{i=1}^{k} \left| \alpha_{ij} - \beta_{ij} \right|$$

Where k is the motif length, α_{ij} represent the nucleotide j at position i of motif α and βij the nucleotide j at position i of motif β.

We consider that a predicted motif instance was relevant if the position is equal or shifted by one base pair from the position of the inserted motif.

Comparison of tool performances

To evaluate the performance of the different motif discovery tools in predicting accurately motifs inserted in simulated datasets, we used the combined Recall and Precision statistic F-measure, which is defined as:

$$F - measure = 2 . \frac{Precision . Recall}{Precision + Recall}$$

Precision is the fraction of retrieved instances that are relevant (Relevant/Retrieved); high precision means that there are more relevant instances that are predicted (true positives) compared to irrelevant ones (false positives). Recall represents the fraction of relevant instances that are retrieved (Relevant/Inserted); high recall means that most of the relevant instances are predicted. A high F-measure value indicates a good compromise between the false positive and false negative prediction rates.

RESULTS

SAMD-ChIP pipeline description

Our approach for the discovery of potential binding sites for the primary TF or TFs that mediate its DNA binding in ChIP sequences takes advantage of the enrichment in these sites near the ChIP peak summits. SAMD-ChIP searches for sets of motifs of specified length *k,* significantly enriched in a user-defined window around the ChIP peak summits compared to a chosen background. Fragments from ChIP-chip/Seq experiments can be submitted to SAMD-ChIP in a fasta or bed file format. Sequences of a specified length centered on the ChIP peak summits are extracted from the genome of reference. Background sequences can be defined by the user as an input parameter or generated by SAMD-ChIP, either by extraction of genomic sequences flanking the ChIP fragments locate at 5kb from the ChIP peak summits or by shuffling the ChIP peak fragment sequences. SAMD-ChIP predicts a set of enriched motifs of length *k* by performing the following steps: motif initialization, clustering, optimization, group specificity evaluation and motif identification (Figure 1). Finally, predicted candidates are reported in a user-friendly web interface.

In the motif initialization step, seed motifs are selected from the central window of a fraction of input sequences and related motif instances are selected from another fraction of input sequences. Both fraction sizes are user-determined. For example, one can specify to select seeds from 40 % of input sequences and to search for seed matches in 60 % of input sequences. We assume that each sequence contains either zero or one motif instance for both seed selection and similar instance detection. Seed matches are assembled into a PWM if their Hamming distance is less than *k/2* and the fold enrichment of each matrix is calculated as the ratio of the number of seed instances in the central window over the number of its instances in the background.

The initialization step generates a set of matrices for each motif length *k*. Because we do not mask motif instances at each initialization round, several matrices may share a fraction of those instances. In order to eliminate redundancy, we implemented a two-step clustering module. The first step consists of grouping similar matrices in the same cluster. In the second step, the matrix with the highest information content (IC) is selected within each cluster.

Selected matrices are then refined by selecting their best occurrences on each sequence, and updated by scanning the fraction of input sequences that were not used during the initialization step. The Refine-Update process is repeated until the matrix remains unchanged (Figure 2).

The non-randomness of the distribution of the matrix in the ChIP sequences is then evaluated using the Kolmogorv-Smirnov test and its fold enrichment compared to a series of 1000 backgrounds generated by shuffling the ChIP peak fragments is evaluated at different cut-offs. SAMD-Chip returns all PWMs predicted to be enriched in a central window of ChIP fragments in a user-friendly interface.

Comparison of benchmark DNA motif discovery tools using simulated datasets

The performance of SAMD-ChIP algorithm was compared with that of existing ChIP-chip and ChIP-seq motif discovery tools MEME-ChIP (32), DREME [131] and TRAWLER (30) using simulated and biological data sets (Table S1). MEME-ChIP implements an EM-based approach, while DREME is based on an enumerative strategy and is designed for the discovery of short motifs only. TRAWLER is based on an enumerative search strategy implemented using a suffix tree. Several recent studies have shown the efficiency of MEME-ChIP in discovering TFBSs and most of these studies have reported that MEME-ChIP performances are higher or comparable to other approaches (32,35). We also conducted a comparative analysis (data not shown) on several ChIP-chip and ChIP-seq data and found that MEME-ChIP [131] performs better compared to the state-of-the-art motif discovery algorithms in predicting short and long motifs. In addition, since our approach

implements an enumerative motif extraction strategy, we evaluated its performances compared to TRAWLER.

First, we aimed to evaluate tool performances in predicting single and composite motifs. To this end, we simulated 100 datasets of 100 DNA sequences of length 300 bp, by using independent and equiprobable nucleotides distributions. The central 100 bp were used as positive sequences and the remaining 200 bp were used as negative sequences (background). We created five motif models (Figure 3, Figures S3 and S4) using the RSAT tool (http://rsat.ulb.ac.be/), by varying motif length (short and long motifs) and motif information content (strong and weak motifs). Each motif was created by compiling 100 instances of specific length. Strong motifs are characterized by highly conserved nucleotides at most positions (Figure 3B and 3D). However, weak motifs have more degenerate positions (Figure 3A and 3C). In addition, we used as a model for composite motif the estrogen receptor response element matrix, a palindrome composed of two blocs of 6 bp separated by a 3 nucleotides spacer (Figure 3E). Next, we inserted instances of each motif at randomly selected positions in 100 sequences of 300 pb length. In order to test the effect of motif abundance, motif instances were inserted in different fractions of positive sequences (10%, 30%, 60% and 80%).

We first verified whether the tested tools predicted correctly the number and the location of inserted instances by measuring the average Hamming distance between the inserted motif instances and the predicted ones (33). For each tool, first we calculated the number of retrieved motif instances for each simulation. Then among those instances, we identified the number of instances that correspond to the inserted ones (relevant instances). To compare tool performances we used Precision, Recall and F-measure statistics, which evaluate the compromise between the rates of false positive and false negative predictions.

When the motif is both long and strong, SAMD-ChIP and MEME-ChIP achieved comparable performances in predicting most instances even if the motif is rare (motif inserted in a small fraction of positive sequences), achieving F-measure values that exceeded 80% (Figure 4A, Long-S). On the other hand, TRAWLER showed very low performances (Figure 4A) in predicting the same motif profile. MEME-ChIP and SAMD-ChIP also showed comparable performances in predicting long and weak motifs (Figure 4B) but MEME-ChIP fails in predicting rare motifs (motifs inserted in only 10% of positives sequences). However, for short and strong motif (short_S) we observed that SAMD-ChIP outperforms both MEME-ChIP and TRAWLER (Figure 4A). DREME, which is part of the MEME suite, implements an exhaustive strategy to search for exact words in the first step (no wildcards are allowed). Then in the generalized step, these words are used as "seeds" to search for similar instances by allowing a fixed number of mismatches. In contrast, SAMD-ChIP allows a number of mismatches d that is motif length dependent ($d \leq$ motif length/2), to find similar motif instances. Consequently, SAMD-ChIP is able to find more motif instances than DREME. In case of short and weak motif (short_W), only SAMD-ChIP and TRAWLER were able to predict motif instances inserted in a fraction of 60% and 80% of positive sequences (Figure 4B). However, TRAWLER F-measures are very low because of its high rate of false positive predictions (Table S2). MEME-ChIP fails in predicting short and weak motifs (Figure 4B). These results highlight the efficiency of SAMD-ChIP in predicting either weak or rare motifs.

The ERE-derived palindrome motifs were used to test whether SAMD-ChIP, MEME-ChIP and TRAWLER are able to predict gap-containing motifs. TRAWLER fails in predicting palindrome motifs, while MEME-ChIP and SAMD-ChIP present comparable performances in predicting palindrome motifs inserted in different fraction of input sequences (Figure 4C).

In the second test, we compared the efficiency of SAMD-ChIP, MEME-ChIP and TRAWLER to predict multiple motifs inserted in the same dataset in different sequence fractions. We also performed 10 simulated experiments, each with 100 simulated

sequences, in which we generated subsets of random and not overlapping positions from positive sequences in which we inserted two motifs from the precedent simulation, motif m_1 of 7 bp (strong and short motif) and m_2 *of* 15 bp (strong and long motif). In the first simulation, we inserted instances of motif m_1 in 70% and of motif m_2 in 40% of sequences and allowed motifs co-localisation in 20% of positive sequences. In the second, we inserted m_1 and m_2 instances in 55% and 80% sequences respectively and allowed motif co-localization in 40% of positive sequences. SAMD-ChIP outperformed both MEME-ChIP and TRAWLER (Figure 4C) in this analysis.

TRAWLER F-measures were much lower than those of SAMD-ChIP and MEME-ChIP in different simulations (Figure 4A and 4C), and TRAWLER failed in predicting palindrome and long motifs. This is mainly due to very low precision values (Table S2). TRAWLER results showed the highest Hamming distances for different simulations (Figure 5). This mainly results from its high level of false positive predictions. In contrast, SAMD-ChIP and MEME-ChIP showed comparable distances for long and palindrome motifs. The same result was observed for short motifs except in case of short motif inserted in a fraction of 60% of positive sequences, where the distance observed for MEME-ChIP predictions is higher than SAMD-ChIP and TRAWLER ones (Figure 5 Short-S_60).

In conclusion, SAMD-ChIP performances are generally better or similar compared to MEME-ChIP performances on simulated data. SAMD-ChIP is more accurate to predict rare or weak motifs, and motifs containing gaps. In addition, both tools are superior to TRAWLER in terms of prediction sensitivity and specificity. For this reason, we discarded TRAWLER for subsequent tests on real data sets.

Effect of background sequences on motif discovery tool performances in ChIP datasets

The fundamental assumption of DNA motif discovery tools is that the distribution of motifs bound by the primary TF and those bound by its partners in TF bound regions must differ from their expected distribution elsewhere in the genome. In general, the exceptionality of these motifs in TF bound regions is measured by the level of their fold enrichment in ChIP bound fragments compared to a specific background model. Therefore, it is important to

use an appropriate background model to provide a correct estimation of motif fold enrichment. In case of binding mediated by a single site, TFBSs present uni-modal and symmetric distributions centred on the ChIP peak summits (48). We used this property to estimate TFBS enrichment using different backgrounds in seven TF DNA binding site datasets derived from genome-wide ChIP-chip or ChIP-seq experiments in MCF7 breast cancer cells (Table S1). These data sets represent genomic fragments bound by the estrogen receptor ERα, the retinoic acid receptor RARα, FOXA1, GATA3, Myc, FOS and AP2. For each TF, we used the 100 bp sequences centred on the ChIP peak summits as positive sequences to perform TFBS discovery using SAMD-ChIP.

First, we used ChIP peak proximal flanking regions (100 bp on each side of the central 100 bp) as background and plotted the distribution of motif positions in the 300 bp centred on the ChIP peak summits (Table1 first column).

We observed that the distribution of TFBS positions around the ChIP peak summits present different profiles, with either broad or sharp peaks. For example, the ERα and FOS binding sites (BS) are distributed in broader peaks compared to AP2 and FOXA1 sites (Table1, first column). The shape of TFBSs distribution could be affected by many parameters such as the length of the sonicated fragments, the TF DNA binding properties the antibodies used???, and the use of sequencing vs microarrays to map the bound regions (44,45). In particular, broader peaks may suggest cooperative interactions between several weak TF binding sites as opposed to recruitment mediated by a single strong site. This disparity in the resolution of the peaks indicates that the use of proximal flanking regions can impact apparent enrichments for TFs presenting distributions with broad peaks depending on the size of the central window.

Next, we examined two different backgrounds for the same datasets, i.e. a background formed by distal flanking regions located at 1kb from the ChIP peak summits, and a background created by shuffling the 200 bp sequences around the ChIP peak summits . We observed that motif enrichment folds were enhanced significantly for both sharp and broad peaks (Table 1 column 2 and 3; the number under the plots represent the fold enrichment

intensities). Thus, these two backgrounds allow better detection of motif fold enrichment. For the following analyses, we used shuffled sequences as a background.

Comparison to benchmark DNA motif discovery tools in biological data sets: identification of primary binding sites

First, we assessed the performance of SAMD-ChIP on ERα bound regions. SAMD-ChIP predicted the ERE motif (15bp) as the most enriched matrix in the 100 bp central peak regions compared to background among all 15 bp long matrices.. The background was created by shuffling the 100 bp flanking sequences at each side of the central 100 bp. The IC content of the ERE matrix was improved after the optimization process (Figure 6A), as is more generally the case for all returned matrices (Figure S5), and the fold enrichment of the optimized matrices is higher compared to the initial ones (data not shown). Several matrices representing ERE motifs predicted for 15bp motif length analysis were grouped together in the same cluster and the most informative matrix was selected (Figure 6B). The optimized ERE matrix is significantly enriched in the central 100bp window compared to background (Figure 7A) and this enrichment is maintained for different matrix cut-offs (Figure 7B).

Next, we compared the ERE matrix discovered by SAMD-ChIP to known models, including the ERE matrices from TRANSFAC (19 bp) and JASPAR (20 bp) as well as a 15 bp ERE matrix, NAR_15, previously generated by compiling a set of validated EREs (46). Next, we selected the most enriched SAMD-ChIP ERE matrices for the same motif lengths as each of these reference matrices. We used these matrices to scan an independent ERα ChIP-seq data set (50). Figure 8A shows the number of sites identified by the different matrices at 80% matrix cut-offs and the matrix logos generated using all ERE instances. We observed that for different motif lengths, SAMD-ChIP matrices reported more ERE sites compared to known ERE matrices (Figure 8B) and their forld enrichment is higher. We then compared the accuracy of the SAMD-ChIP ERE matrices and known ERE models to identify validated EREs. From a collection of experimentally validated EREs (47), we

extracted a 50 bp sequence centred on each ERE, and scanned these sequences using all ERE matrices by varying the matrix cut-off between 60% and 90%, with 10% increments. We found that SAMD-ChIP matrices provide the best trade-off between Precision and Recall (Figure 8C). These results demonstrate that SAMD-ChIP might enhance significantly the quality of TFBS known models.

Finally, we ran SAMD-ChIP, MEME-ChIP and DREME on each TFBS data set to discover enriched motifs for different motif length (6 bp to 20 bp) using shuffled sequences as background. Each predicted motif is characterized by its fold enrichment in the 100bp central window compared to the background and its group specificity *P-value*. Next, we evaluated the enrichment of each predicted motif for different matrix cut-offs. ERα, FOXA1, AP2 and FOS BS were predicted by both tools as the top enriched motifs in their respective data sets. However, GATA3 BS was predicted only by SAMD-ChIP (Figure 9). Retinoic acid receptor (RARα) forms heterodimers with members of the RXR nuclear receptor family and binds to two direct repeat DNA motifs (RGGTCA) with 1, 2 or 5bp spacers (42,52,53). Only SAMD-ChIP was able to discover RARα BS. Interestingly, SAMD-ChIP predicted different forms of DR motifs (DR5, DR2 and DR0) in RARα bound regions. These findings demonstrate the efficiency of SAMD-ChIP in predicting long and gapped motifs. While the enrichment of other primary TFBSs in their respective data sets are around 30 (Table 1), the enrichment of DR motifs in RARα bound regions is around ~ 4 fold enrichment. We also observed the enrichment of other putative TFBSs like AP1, FOXA1, ERE, SP1 and AP2 motifs in RARα bound regions (data not shown). These results suggest that a fraction of RARα DNA binding is mediated through protein-protein interactions. Interestingly, ERE motifs and AP2 BS were predicted only by SAMD-ChIP in RARα bound regions. We observed that the enrichment of AP1 BS is higher in RARα bound regions compared to its fold enrichment in ERα bound regions (2 times mention precise folds). However, FOXA1 and AP2 BSs showed comparable fold enrichment in the two data sets.

Analysis of the Myc dataset (309 Myc bound regions) reveals the weakness of current motif discovery algorithms to identify weak and rare motifs (Figure 10). Indeed, no motifs were predicted by DREME in Myc bound regions. However, two distinct motif profiles were predicted by MEME-ChIP, identified by STAMP as Ras response element (RREB1) and inhibitor of DNA binding 1 (Id1) binding sites. The RREB1 motif was also predicted by SAMD-ChIP as the top enriched motif for 15bp motif length analysis in Myc bound regions, whereas the Id1 motif was not found by SAMD-ChIP because this motif is not enriched in the 100 bp central window with a threshold higher than 1.5. Despite the small number of Myc bound regions, SAMD-ChIP was able to predict a 10 bp Myc-like binding site with small but significant fold enrichment (1.84) and a significant group specificity *P-value* (8.6e-03) in the 100 bp central window around the ChIP peak summits. We re-analyzed Myc bound regions using a 200 bp central window length and found the same Myc-like matrix (~3.3 fold enrichment) and a total of 68 occurrences. According to Huas and coworkers (39), motif search analysis using the same Myc bound regions and the Myc TRANSFAC matrix reveals only 34 Myc binding sites in the 500bp regions around the Myc ChIP peak summits. This result confirms the presence of putative Myc binding sites in these regions. Thereby, SAMD-ChIP appears more accurate to discover rare motifs compared to MEME-ChIP and DREME. Interestingly other putative TFBS such as EBF1 (PMID: 21735360), Tcfcp2l1 (PMID: 20158869) and Myf (PMID: 1850105) BS were predicted only by SAMD-ChIP in Myc bound regions (Figure 10). Of note, the enrichment of EBF1 motif in Myc bound regions has been previously reported (38).

Comparison of DNA motif discovery tools using Real data sets: Secondary motifs discovery

To compare tool performances in predicting secondary motifs, we compared motifs predicted by SAMD-ChIP, MEME-ChIP and DREME in ChIP-chip/Seq data for AP2, FOXA1, GATA3, RARα and FOS in MCF7 cells treated with estrogen (E2). Several studies have demonstrated the interplay between ERα, RARα, FOXA1 and AP1 in MCF7 breast cancer cells (40, 41, 43). ERα bound fragments have been shown to be enriched in binding sites for FOXA1 and AP1, suggesting recruitment of ERα to DNA via these TFs or

cooperativity for DNA binding. Several motifs identified as FOXA1, AP1, SP1, GATA3 putative BS were discovered by multiple tools in ERα bound regions (Figure 11). Interestingly, only SAMD-ChIP identified RARα putative BS (DR5 and DR1) in ERα bound regions (Table 2). Ross Iness and coworkers (49) showed that in the presence of estrogen, ERα and RARα can co-occupy the same regions in the genome. Thus, the enrichment of DR motifs in ERα bound regions is consistent with these observations while suggesting tethering of ERα by RARα bound to its sites as a contributing mechanism for co-occupancy. These results also highlight the efficiency of our approach in identifying motifs with gaps.

We further examined the potential converse enrichment of EREs, FOXA1, FOS, GATA3, RARα and AP2 binding sites in TFs bound regions. SAMD-ChIP predicted EREs while MEME-ChIP identified only half-EREs in all these data sets. Interestingly, AP1 and FOXA1 BS were predicted in all data sets by both MEME-ChIP and SAMD-ChIP. The enrichment of FOXA1 and GATA3 motifs in ERα bound regions and the prediction of ERE motifs in both FOXA1 and GATA3 bound regions suggest either site co-localization or reciprocal tethering mechanisms leading to chromatin loop formation. In addition, the reciprocal enrichment of ERα and AP1, AP2 and RARα binding sites suggests an even more extensive network of cooperative TF interactions. Several other motifs were predicted in ERα bound regions by both tools, such as FOXA1, AP1, SP1 and EBF1 putative BS. However, AP2 motifs were predicted only by SAMD-ChIP (Figure 11)

Comparison of DNA motif discovery tool performances in predict new motifs

Two new motif profiles were predicted by both SAMD-ChIP and MEME-ChIP in ERα bound regions. The first motif is composed by two C(T/A)G boxes separated by 1 nucleotide spacer and a second motif is composed by the two C(T/A)G boxes separated by 2 nucleotides spacer (Figure 11). The motif with one spacer was also predicted in FOXA1 and RARα bound regions. Similar motifs composed by two C(T/A)G boxes separated by 3

and 4 nucleotides spacer were predicted only by SAMD-ChIP in FOS, ERα and RARα bound regions (Table 2). These motifs are similar to the motifs that compose the Myf binding site, but with different spacings when comparing the fold enrichment of these new motifs in all data sets, we did not observe any bias to a specific form (data not shown). Thus, they could be bound by the same TF without any preference for a specific form or may be targets for different TFs. No known TFs were predicted to bind these motifs by STAMP. Further analysis will be needed to determine whether these represent variant forms of known TFBSs or sites bound by yet uncharacterized TFs.

Several SAMD-ChIP unique motifs are palindromes or composite motifs (Table 2, Table S3 and S4). Many of them were found in different TF data sets. Indeed, SAMD-ChIP predicted a new palindrome motif with three nucleotides spacer in both GATA3 and ERα bound regions (SWGATnnnATCWS). This may represent a dimeric variant of a GATA site. SAMD-ChIP detected enrichment of a new motif composed by a half ERE (RGGTCA) and a C(T/A)G motif in ERα, FOS and GATA3 bound regions respectively (Table 2). This result may suggest binding of ERα monomers with unknown partner(s) binding to C(T/A)G motifs. In addition, using SAMD-ChIP we predicted composite motifs containing both a half ERE (RGGTCA) and a forkhead motif with variable spacers in ERα bound regions (Table 2). This motif is also enriched in RARα bound regions. Many studies have reported the pioneer role of FOXA1 in mediating ERα DNA binding (40, 43) but the interaction mechanisms between these two TFs are unclear. The identification of this composite motif supports the idea of interactions between ERα and FOXA1.

SAMD-ChIP predicts two composite motifs in RARα and ERα bound regions. The first motif is composed by an AP1 BS and a C(T/A)G motif separated by three nucleotides spacer and the second motif contains a half ERE motif and AP1 BS separated by one nucleotide spacer (Table 2). AP1 is a heterodimer formed by two subunits, FOS and JUN proteins. AP1 is known to mediate ERα and RARα DNA binding by tethering (51). The

colocalization of half ERE and AP1 motifs may represent an alternative mode of binding where both interacting TFs are bound to DNA.

In conclusion, analysis of simulated and real data sets demonstrates that SAMD-ChIP outperforms the state-of-the-art *do novo* motif discovery tools in predicting a vast array of single as well as composite motifs. SAMD-ChIP predicts efficiently weak and rare motifs. Systematic use of SAMD-ChIP in ChIP datasets is expected to enhance significantly known TFBS matrices and to lead to discovery of new motifs not found in TRANSFAC and JASPAR data bases. Ultimately, identification of TFs binding these new motifs will greatly enhance our understanding of transcription regulatory networks.

DISCUSSION

In this paper we describe SAMD-ChIP, a new hybrid algorithm for DNA motif discovery adapted to ChIP-chip and ChIP-sequencing data. This approach combines the strengths of both probabilistic and deterministic approaches to identify motifs enriched close to the ChIP peak summits. SAMD-ChIP combines both a stochastic strategy to select ChIP fragments for seed identification and an enumerative strategy to search for motif instances. First, a motif is initialized by selecting a seed from a window centred on a fragment peak summit. This seed represents the first instance within the motif. Next, we search for similar instances according to SAMD-ChIP sequence similarity criterions, using the remaining fragments. The best instance, which could be located at any position on the fragment, is selected. All found instances are compiled in a PWM representing the initial motif. These initial motifs are then optimized using an EM heuristic. To our knowledge, this strategy is specific to SAMD-ChIP.

Depending on the number of input ChIP fragments *N* and the objectives of the analysis (looking for highly enriched motifs or for motifs present in a small fraction of input sequences), SAMD-ChIP enables us to use either all or a fraction of input fragments to create initial models. In case we selected a fraction of *f* fragments, the set of remaining sequences (*N-f*) are used to update initial motifs during the optimization step (motif update). In doing so, we ensure that SAMD-ChIP uses all input fragments and therefore maximize the chance to predict *bona fide* motifs.

Initial motifs are selected according to their fold enrichment in a selected central window around the ChIP peak summits compared to what is expected by random in a background data set. Using this prior information generates good starting models for the next steps and consequently will increase the probability of finding correct and optimal solutions.

During SAMD-ChIP initialization step, similar instances are grouped together to form initial motifs. We used a similarity threshold *d* at least equal to motif length/2 ($d \leq$ motif length/2). Using a threshold which depends on motif length increases the chance to find

more similar instances for each motif. This property is very important for finding weak motifs (motifs with small number of conserved positions) and motifs containing gaps.
In order to evaluate motif exceptionality, SAMD-ChIP implements two statistics. The first one represents the motif fold enrichment, which is the normalized ratio of the number of motif instances in a selected window around the ChIP peak summits compared to the number of motif instances in a selected background. The second statistic is the group specificity *P-value*. This *P-value* represents the probability to reproduce motif distribution profile randomly. Juntao Li and co-workers (58) proposed an approach based on the estimation of motif distribution. They implemented two distribution parameters, Kurtosis and Skewness to evaluate the significance of predicted motifs. However, motif distribution presents different peak shape, sharp and broad. In addition, several TFs tend to interact with others partners to bind DNA (composite motifs). Thus, It becomes difficult to distinguish correctly between different peaks in case the region contains more than one putative TFBS.
To search for multiple motifs, several tools proceed by masking all motif instances before looking for additional motifs (17,18,19,29,34). This could be problematic because the order in which motifs are masked can have a major effect on the final result. Our approach allows redundancy during the initialization process. However, in the next step, redundant PWMs are clustered and the most informative one is selected.

SAMD-ChIP implements a matrix search pipeline to test the enrichment of predicted matrices for different matrix cut-offs. This is an important step since we expect that the enrichment of potential TFBSs should be observed at several matrix cut-offs.
Focusing on real data analysis, we observed that the central window length and the background model are crucial parameters for TFBS discovery in ChIP-chip and ChIP-seq data. For example, using a 100 bp central window length around the ChIP peak summits and the proximal flanking windows as background, we were able to predict the AP2 and FOXA1 BS with very significant fold enrichment (motifs with sharp distributions). However, ERα and FOS BS present broader distribution profiles, underestimating the fold

enrichment (Table 1). Therefore when using flanking windows as background underestimate their fold enrichment, TFBS discovery may be affected by several parameters at different levels (nature of the TFBS and its mode of binding, length of sonicated fragments, type of arrays used, peak identification procedure). We thus suggest using ChIP peak distal flanking regions or random sequences as background, as these are less sensitive to these parameters and lead consistently to higher enrichment folds. To this end, SAMD-ChIP tool enables to select different background models (user defined background, distal background or background created by shuffling input sequences) and different central window lengths.

Although SAMD-ChIP performance compared well with the state-of-the-art motif discovery tools, the some of its assumptions may still limit its performance. For example, we assume that each sequence may contain one or zero instance for each motif. Further development of the SAMD-ChIP tool will examine potential gains in accuracy by allowing identification of zero, one or more motif instances per sequence as implemented in MEME. Recently, Xiaotu et al. presented the POSMO tool for DNA motif discovery in ChIP-chip and ChIP-seq data (59). POSMO is based on an enumerative strategy. They showed that POSMO outperforms established methods such as MEME-ChIP in term of time complexity. However, this approach is designed only for the discovery of short motifs (motif length <10) and performs poorly on weak motif and motifs containing gap. We tested POSMO tool on ERα, RARα and FOXA1 ChIP-seq data sets. POSMO predicts FOXA1 motif in FOXA1 data set but fails in predicting ERE and DR motifs in ERα and RARα data sets respectively (data not shown).

In this study we overcame some of the obstacles in TFBS motif discovery, such as decrease the number of false positives, predicting rare motifs and motifs with gap such motifs bound by nuclear receptors by developing an accurate and fully automated approach to evaluates motif enrichment in the vicinity of TF bound peak regions through successive motif

discovery, clustering, optimization and identification steps. SAMD-ChIP outperforms MEME-ChIP, DREME and TRAWLER at different levels. It predicted a larger number of validated motifs in simulated or experimental datasets, due to more accurate motif predictions and better performance on weak or rare motifs and motifs containing gaps. SAMD-ChIP predicts only motifs that are enriched in the selected central window. Thus, it presents less false positive predictions compared to other tools. In experimental datasets derived from MCF7 cells, SAMD-ChIP identified the reciprocal enrichment in binding sites for ERα and cooperative TFs, in particular identifying an enrichment in ERE binding sites in datasets for FOXA1, FOS, GATA3, RARα and AP2 and additional reciprocal enrichments in some of these TF datasets such GATA3-FOXA1, FOXA1-AP1 and GATA3-AP1 thus shedding additional light on the interplay between ERα and its partners. Interestingly, many new composite motifs were predicted only by our approach, possibly reflecting hetero-dimerization or TF-TF interactions leading to preferred association of their binding sites. Thus SAMD-ChIP tool should prove useful for large-scale data analysis such as ChIP-chip and ChIP-sequencing and facilitate our understanding of TF regulatory networks.

ACKNOWLEDGEMENTS

We thank Prof. F Major for advice on the algorithm and critical readin of this manuscript.

Funding: CIHR operating grant MT13147 to Dr Mader.

Conflict of Interest: none declared.

REFERENCES FORMAT ?

1. Hamming (1950). "Error detecting and error correcting codes." BSLG.
2. Bailey, T. L. and C. Elkan (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." Proc Int Conf Intell Syst Mol Biol 2: 28-36.

3. Thijs, G., M. Lescot, et al. (2001). "A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling." Bioinformatics 17(11751219): 1113-1122.
4. Tompa, M., N. Li, et al. (2005). "Assessing computational tools for the discovery of transcription factor binding sites." Nat Biotechnol 23(1): 137-144.
5. Sandve, G. K. and F. Drablos (2006). "A survey of motif discovery methods in an integrated framework." Biol Direct 1: 11.
6. Stormo, G. D. (2000). "DNA binding sites: representation and discovery." Bioinformatics 16(1): 16-23.
7. Zhang, M. Q. and T. G. Marr (1993). "A weight array method for splicing signal analysis." Comput Appl Biosci 9(8293321): 499-509.
8. Stormo, G. D., T. D. Schneider, et al. (1986). "Quantitative analysis of the relationship between nucleotide sequence and functional activity." Nucleic Acids Res 14(16): 6661-6679.
9. Horton, P. B. and M. Kanehisa (1992). "An assessment of neural network and statistical approaches for prediction of E. coli promoter sites." Nucleic Acids Res 20(1508724): 4331-4338.
10. Mahony, S. and P. V. Benos (2007). "STAMP: a web tool for exploring DNA-binding motif similarities." Nucleic Acids Res 35(17478497): 253-258.
11. Kummerfeld, S. K. and S. A. Teichmann (2006). "DBD: a transcription factor prediction database." Nucleic Acids Res 34(Database issue): D74-81.
12. Sandelin, A., W. Alkema, et al. (2004). "JASPAR: an open-access database for eukaryotic transcription factor binding profiles." Nucleic Acids Res 32(Database issue): D91-94.
13. Matys, V., E. Fricke, et al. (2003). "TRANSFAC: transcriptional regulation, from patterns to profiles." Nucleic Acids Res 31(12520026): 374-378.

14. Liu, X. S., D. L. Brutlag, et al. (2002). "An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments." Nat Biotechnol 20(12101404): 835-839.

15. Liu, X., D. L. Brutlag, et al. (2001). "BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes." Pac Symp Biocomput: 127-138.

16. Lawrence, C. E., S. F. Altschul, et al. (1993). "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignments." Science 262(8211139): 208-214.

17. Roth, F. P., J. D. Hughes, et al. (1998). "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation." Nat Biotechnol 16(9788350): 939-945.

18. Bussemaker, H. J., H. Li, et al. (2000). "Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis." Proc Natl Acad Sci U S A 97(10944202): 10096-10100.

19. Stormo, G. D. and G. W. Hartzell, 3rd (1989). "Identifying protein-binding sites from unaligned DNA fragments." Proc Natl Acad Sci U S A 86(4): 1183-1187.

20. Waterston, R. H., K. Lindblad-Toh, et al. (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature 420(12466850): 520-562.

21. Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature 409(11237011): 860-921.

22. Ren, B., F. Robert, et al. (2000). "Genome-wide location and function of DNA binding proteins." Science 290(11125145): 2306-2309.

23. Iyer, V. R., C. E. Horak, et al. (2001). "Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF." Nature 409(11206552): 533-538.

24. Johnson, D. S., A. Mortazavi, et al. (2007). "Genome-wide mapping of in vivo protein-DNA interactions." Science 316(17540862): 1497-1502.

25. Robertson, G., M. Hirst, et al. (2007). "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing." Nat Methods 4(17558387): 651-657.

26. Laajala, T. D., S. Raghav, et al. (2009). "A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments." BMC Genomics 10(20017957): 618-618.

27. Wilbanks, E. G. and M. T. Facciotti (2010). "Evaluation of algorithm performance in ChIP-seq peak detection." PLoS One 5(7).

28. Zhao, Y. and G. D. Stormo Quantitative analysis demonstrates most transcription factors require only simple models of specificity, Nat Biotechnol. 2011 Jun 7;29(6):480-3. doi: 10.1038/nbt.1893.

29. D'Haeseleer, P. (2006). "How does DNA sequence motif discovery work?" Nat Biotechnol 24(8): 959-961.

30. Ettwiller, L., B. Paten, et al. (2007). "Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation." Nat Methods 4(17589518): 563-565.

31. Thomas-Chollier, M., M. Defrance, et al. (2011). "RSAT 2011: regulatory sequence analysis tools." Nucleic Acids Res 39(Web Server issue): W86-91.

32. Machanick, P. and T. L. Bailey (2011). "MEME-ChIP: motif analysis of large DNA datasets." Bioinformatics 27(12): 1696-1697.

33. Hu, M., J. Yu, et al. (2010). "On the detection and refinement of transcription factor binding sites using ChIP-Seq data." Nucleic Acids Res 38(20056654): 2154-2167.

34. Hu, J., B. Li, et al. (2005). "Limitations and potentials of current motif discovery algorithms." Nucleic Acids Res 33(15): 4899-4913.

35. Das, M. K. and H. K. Dai (2007). "A survey of DNA motif finding algorithms." BMC Bioinformatics 8 Suppl 7: S21.

36. Valouev, A., D. S. Johnson, et al. (2008). "Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data." Nat Methods 5(9): 829-834.

37. Shim, H. and S. Keles (2008). "Integrating quantitative information from ChIP-chip experiments into motif finding." Biostatistics 9(1): 51-65.

38. Zeller, K. I., X. Zhao, et al. (2006). "Global mapping of c-Myc binding sites and target gene networks in human B cells." Proc Natl Acad Sci U S A 103(47): 17834-17839.

39. Hua, S., C. B. Kallen, et al. (2008). "Genomic analysis of estrogen cascade reveals histone variant H2A.Z associated with breast cancer progression." Mol Syst Biol 4(188): 15.

40. Lupien, M., J. Eeckhoute, et al. (2008). "FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription." Cell 132(6): 958-970.

41. Joseph R, O. Y., Huss M, Sun W, Kong SL, Ukil L, Pan YF, Li G, Lim M, Thomsen JS, Ruan Y, Clarke ND, Prabhakar S, Cheung E, Liu ET. (2010). "Integrative model of genomic factors for determining binding site selection by estrogen receptor-α." Mol Syst Biol.

42. Balmer, J. E. and R. Blomhoff (2005). "A robust characterization of retinoic acid response elements based on a comparison of sites in three species." J Steroid Biochem Mol Biol 96(5): 347-354.

43. Kong, S. L., G. Li, et al. (2011). "Cellular reprogramming by the conjoint action of ERalpha, FOXA1, and GATA3 to a ligand-inducible growth state." Mol Syst Biol 7(526): 59.

44. Gilchrist, D. A., D. C. Fargo, et al. (2009). "Using ChIP-chip and ChIP-seq to study the regulation of gene expression: genome-wide localization studies reveal widespread regulation of transcription elongation." Methods 48(4): 398-408.

45. Buck, M. J. and J. D. Lieb (2004). "ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments." Genomics 83(3): 349-360.

46. Bourdeau, V., J. Deschenes, et al. (2008). "Mechanisms of primary and secondary estrogen target gene regulation in breast cancer cells." Nucleic Acids Res 36(17986456): 76-93.

47. Vega, V. B., C.-Y. Lin, et al. (2006). "Multiplatform genome-wide identification and modeling of functional human estrogen receptor binding sites." Genome Biol 7(16961928).

48. Barrett, C. L., B. K. Cho, et al. (2011). "Sensitive and accurate identification of protein-DNA binding events in ChIP-chip assays using higher order derivative analysis." Nucleic Acids Res 39(5): 1656-1665.

49. Ross-Innes, C. S., R. Stark, et al. (2010). "Cooperative interaction between retinoic acid receptor-alpha and estrogen receptor in breast cancer." Genes Dev 24(2): 171-182.

50. Welboren, W. J., M. A. van Driel, et al. (2009). "ChIP-Seq of ERalpha and RNA polymerase II defines genes differentially responding to ligands." Embo J 28(10): 1418-1428.

51. Petz, L. N., Y. S. Ziegler, et al. (2002). "Estrogen receptor alpha and activating protein-1 mediate estrogen responsiveness of the progesterone receptor gene in MCF-7 breast cancer cells." Endocrinology 143(12): 4583-4591.

52. Durand, B., M. Saunders, et al. (1992). "All-trans and 9-cis retinoic acid induction of CRABPII transcription is mediated by RAR-RXR heterodimers bound to DR1 and DR2 repeated motifs." Cell 71(1): 73-85.

53. Leid, M., P. Kastner, et al. (1992). "Purification, cloning, and RXR identity of the HeLa cell factor with which RAR or TR heterodimerizes to bind target sequences efficiently." Cell 68(1310259): 377-395.

54. Wu, S., J. Wang, et al. (2010). "ChIP-PaM: an algorithm to identify protein-DNA interaction using ChIP-Seq data." Theor Biol Med Model 7: 18.

55. Brown, C. D., D. S. Johnson, et al. (2007). "Functional architecture and evolution of transcriptional elements that drive gene coexpression." Science 317(5844): 1557-1560.

56. Frith, M. C., M. C. Li, et al. (2003). "Cluster-Buster: Finding dense clusters of motifs in DNA sequences." Nucleic Acids Res 31(13): 3666-3668.

57. Zhou, Q. and W. H. Wong (2004). "CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling." Proc Natl Acad Sci U S A 101(33): 12114-12119.

58. Li, J., L. Zhu, et al. (2011). "Deciphering transcription factor binding patterns from genome-wide high density ChIP-chip tiling array data." BMC Proc 5 Suppl 2: S8.

59. Ma, X., A. Kulkarni, et al. (2012). "A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information." Nucleic Acids Res 40(7): 6.

Figure 1. SAMD-ChIP pipeline.

Figure 2. Motif optimization process. PWM0 represents the initial matrix.

PWM_r and PWM_u represent the refined and the updated matrices respectively.

Figure 3. Simulated motifs. A and B represent weak and strong short motifs of length 7bp. C and D represent weak and strong long motifs of length 15bp . E represents a palindrome motif of 15bp length. Motif logos were generated using R package "seqLogo".

Figure 4. SAMD-ChIP, MEME-ChIP and TRAWLER F-measure averaged over the 100 simulations.

(A) Long strong (Long-S) and Short strong (Short-S) motifs. (B) Long weak motif (Long-W) and Short weak (Short-W) motifs. (C) Palindrome (Pal) and two motifs simulations. All motifs are inserted in different fractions of input sequences (10%, 30%, 60% and 80%).

Figure 5. Hamming distance between implemented and predicted motifs using MEME-ChIP, SAMD-ChIP and TRAWLER.

The y-axis represents the distance between the predicted and the inserted motifs for different sequence fractions (30%, 60% and 80%). The x axis represents motif categories: Long-S (long and strong motifs); Long-W (long and weak motifs); Pal (palindrome motifs); Short-S (short and strong motifs).

Figure 6. ERE matrix predicted in ERα bound regions.

(A) ERE matrix before and after the refinement process. (B) Clustering of predicted ERE similar matrices predicted in ERα bound regions using shuffle sequences as background

Figure 7. ERE matrix distribution in ERα bound regions.

(A) Distribution of ERE matrix identified by SAMD-ChIP in 300bp sequences centred on ERα ChIP-peak summit. ERE fold enrichment is measured as the ratio of the number of ERE occurrences in the central 100bp and the number of ERE occurrences in background. GP pvalue represents the ERE group specificity *P-value* (B) SAMD-ChIP ERE matrix fold enrichment for different matrix cut-offs (40%-90%). The table represents the number of matrix occurrences for each matrix cut-off. Fisher and Khi2 p-values test for the significance of ERE distribution in the central the 100 bp compared to background.

Figure 8. Comparison of ERE matrices predicted by SAMD-ChIP versus known ERE matrices.

(A) ERE matrices logos generated by compiling instances found using the various ERE matrices at 80% matrix cut-off. (B) Distribution of ERE sites and fold enrichment reported for 80% matrix cut-off using different ERE matrix profiles (C) comparison of the performances of ERE matrix predicted by SAMD-ChIP to known ERE models in predicting validated EREs using matrix cut-offs of 60, 70, 80 and 90%. We used a 50 bp sequences flanking a set of validated EREs (Vega et al, 2006) as positive sequences and shuffle background as negative sequences.

Figure 9. SAMD-ChIP and MEME-ChIP predictions in ERα, AP2, FOXA1, FOS, GATA3, and RARα bound regions.

NA means no predictions were found by the corresponding tool.

Figure 10. SAMD-ChIP and MEME-ChIP predictions in Myc bound regions.

TFBS matches in TRANSFAC/JASPAR are given for each prediction. NA means no match was found in TF databases.

Figure 11. Putative TFBSs predicted in ERα bound regions by SAMD-ChIP and MEME-ChIP.

Table 1. Effect of background models on motif enrichment.

Distribution of EREs and FOXA1, AP2 and AP1 binding sites in the 100 bp central window in their respective bound regions using different backgrounds (proximal , distal and shuffle background)

Table 2. SAMD-ChIP unique motifs predicted in different data sets.

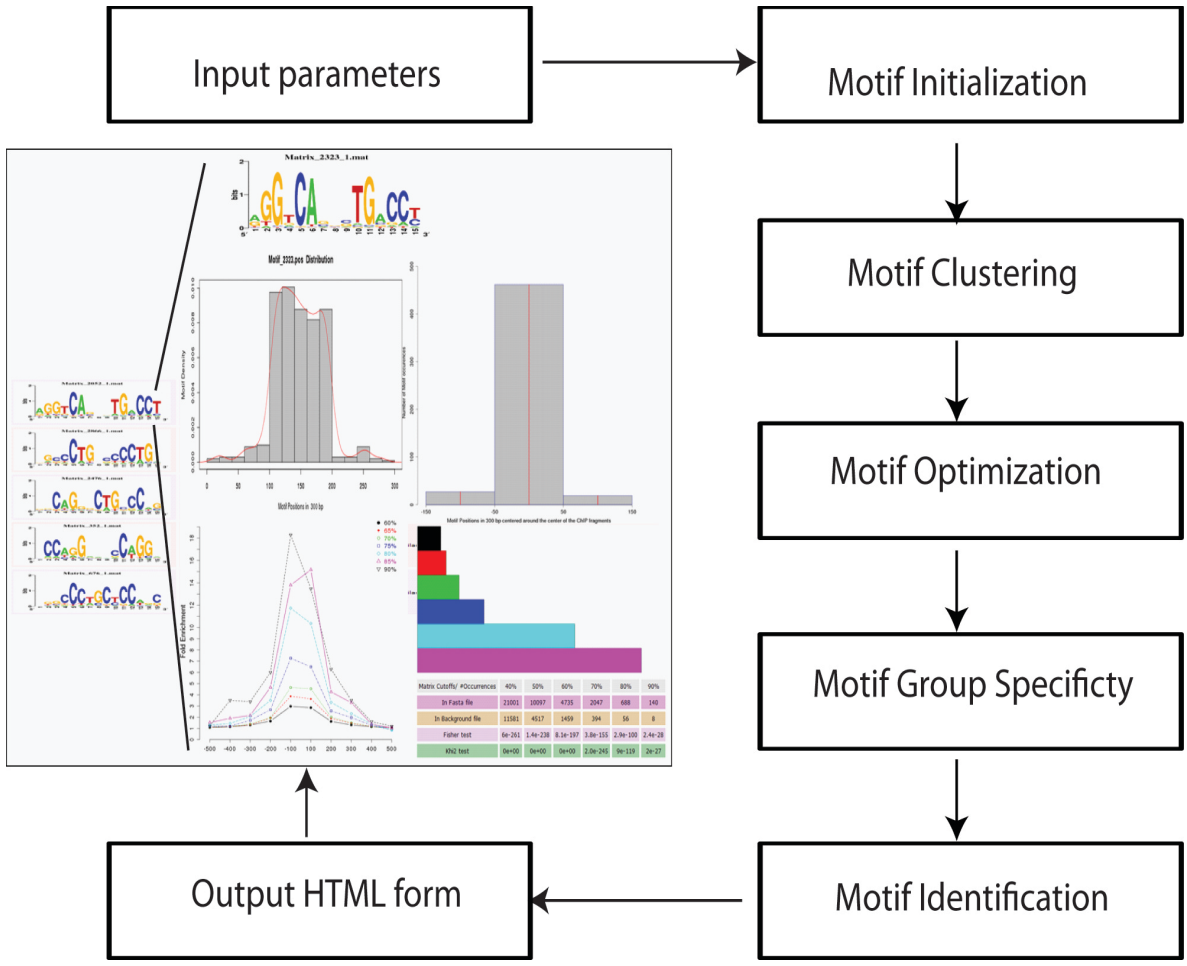


Figure 1. SAMD-ChIP pipeline.

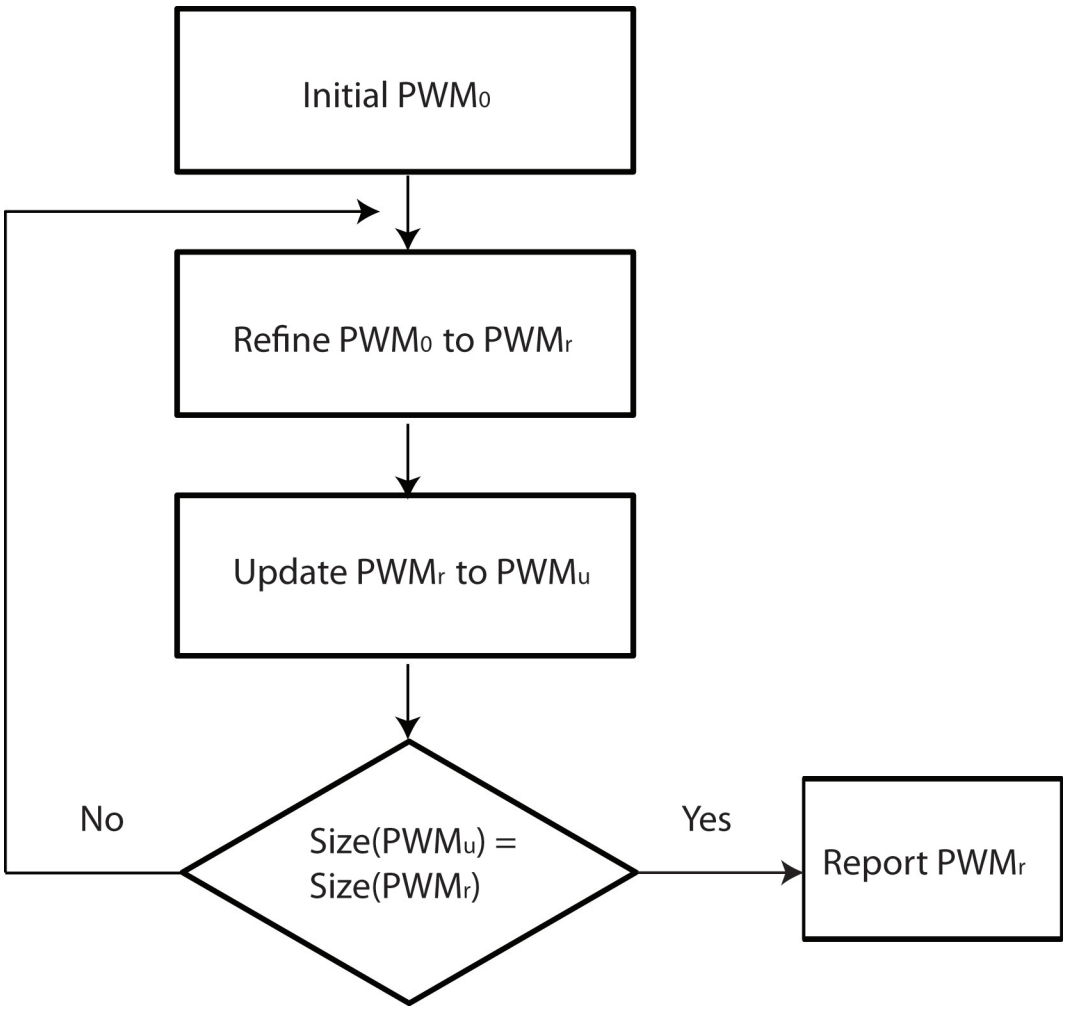


Figure 2. Motif optimization process. PWM0 represents the initial matrix.

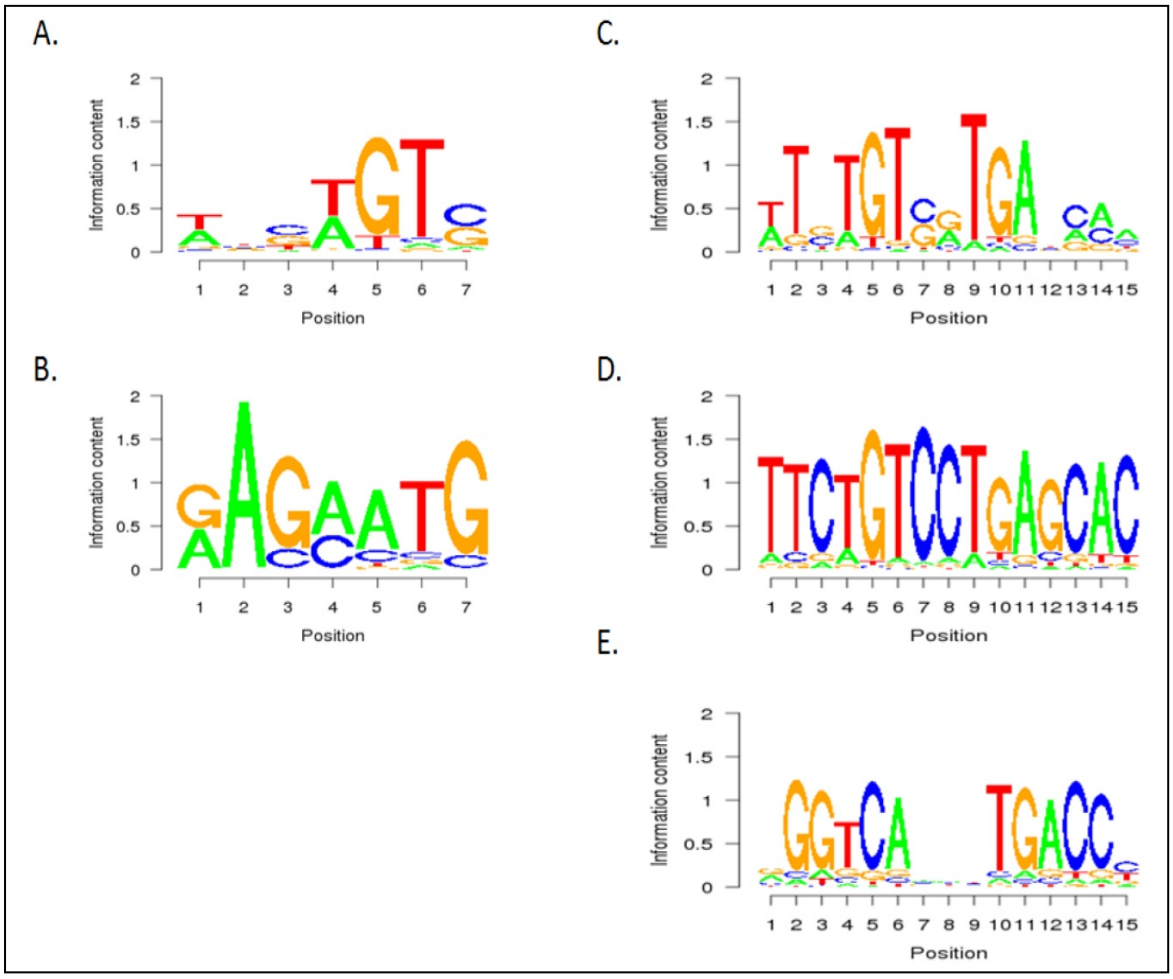


Figure 3. List of simulated motifs.

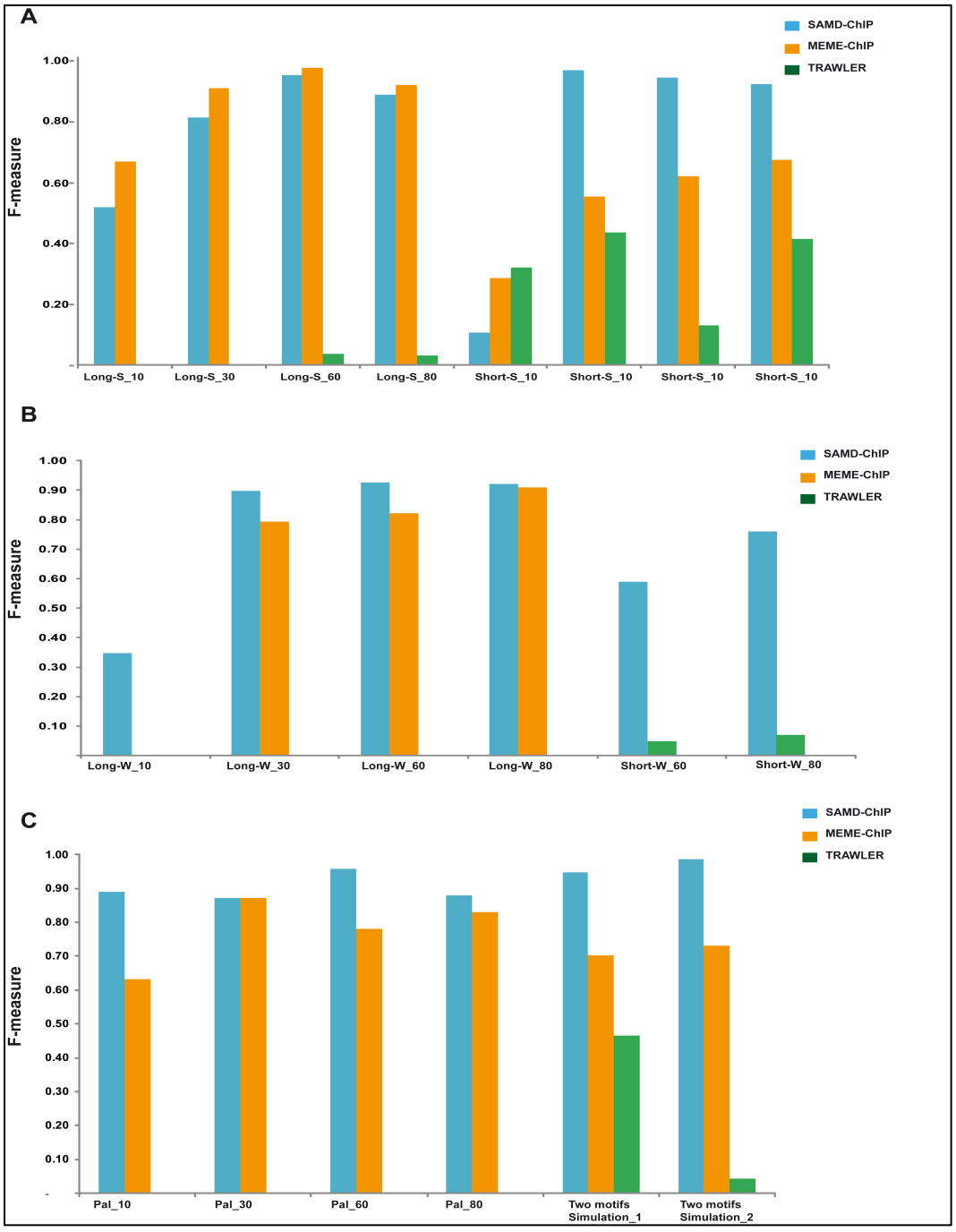


Figure 4. SAMD-ChIP, MEME-ChIP and TRAWLER F-measure averaged over the 100 simulations.

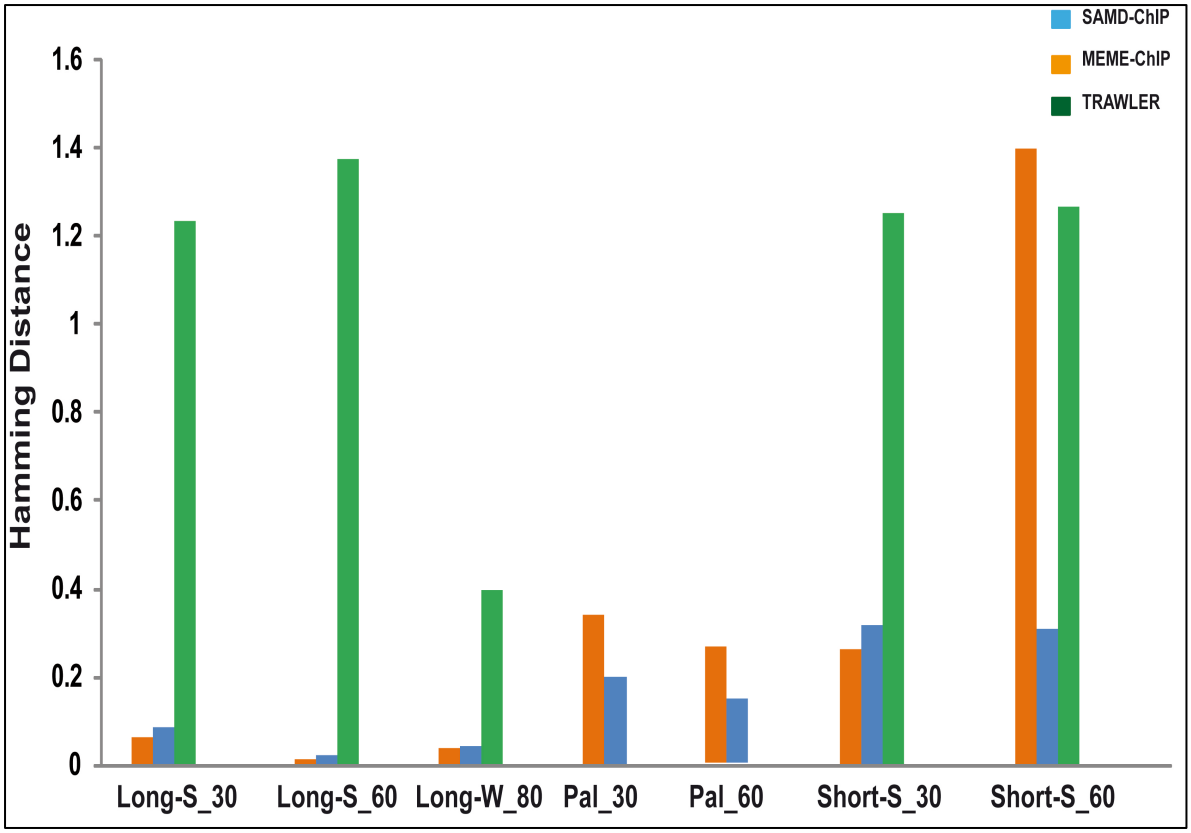


Figure 5. Hamming distance between implemented and predicted motifs using MEME-ChIP, SAMD-ChIP and TRAWLER.

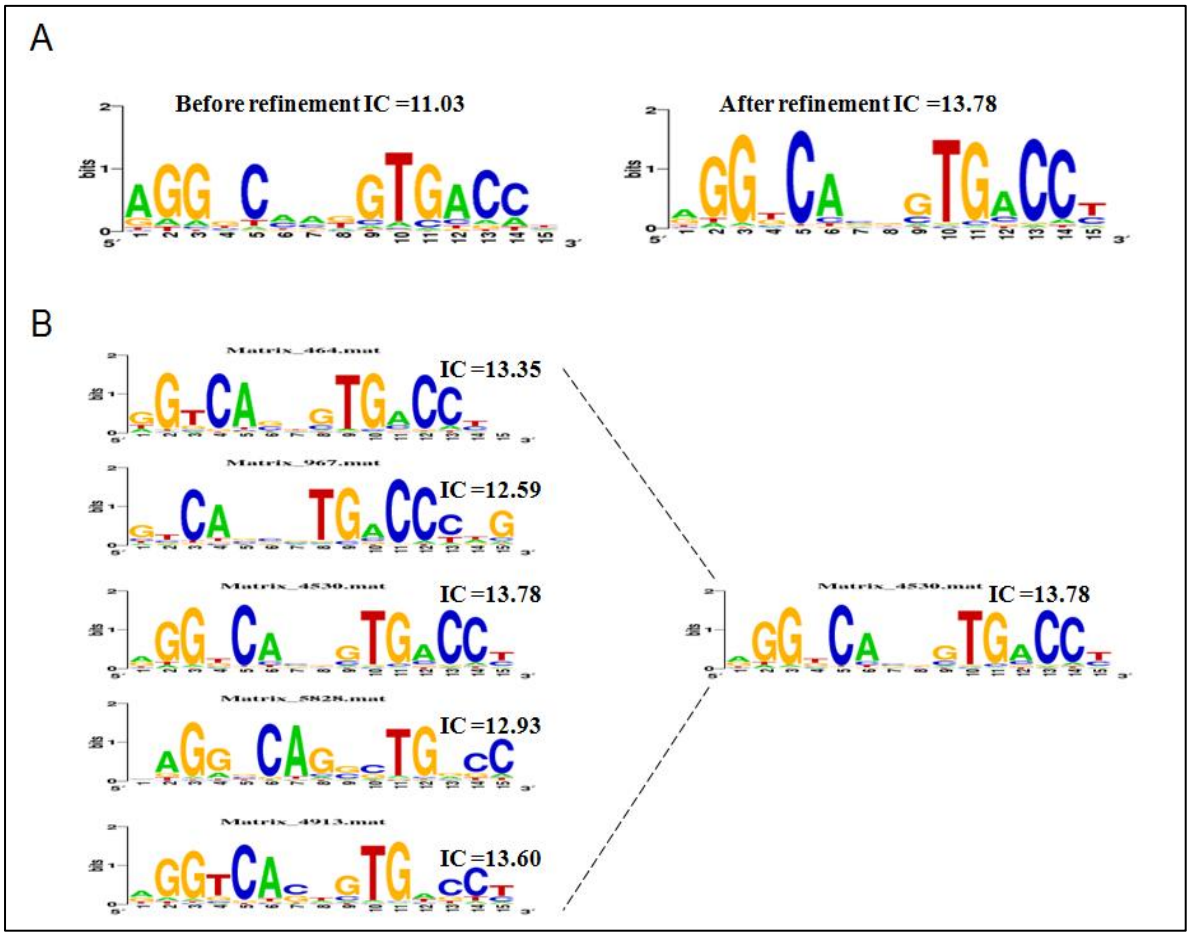


Figure 6. ERE matrix optimization.

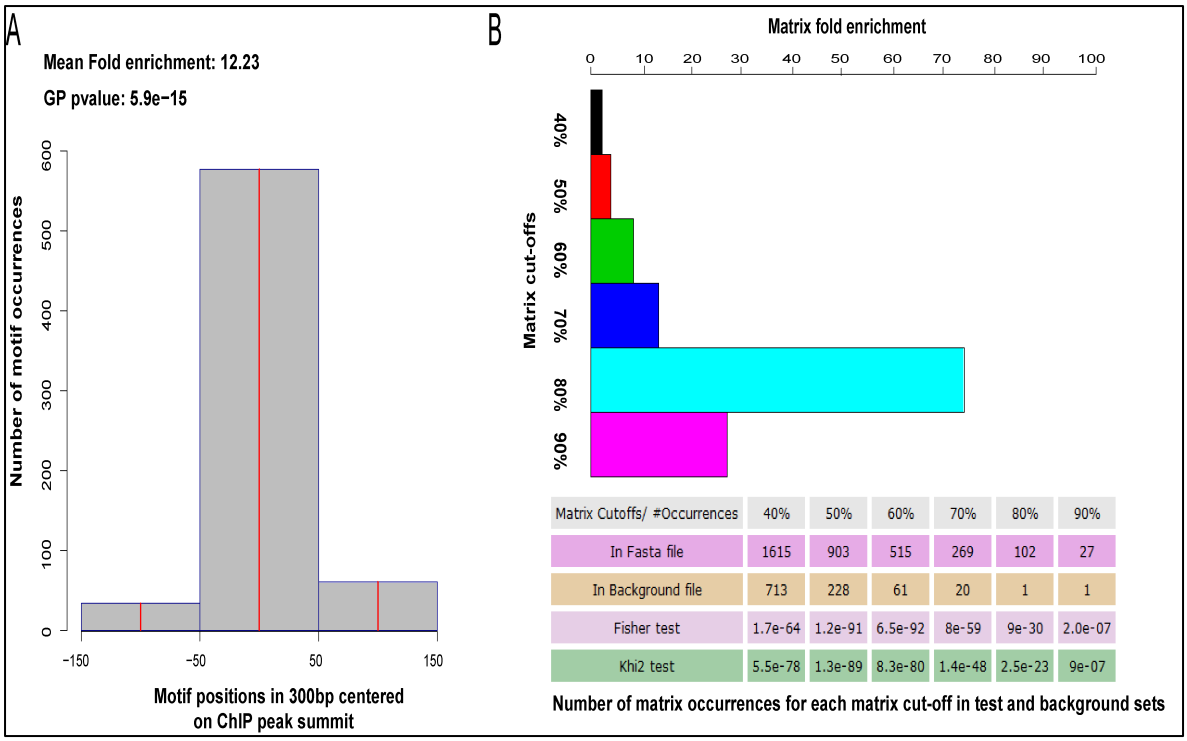


Figure 7. ERE matrix distribution in ERα bound regions.

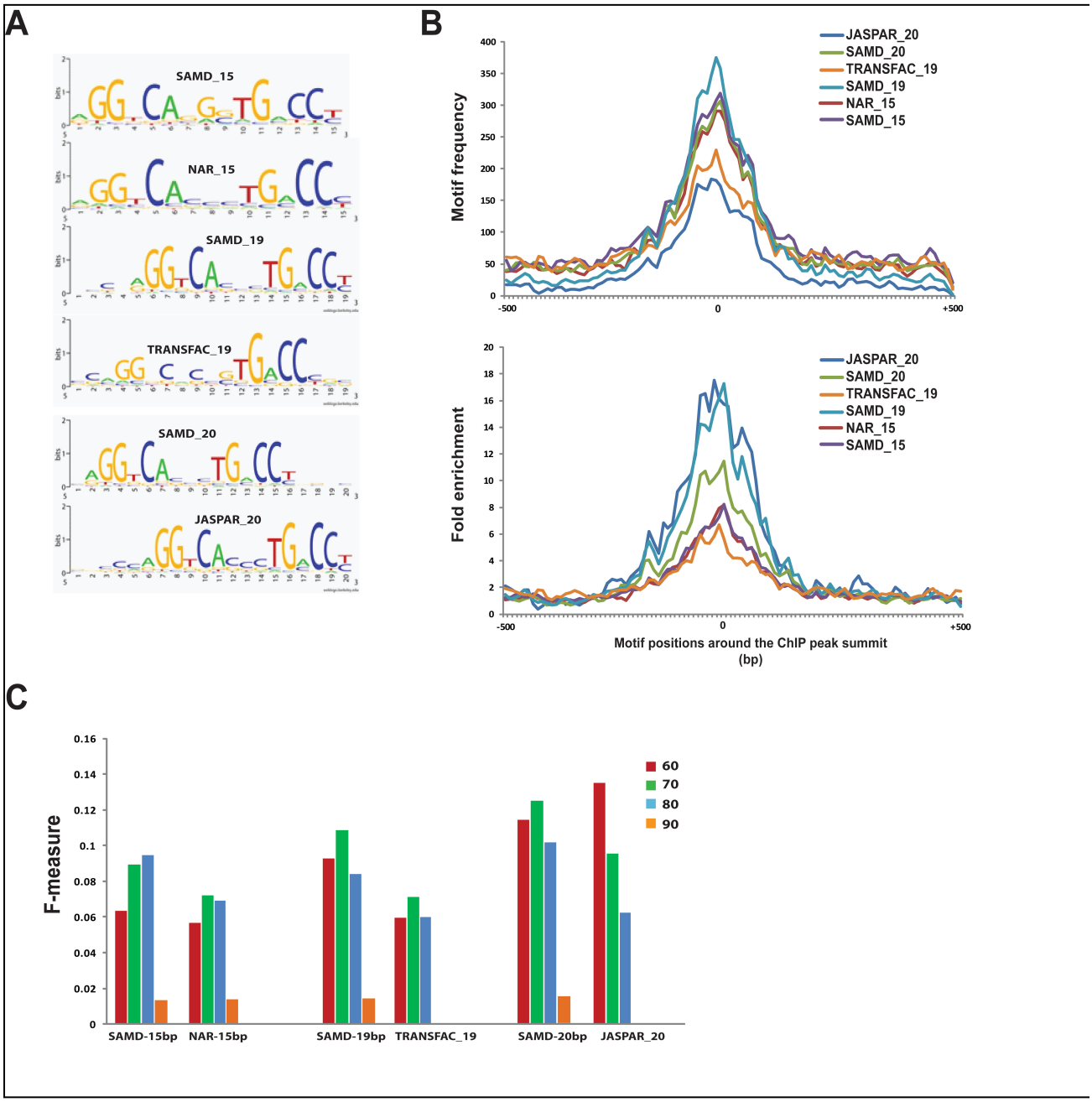


Figure 8. Comparison of ERE matrices predicted by SAMD-ChIP versus known ERE matrices.

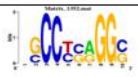
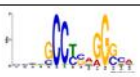
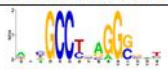
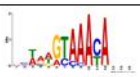
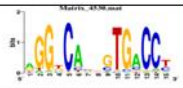
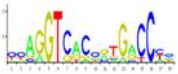
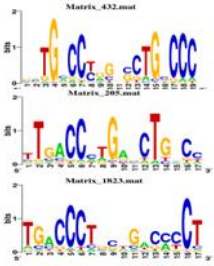
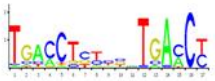
TF name	SAMD-ChIP	MEME	DREME	TF logo in TRANSFAC/JASPAR
AP2 (10bp)				
FOXA1 (9bp)				
FOS (7bp)				
GATA (7bp)		NA		
ERα (15bp)			NA	
RARα DR5, DR0 and DR2 motifs		NA	NA	

Figure 9. SAMD-ChIP and MEME-ChIP predictions in ERα, AP2, FOXA1, FOS, GATA3, and RARα bound regions.

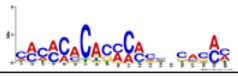
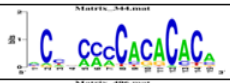
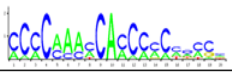
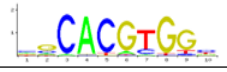
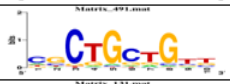
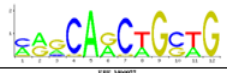
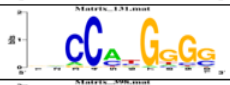
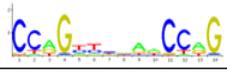
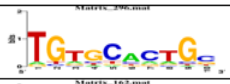
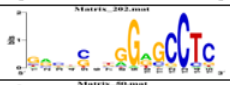
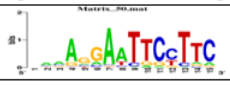
MEME-ChIP	SAMD-ChIP	Matches in TRANSFAC/JASPAR	TFs logo in TRANSFAC/JASPAR
		MA0073.1_RREB1	
		MA0147.1_Myc	
		MA0055.1_Myf	
		MA0154.1_EBF1	
		MA0145.1_Tcfcp2l1	
		NA	NA
		NA	NA
		NA	NA
		NA	NA
		NA	NA

Figure 10. SAMD-ChIP and MEME-ChIP predictions in Myc bound regions.

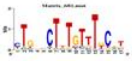
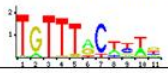
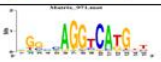
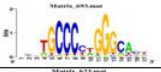
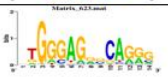
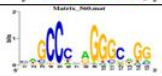
DREME	MEME	SAMD-ChIP	Matches in TRANSFAC/JASPAR	TFs logo in TRANSFAC/JASPAR
			MA0148.1 _FOXA1	
			Half ERE	
			MA0099.2_AP1	
	-		MA0154.1_EBF1	
	-		MA0079.2_SP1	
			MA0079.2_SP1	
			MA0055.1_Myf	
			NA	NA
			NA	NA
			M00189_AP2	

Figure 11. Putative TFBSs predicted in ERα bound regions by SAMD-ChIP and MEME-ChIP.

Table 1. Effect of background model on the observed enrichment.

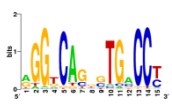
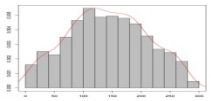
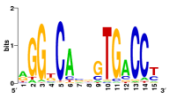
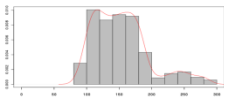
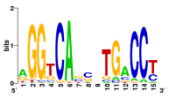
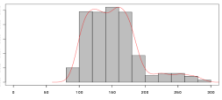
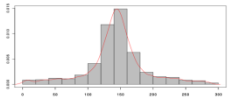
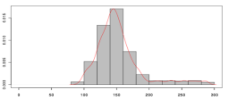
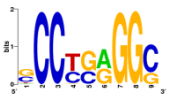
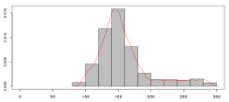
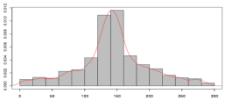
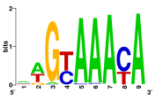
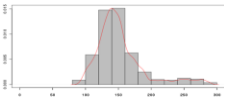
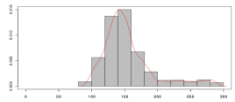
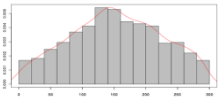
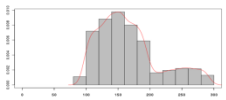
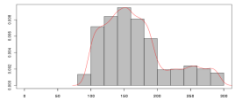
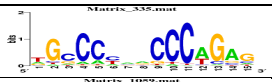
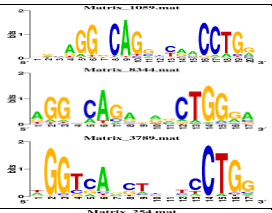
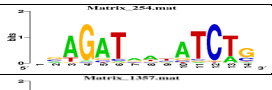
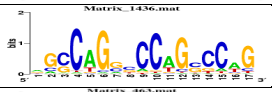
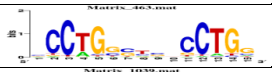
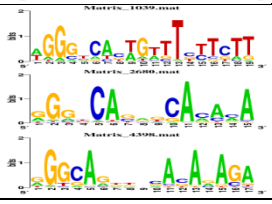
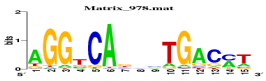
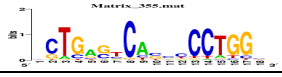
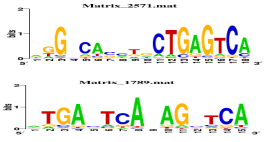
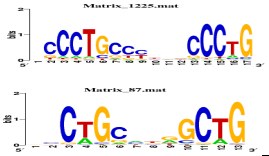
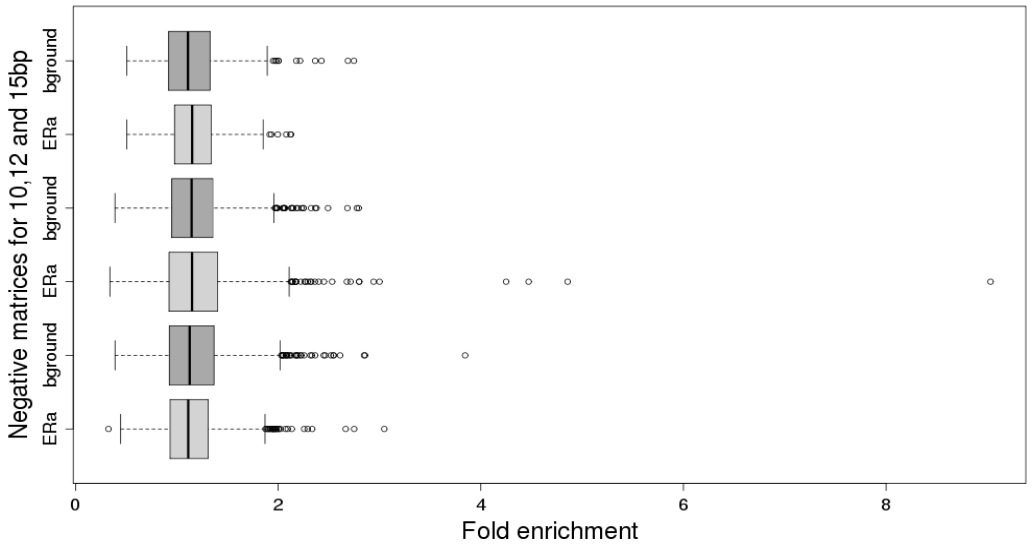
Proximal background		Shuffle background		Distal background	
Logo	Distribution	Logo	Distribution	Logo	Distribution
	 ~2		 ~13		 ~18
	 ~9		 ~46		 ~41
	 ~6		 ~29		 ~35
	 ~2		 ~27		 ~29

Table 2. SAMD-ChIP unique motifs predicted in different data sets

Motif logo	Predicted in
	GATA3, ERα, FOS (half ERE, Ct/aG box)
	ERα (Half ERE, Ct/aG box)
	GATA3, ERα (unknown motif)
	FOS (IR5 motif)
	ERα (unknown motif)
	ERα (unknown motif)
	ERα (Half ERE, Fkh motif)
	ERα (DR5)
	ERα (DR1)
	RARα (ERE)
	ERα, RARα (AP1 motif, Ct/gG motif)
	ERα, RARα (AP1 motif, Half ERE)
	RARα (unknown motifs)

A



B

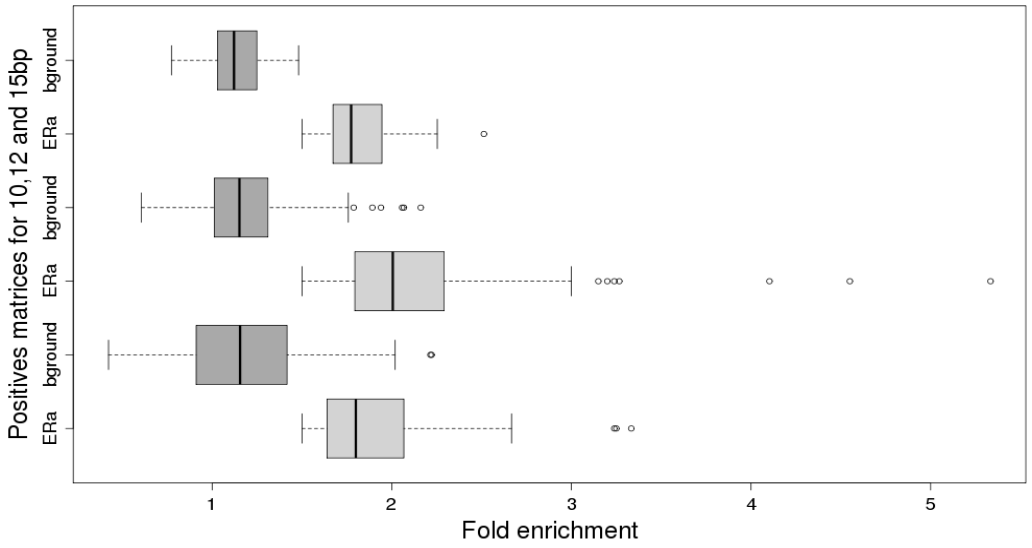


Figure S1

Fold enrichment distribution in input and background data sets. Fold enrichment for positive (fold ≥1.5) and negative (fold <1.5) matrices extracted from ERα bound regions in MCF7 cells. (A) Fold enrichment box plots for not enriched matrices (negatives) in ERα bound regions and in background respectively (B) fold enrichment box plots for enriched matrices (positives) in ERα bound regions and in background respectively.

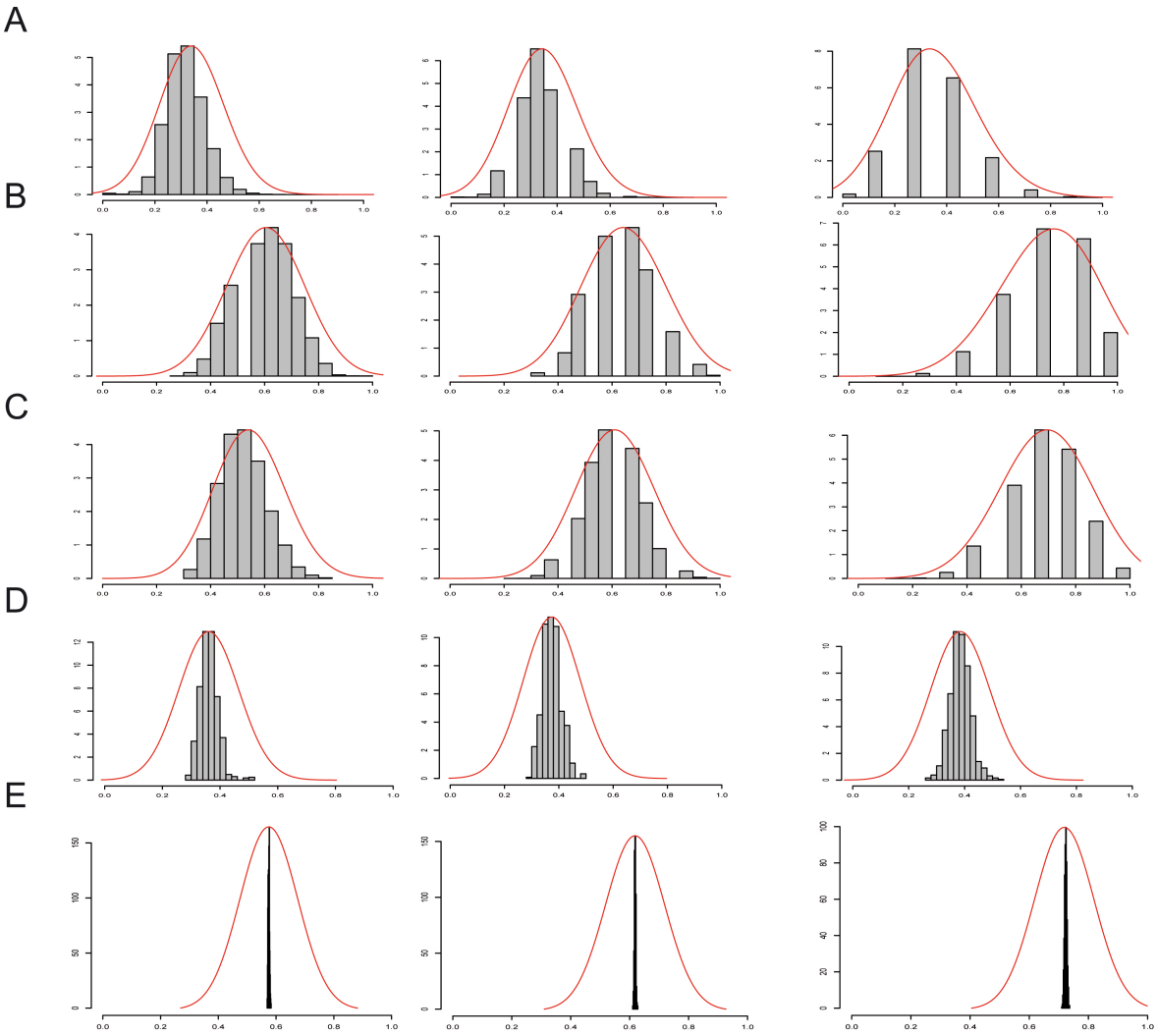


Figure S2

Matrix similarity threshold analysis. (A) Similarity distribution between random sequences for different length (20, 15, 9bp). (B) Similarity distribution within a set of JASPAR matrices for different length (20, 15, 9bp). (C) Similarity distribution within a set of SAMD-ChIP matrices for different length (20, 15, 9bp). (D) Similarity distribution between a set of distinct SAMD-ChIP matrices (matrices representing different motif profiles) for different motif length (20, 15, 9bp). (E) Similarity distribution between a set of similar SAMD-ChIP matrices (matrices representing same motif profiles) for different motif length (20, 15, 9bp).

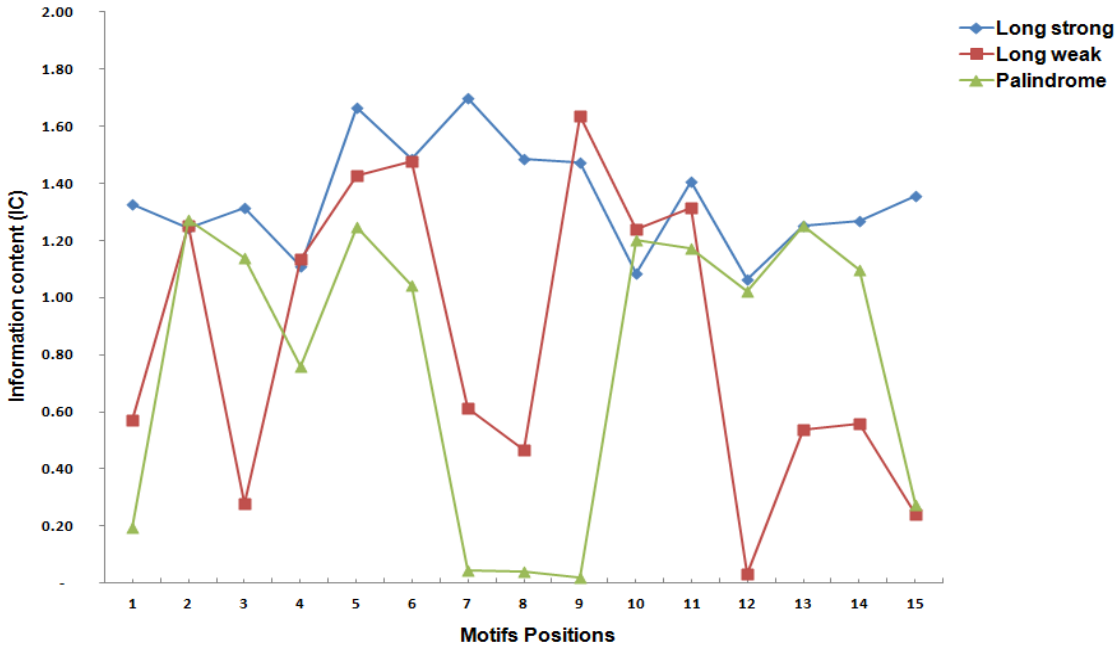


Figure S3

Motifs information content (IC) distribution for long motif (strong and weak) and palindrome motif used in simulation analysis.

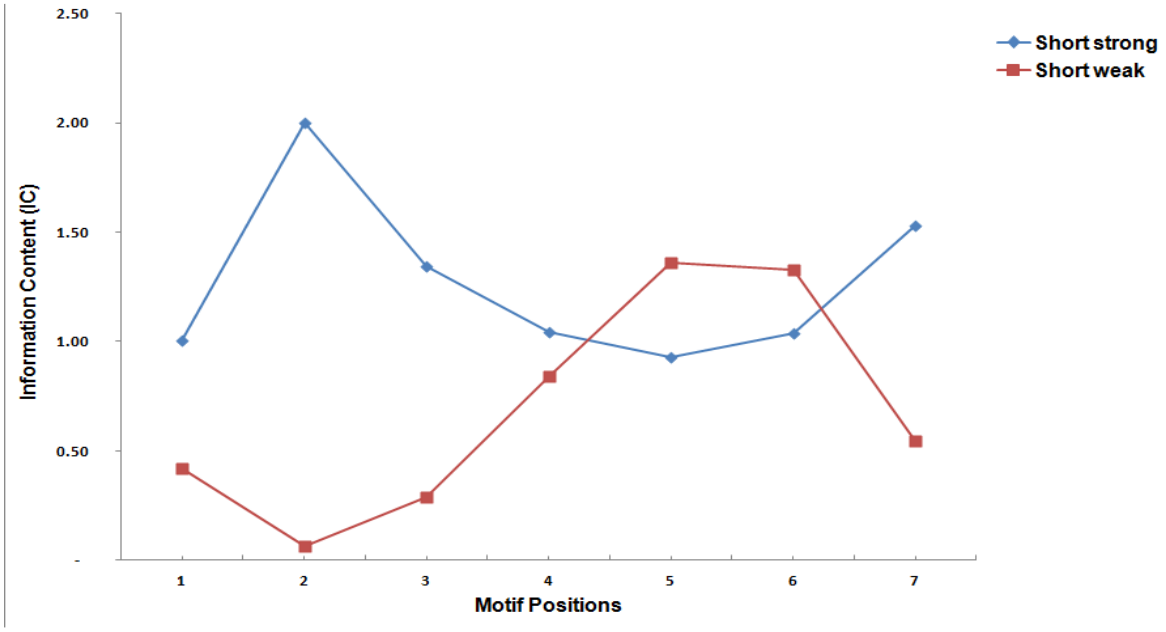


Figure S4

Motifs information content (IC) distribution for short motif (strong and weak) used in simulation analysis.

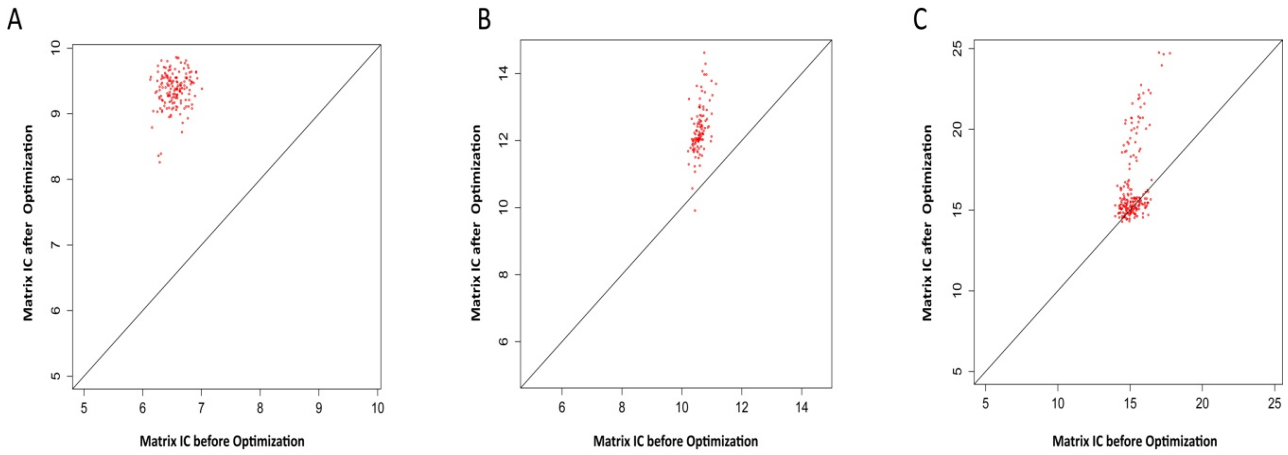


Figure S5

Matrix IC before and after optimization process for different motif lengths. (A) 8bp (B) 15bp (C) 20bp.

Table S1. Biological data sets description

TF name	# Regions	E2-treated	Experiment
ERα (Carroll et al., Nat Genet. 2006; Ross Iness et al., Genes Dev. 2010)	3134	45mn/1h	Common regions between ChIP-chip and ChIP-seq
RARα (Ross Iness et al., Genes Dev. 2010)	Top 4000	45mn	ChIP-seq
FOS (Josheph R et al., Mol Syst Biol. 2010)	Top 4000	3h	ChIP-seq
FOXA11/GATA3/AP2 (Tan et al., EMBO J. 2011)	Top 4000	45mn	ChIP-seq
MYC (Huas et al; Mol Syst Biol. 2008)	309	2H	Chip-chip

Top 4000: for the motif discovery analysis, we used the 4000 most enriched fragments.

Table S2. Benchmark results of simulated data

Simulation	SAMD-ChIP		MEME-ChIP		TRAWLER	
	Ret	Rel	Ret	Rel	Ret	Rel
Pal-10	8	8	9	6		
Pal-30	32	27	25	24		
Pal-60	61	58	40	39		
Pal-80	79	70	74	64	191	60
Two-motifs-m1(15bp)	61	55	53	38	18	17
Two-motifs-m2(7bp)	82	80	57	50	62	3
Long-w-10	15	4	12	0		
Long-w-30	28	26	23	21		
Long-w-60	61	56	47	44	196*	57
Long-w-80	85	76	74	70	187*	53
Long-s-10	17	7	11	7		
Long-s-30	44	30	25	25	71	0
Long-s-60	60	57	57	57	52	2
Long-s-80	78	70	68	68	50	2
Short-s-10	9	1	11	3	15	4
Short-s-30	32	30	17	13	85	25
Short-s-60	65	59	40	31	188*	16
Short-s-80	87	77	60	47	195*	57
Short-w-10						
Short-w-30					30	1
Short-w-60	38	29			24	2
Short-w-80	67	56			42	4

Ret (Retrieved): represents the total number of predicted instances for each simulation.

Rel (Relevant): Number of correct predictions.

Implemented: the number of implemented instances for each simulation

w for weak , s for strong and pal for palindrome

All these predictions are averaged over 100 simulated studies.

*Most of TRAWLER predictions are shifted by more than one position.

Table S3. Composite motifs predicted in ERα bound regions

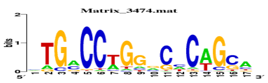
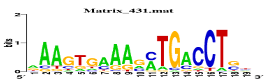
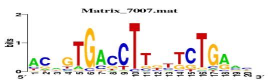
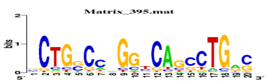
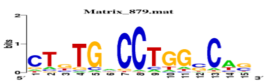
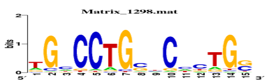
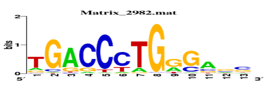
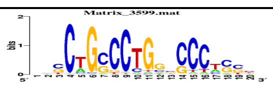
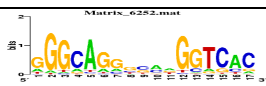
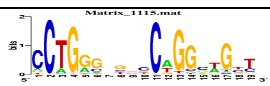
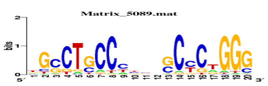
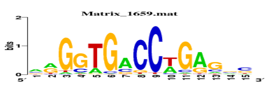
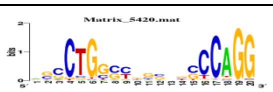
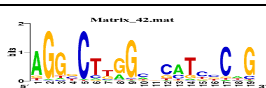
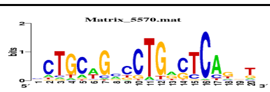
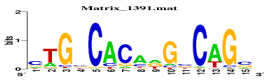
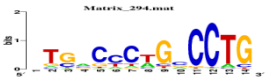
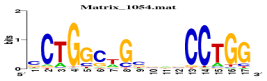
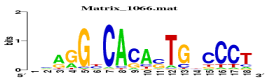
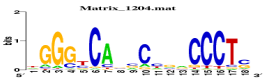
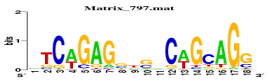
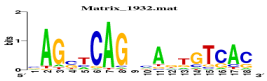
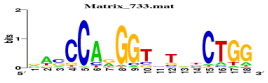
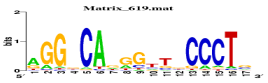
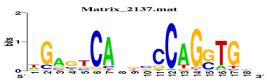
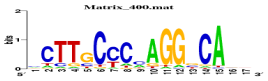
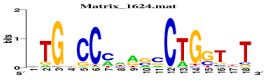
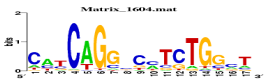
		
		
		
		
		
		

Table S4. Composite motifs predicted RARα bound regions

Chapitre 7

Inférence des interactions entre les facteurs de transcription à partir de la co-localisation de leur sites de fixation à l'ADN

Dans le second manuscrit, nous avons étendu notre analyse à la découverte et la prédiction des interactions entre facteurs de transcription. Nous avons développé un module qui prédit les interactions entre FTs en se basant sur la co-localisation de leurs sites de fixation à l'ADN à partir des données de ChIP-chip et de ChIP-seq.

Par définition, un motif composite est un motif composé de deux blocs (sites) séparés par un certain nombre de positions. Chaque bloc correspond à un site lié par un FT donné. La présence d'un site composite suggère une interaction entre deux FTs. Cette interaction peut être soit directe ou indirecte.

Dans le cas des sites composites, l'espacement entre les deux sites est variable, allant d'une position jusqu'à une centaine de paires de bases. Par conséquent, tenter de prédire ces sites est une tache très ardue à cause de la taille et la composition de l'espacement. Notre approche de prédiction des éléments composites se fait en deux étapes. Dans la première étape, on applique notre outil de découverte de motifs SAMD-ChIP pour prédire les motifs enrichis à proximité des régions de ChIP-chip/seq pour certaines longueurs de motifs (par exemple de 6-20 bp). Ensuite, on estime la probabilité de co-localisation des paires de motifs prédits à partir de leurs distributions dans une fenêtre autour du centre des régions liées. Ceci est réalisé en mesurant deux paramètres : la distance entre les deux sites et la taille de l'espacement entre eux. Il est à noter que cette approche peut être utilisée avec n'importe quel autre outil de découverte de motifs qui utilise les PWMs pour modéliser les

motifs prédits. Elle peut s'appliquer aux modèles de TFBSs existants dans les banques de TFs comme TRANSFAC et JAPSAR.

Inference of transcription factor cooperativity from ChIP-chip and ChIP-sequencing data

M. AID[1,2], D. LAPERRIÈRE[1,2], S. MADER[1,2]
[1]Biochemistry department, Montreal University, Quebec, Canada,
[2]IRIC, Institute of Research in Immunology and Cancer, Montreal University, Montreal, Quebec, Canada

ABSTRACT

Results from genome-wide chromatin immunoprecipitation experiments have revealed that DNA binding sites of different transcription factors (TFs) are enriched to variable levels in the same datasets, suggesting that TF-TF interactions are critical for transcriptional regulation. Secondary TF binding sites may be enriched due to their role as tethering factors recruiting the primary TF on DNA. On the other hand, they may bind DNA and cooperate with the primary TF, either through the formation of chromatin loops or due to TF-TF interactions on composite sites. We have developed an approach to investigate further the nature of interactions between TFs enriched in the same datasets and the presence of composite motifs. This approach uses TFBS matrices predicted to be enriched in the center of ChIP-chip and ChIP-sequencing regions by TFBS motif search or discovery tools and first assesses co-localization in a subset of ChIP fragments and the distribution of spacing lengths between co-occuring motifs. Finally, enrichment of the composite motif in the central window of ChIP fragments is assessed within the entire dataset and compared to those of the individual single motifs to assess cooperativity between the two factors. We validated this approach on binding sites enriched in datasets of ERα, RARα, FOXA1, AP1,

AP2 and GATA3 obtained in MCF7 breast cancer cells. We observed that EREs are highly enriched in several datasets in addition to those generated with antibodies against ERα, including FOXA1 and RARα. On the other hand, binding sites for FOXA1 and RARα are less enriched in ERα datasets than in their own, suggestive of a greater capacity of ERα to act as tethering factor. ERE sequences enriched in different datasets were similar, although biases for some nucleotides were observed in different datasets, suggesting that the primary sequence of the response elements may affect selection of cooperating TFs. While co-localization of EREs, EREs and AP1 binding site, EREs and FOXA1 binding site was not observed, half-EREs were as expected co-enriched preferentially as palindromes with 3 bp spacing both in ERα, but also in RARα bound regions. Finally, while AP1 and half-EREs were co-enriched in 1 kb ChIP regions in the ERα and RARα datasets with a preferred spacing of 46 and 48 base pairs, they were not enriched in the central ChIP window, reflecting their presence in Alu repeats located close to but outside of the main peak. This new module can be used in combination with any TFBS discovery tool including our recently described SAMD-ChIP pipeline, and provides a practical approach to improve our understanding of TF-TF interactions.

INTRODUCTION

Recent developments in functional genomic approaches (1-3), in particular chromatin immuno-precipitation coupled to DNA microarrays (ChIP-chip) or DNA high-throughput sequencing (ChIP-seq), have provided high resolution genome-wide maps of *in vivo* protein-DNA interactions. These experiments identify genomic regions immuno-precipitated in the presence of an antibody specific for a given transcription factor (TF). Matrix search algorithms can then be applied to identify known TF binding sites (TFBS) in those regions (4,5). In addition, DNA motif discovery approaches can define matrices of enriched sites without prior knowledge (6). These approaches generally yield a series of matrices enriched in TF bound regions, suggesting complex TF-TF interactions. However, they cannot distinguish direct versus indirect TF-DNA binding and do not provide information on the nature of the interaction between the enriched TFBSs.

Most TFs exert their function as part of large transcriptional complexes (7). Their DNA binding sites tend to cluster together in the same cis-regulatory regions (8). These interactions are critical for TFs to achieve their roles (9). TFs can form homodimers within the same family such as estrogen receptor alpha (ERα), which binds to palindromic estrogen response elements (ERE) (10) or can form heterodimers such as FOS/JUN (11) and MYC/MAX (12). TF interactions across families can be also be mediated by the formation of DNA-binding complexes of multiple proteins, for instance within enhancers (13). The spacing between sites bound by interacting TFs is dependent on the structural constraints of the interacting TFs (14,15). However, TF-TF interactions can take also place via the formation of chromatin loops between more distant sites. In this case, each TF may be found associated with its own site and that of the interacting TF if the cross-linking conditions allow detection of this interaction, but these sites will likely be separated during the fragmentation process. Finally, TFs may be recruited to DNA via protein-protein interaction with other TFs without binding to its own site (tethering). All these models are difficult to distinguish without additional experimental data.

Complementary computational approaches have been developed to predict cis-regulatory modules (CRM), defined as groups of DNA motifs that appear together in closely genomic positions more often than what is expected by chance. These approaches aim at predicting association between single signals and the spatial relationship between them (16-24, ADD BLANCHETTES GENOME RESEARCH), a challenging experimental and computational task in the limited prior knowledge (5,25,26). Recent analysis of ChIP-chip and ChIP-seq data shows that if two TFs are co-associated, their ChIP peaks (or their binding sites) are not only in close proximity with each other, but the relative distance of each TFBS in respect to the other exhibits a normal-like distribution (27-29). Based on this observation, Zhizhuo Z. et al. conceived CENTDIST for TF co-binding search around the ChIP peak summits (16). This tool takes as input a set of genomic locations representing ChIP-seq/chip peaks (chromosome-peak summit position) and a list of TFBS matrices (PWMs) from TF databases (4) and infer co-TF-binding. CENTDIST computes the distribution of motif occurrences with respect to the peak summit for each motif using different matrix

cut-offs. Then, it estimates the set of parameters that maximizes the TF co-binding score. Another approach, SpaMo, aims to detect the enrichment of motif spacing rather than the enrichment of motif occurrences (30). SpaMo scans input sequences using the PWM representing the primary TF and selects the best matrix hit on each sequence. Next, on each sequence, SpaMo searches for secondary motif hits using matrices from TF databases and calculates the distance between the primary and the secondary hits on each sequence. However, these tools are limited to known TFBS models and cannot be easily used with a series of input matrices obtained from motif discovery tools.

Here, we developed a TF-TF interaction module that can be used in association with either TFBSs search or discovery programs. Recently, we proposed SAMD-ChIP, a new computational approach for DNA motif discovery adapted to ChIP-chip and ChIP-seq data. Our approach takes advantages of TFBS enrichment in the vicinity of the ChIP peak summits. SAMD-ChIP (Aid., et al., in preparation) predicts enriched motifs in a selected window centered on the ChIP-peak summits. However, the computational efficiency of our approach decreases when looking for long motifs ($\geq$ 20 bp length). This class of motifs generally corresponds to a module composed by two or more TFBSs, which are separated by variable spacer length. It is more suitable to perform first motif discovery for short motifs ($\leq$ 20 bp), then infer eventual co-binding of TFs based on the co-localization of their binding sites in a specific window of the ChIP peak summits more frequently than what is expected by chance. We therefore developed an approach that assesses both co-enrichment of TFBSs in ChIP regions and determines whether specific arrangements of these motifs as direct repeats, inverted or everted repeats are enriched in these regions. Finally, enrichment of the resulting potential composite motifs in the ChIP peak regions is compared to those of the individual motifs to assess the collaborative nature of the association between these TFs. We show that this approach identifies correctly known association of sites such as the direct repeats or palindromes of PuGGTCA motifs found in nuclear receptor binding sites, and also detects association of motifs within fixed distance such as observed in repeat regions found in the vicinity, but not enriched in the center of ChIP peak regions.

MATERIALS AND METHODS

ChIP data sets description

To gain insight into the transcriptional regulatory network in MCF7 breast cancer cells, we collected a set of ChIP-chip and ChIP-seq data sets against ERα, FOXA1, FOS, RARα, AP2y and GATA3 in MCF7 breast cancer cells treated by estrogens (E2) (Table S1). We re-analysed these data sets using our *do novo* motif discovery tool SAMD-ChIP (AID et al., in preparation) to identify putative TF binding sites ranging from 6 to 20 bp, in a fixed window length (100 bp) around the ChIP peak summits for all data sets. Next, for each TF data set, we compared the discovered motifs to TF databases (4) using STAMP tool (31).

Screening for TFBSs

Human Genomic sequences +/- 500 bp around the center the ChIP regions of each datasets were extracted from the UCSC Genome Browser Database (hg18, Mar. 2006) (32). Custom ERE, RARE matrices and matrices from TRANSFAC 2010.2 (5) were used to screen these sequences for transcription factor binding sites using a base score cut-off of 60% and 5% increments. For each transcription factor, four reference cut-offs were chosen with an average number of predictions between 2 and 0.10 per 20 Kbp around the TSS of all genes as described previously (33). Background datasets used to calculate fold enrichment were generated with Ushuffle tool (34).

Inference of TFBSs co-localization

In order to predict TFBS co-localization, we extracted for each matrix pair all their co-occurrences located on common fragments. To do so, we first used a home matrix-search pipeline adapted to ChIP-chip and ChIP-seq data to scan the TF bound regions for different matrix cut-offs. Next, for each matrix-cut-off, the significance of the size of the fraction containing both TFBSs versus those of the fractions containing each motif is evaluated using a Fisher exact test. Next, we measured the distance between the start positions of each motif for significantly co-occuring motif pairs, distinguishing combinations as direct

repeats, inverted or everted repeats. Then, we measured the spacing between these motifs. In case the two matrices represent motifs which form a composite motif, the distance between their occurrences will form a peak at a specific value. The probability of non-random occurrence of the observed peaks in spacing or distances is measured by a permutation test under *R* in which the number of times the optimal spacing/distance is observed in 1000 random permutations of the site positions (35). Finally, we evaluated the enrichment of the composite motif in the vicinity of ChIP peak summits.

RESULTS

Co-enrichment of several transcription factor binding sites in MCF7 ChIP datasets

Estrogen signalling controls gene regulatory networks via the estrogen receptor alpha, which acts a ligand-modulated transcription factor (33,36). Several genome-wide studies have identified ERα bound regions in MCF7 breast cancer cells and have revealed a co-enrichment of other transcription factor binding regions, including those for RARα, FOXA1, AP1, AP2 and GATA3, in the chromatin fragments associated with this receptor (37-42). Conversely, enrichment of EREs has been observed in datasets of RARα, FOXA1, GATA3, AP1 and AP2 obtained also in MCF7 cells. The corresponding TFs may thus cooperate with ERα in the control of breast cancer cell proliferation. However the functional interplay between these transcription factors still remains incompletely understood.

To examine the mode of TF interactions, we systematically compared the enrichment of binding sites for all these TFs in the corresponding datasets. Due to the difference in size data sets, we selected the top 4000 ChIP fragments for all data sets except for ERα, where we used the whole data set. Enrichment of TFBSs was calculated in the 1kb around the ChIP peak summits for all data sets by using shuffle sequences as background.

Co-enrichment of DNA bound regions for FOXA1, AP1, AP2 and GATA3 has been previously reported (37-42) and conversely enrichment of EREs has been observed in datasets of RARα, FOXA1, GATA3, AP1 and AP2. To determine the extent of

transcription factor co-localization in regulatory regions, we compared relative site enrichments in the peaks of bound ChIP regions in these datasets (Figure 1). Strikingly, EREs were highly enriched (>10-fold at a cut-off of 75%) in most datasets; in particular, in the FOXA1 and RARα datasets. Note that the enrichment in EREs in the ERα bound regions may be underestimated as it is expected to be high-affinity than in the overall data set. Remarkably, however, enrichment of EREs in the RARα and AP2 data sets was higher than that of the binding sites of these TFs at the same matrix cut-off, while it was lower but still comparable for those of AP1 and FOXA1 binding sites in the data sets of the corresponding TFs (18 vs 25 and 36 vs 50, respectively, see Figure 1). Enrichment of EREs was however comparatively low in GATA3 dataset. These results, suggest a high relative affinity of all examined TFs expect for GATA3 with EREs, whether due to direct or indirect binding. On the other hand, enrichment in binding sites for other TFs in the ERα dataset was relatively low (2-3 fold compared to 26 fold for EREs for most TFs except 5-6 fold for FOXA1). Together, these observations suggest that ERα has a greater selectivity for its own sites than other transcription factors, including RARα, which like ERα is a member of the same family of nuclear receptors and recognises similar types of sites. Note that the greater frequency of occurrences in the genome of short sites like those of AP1, AP2 and FOXA1 compared to EREs or RAREs results in association of ERα with a large number of sites that contain response elements other than EREs, even with a low affinity. These results suggest that other TFs may be recruited to EREs in part via interactions with ERα. These interactions may involve tethering of these factors to the ERE or co-binding of the TFs and ERα to their respective sites. Indeed, we observed significant overlap between DNA bound regions for all these TFs (Figure 2 A and B and Figure S2).

We then assessed whether co-localization of EREs and other TF binding sites is observed within ChIP regions. In spite of the mutual enrichment of EREs, FOXA1 and AP1 sites in their datasets, co-occurrences of EREs with either FOXA1 or AP1 sites was not statistically significant in the ERα data set. No significant pair-wise association of EREs were detected either, consistent with most ChIP fragments being centred on one (or none) rather than several EREs.

Similarly no association was observed between EREs and DR5 elements, which mediate DNA binding of RARα, in either ERα or RARα datasets. We further examined the association of sites in ChIP regions found in common between the two datasets. The RARα dataset contains 50% of total regions (2463/4867) in common with ERα dataset. In those regions, EREs were found in 402 regions (65% cut-off) and DR5 in 356 regions (65% cut-off), and 68 regions contained both binding sites (Figure 3). This suggests that the co-localization of these TFs is not mediated by co-occurrence of binding sites, denoting a lack of significant association between these sites for different cut-offs. These results suggest that the co-localization of TFs is not mediated by co-occurrence of their binding sites, but possibly either via tethering of one factor to the other bound to its site, or via formation of chromatin loops between different ChIP regions resulting in reciprocal association of TFs to each other's sites.

Nucleotide bias for EREs enriched in different datasets

Another possibility for the co-association of two transcription factors is cross-recognition of each other's sites. This possibility is especially plausible for TFs belonging to the same family such as ERα and RARα, which both recognize PuGGTCA motifs. We therefore investigated whether ERE motifs enriched in these datasets display a similar nucleotide bias at the different positions (Figure 4). To further examine nucleotide bias, we compare the enrichment of matrices with a nucleotide bias at different positions using different cut-offs. Preferences exhibits were similar in both datasets, with the best tolerated substitution at position 1 and the least favourable at position 3, 5 and 6. However, a relative preference for T at position 5 was observed in the ERα dataset (rank 9 with 10.6 fold enrichment vs rank 13 with 5 fold enrichment at high cut-off, while a preference for C at position 2 and 4 was observed in RAR dataset (ranks 6 and 9 vs 14 and 13 respectively, with enrichment of 8.7 and 7.5 vs 7.8 and 8.5).

Association of half EREs with other motifs

Our motif discovery tool identified several composite motifs containing half ERE as weakly, but enriched in datasets for ERα and RARα. These include combination of half

EREs with AP1 or FOXA1-like binding sites (Table 1). We therefore conducted further studies to assess whether composite sites may mediate co-association of TFs with ERα by assessing co-enrichment of half ERE with other sites. We first verified that half EREs are significantly associated with each other in the ERα and RARα datasets. EREs are palindrome motifs composed by two half EREs separated by a 3 nucleotides spacer (RGGTCAnnnTGACCY) (33).

In ERα bound regions, the half ERE motif was considerably less enriched than the full ERE motif, reflecting the lower affinity of the receptor for these sites (Figure 5). However, these motifs are frequent and pairs of half EREs were identified in a large number of ChIP fragments. Co-localization (presence of at least two motifs) was statistically significant, and as expected statistically significant bias for a distance between repeats of 14 bp and 3bp spacer was observed in a palindromic conformation (Figure 6). No other configuration appeared enriched in ERα bound regions for palindromic with 3bp spacer. Two motifs composed by a half ERE and an AP1 BS separated by different spacer length were identified by SAMD-ChIP as enriched in both ERα and RARα datasets (Figure 7). However, it is possible that other spacings may also be enriched. Since composite motifs with two different spacer lengths were identified, we first generated for each matrix different forms by varying the spacer length, we set the spacer to N (the same frequency for each nucleotide). Next, we searched for these two matrices in both ERα and RARα bound regions for different matrix cut-offs (75% to 85%). We observed that the two matrices showed significant enrichment for different matrix cut-offs in both ERα and RARα bound regions. The motif identified as Matrix_1789 presents higher fold enrichment for longer spacer lengths (Figure 8). However, the second motif (Matrix_2571) showed variable enrichment depending on the spacer length (Figure 9).To determine whether association of half EREs and AP1 BS is flexible or whether an optimal configuration exists that was not detected by the motif discovery program, we assessed the distribution in spacer lengths between half-ERE motifs and AP1 BS in both ERα and RARα bound regions (1kb window) using our co-localization module. In ERα bound regions, a spacer length of 46 and 48 bp showed the highest peak intensities (Figure 10 A). As AP1 BS is palindromic, similar results were obtained with the different

arrangement of the motifs. This result was confirmed in RARα bound regions with higher enrichment (Figure 10 B). Of note however, matrices composed by a half ERE motif and an AP1 BS with 46 and 48 bp spacer lengths do not exhibit a significant enrichment in either ERα and RARα ChIP peak summits. These results denote an abundance of these composite motif in the vicinity of ChIP regions (within 1 kb), but not in the center (Figure 11 A and B). Analysis of the sequence of these composite motifs identified their overlap with Alu repeats. SAMD-ChIP identified two composite motifs in ERα bound regions composed by a half ERE motif and an AT rich motif (Figure 12 A and B). Since the forkhead family protein bind to an AT rich motif, we hypothesized that this motif may be bound by FOXA1, which is known to be enriched in ERα bound regions. Several studies highlighted the pioneer role of FOXA1 in mediating ERα binding and activity in MCF7 cells (43). However the existence of specific composite half ERE-FOXA1 sites has not been investigated. We therefore assessed the co-enrichment of half ERE and FOXA1 motif in the ERα and RARα bound regions. We observed a significant co-enrichment of half EREs and the FOXA1-like motif. However, we did not find a bias for a specific spacer length between the two motifs (Figure 12 C). Yet, the prediction of a composite motif in ERα and RARα bound regions may suggest a cooperative binding between ERα/RARAα and FOXA1 (Figure 13 A and B).

Analysis of ERα and RARα bound regions reveals the weak enrichment of new composite motifs, composed by two Ct/gG boxes and separated by variable spacer length (Figure 14 A and B). Based on motif enrichment measured by SAMD-ChIP in these data sets, we did not observe a bias for a specific spacer length. The observed fold enrichment varies from 1.5 to 2, depending on the spacer length. The presence of these different composite motifs may suggest the presence of a TF that binds to a Ct/aG box motif as a homo-dimer with variable spacer length.

Next, we sought to establish a correlation between the predicted composite motif formed by a half ERE and a Ct/aG motif and the presence of a new ERα/RARα partner which bind to a Ct/aG DNA motif (Figure 15 A and B). First, we used our co-localization pipeline to estimate the optimal spacer length between the half ERE and the Ct/aG motif in both ERα

and RARα bound regions. In both data sets, we did not observe any bias for a specific spacer length between a half ERE and the Ct/gG motif. Indeed, the Ct/aG motif is enriched in both ERα and RARα as shown in Figure S3 and S4. However, the spacer length showed a uniform distribution in a window of 100 bp (Figure S3 and S4). The distribution of the spacer length between the two motifs showed a peak at 0 and 2 bp due to the high proportion of overlap between the two motifs (Figure S3 and Figure S4).

Now, we aimed to verify if motifs composed by half ERE and the Ct/aG box are variations of ERE motif or not. First, we analysed the motif identified as Matrix_3789. This motif could be considered as a degenerate form of an ERE with a conserved half site and a degenerate one. To test this observation, we created two new matrices by increasing the spacer length by 1 and 2 nucleotides respectively. Then, we scanned the ERα bound regions using these matrices and the original one for the same matrix cut-offs. As shown in Figure 15 A, the fold enrichment decreased when increasing the spacer length. Thus, we conclude that Matrix_3789 is a variation of an ERE and may be bound by ERα with low affinity (44).

Next, we asked whether the second motif identified as Matrix_8344 is also a variation of an ERE motif or a new composite motif containing both half ERE and a new motif. To answer this question, we first generated three new matrices from the Matrix_8344 as follow: the first matrix is similar to Matrix_8344 where we replaced the spacer by N (any nucleotide) and we eliminated the two last positions, the two other matrices were created by increasing the spacer by 1 and 2 nucleotides. All these matrices were then used to screen ERα bound regions. The result showed that when increasing the spacer the fold enrichment is increased (Figure 15 B). We also found that the composite motif is more enriched than each motif alone (data not shown). All these results may suggest a cooperative interaction mechanism between ERα and a TF that binds to the Ct/aG motif and ERα DNA binding may be stabilized by the presence of this new partner.

DISCUSSION

In higher eukaryotes, physical interactions between TFs are considered as one of the most important mechanisms governing transcription. In this paper we aim to infer TFs co-localization from ChIP-chip and ChIP-seq data by exploiting the spatial proximity between bound motifs. We presented a new module to identify TFBSs co-localization from ChIP-chip and ChIP-seq data. This module will be implemented as part of our *do novo* motif discovery approach SAMD-ChIP (Aid et al., paper in preparation). This module uses a pair of TFBSs and reports the distribution of their occurrence positions in a window around the ChIP-peak summits. It will also estimate the distance between the start positions for each pair of occurrences and the spacer length between them. It is of interest to note that this module can also be applied on a set of matrices from TFBS data bases and is not limited only to SAMD-ChIP outputs. Our approach is not limited by the TFBSs proximity assumption. Users could select a central window length according to the analysed data to perform motif discovery before inferring motifs co-localization. In addition to the estimation of the optimal spacer length and the distance between two TFBSs, our pipeline also ensures that the composite motif is enriched in the vicinity of the ChIP peak summits. Finally, the enrichment of the composite motif is compared to those of individual motifs to assess cooperativity between the two TFs.

First, we validated our module by identifying the optimal spacer length and the distance between two half EREs forming the ERE motif bound by ERα. Analysis of several composite motifs, predicted by SAMD-Chip in both ERα and RARα bound regions, highlights the effect of the pacer length and composition on composite motif enrichment.
Different mechanisms can mediate cross-association of TFs on other TF binding sites. TFBSs can co-localize in proximity to each other to mediate TF DNA binding cooperativity. This mechanism was validated on half EREs which form the ERE motif bound by ERα. We validated the co-localization of these two motifs and predicted the optimal distance and spacer between them.

Next, we assessed whether the prediction of a composite motif formed by a half ERE and an AP1 binding site could suggest a cooperative DNA binding between ERα and AP1. Using our co-localization module, we predicted two optimal spacer lengths (46 and

48 bp) between the half ERE and the AP1 BS in both ERα and RARα bound regions. However, these composite motifs are not enriched in the vicinity of the ChIP peak summits. We observed that most of these sites overlap with ALU repeats. These results suggest that repeats may play an important role in TF DNA binding.

Using SAMD-ChIP, we also predicted a composite motif formed by both half ERE and a FOXA1-like motif. This motif has a 2-fold enrichment in both ERα and RARα bound regions. However, our analysis did not reveal a bias for a specific space length and distance between the half ERE and the FOXA1-like motif. It was reported before that FOXA1 plays a pioneer role in mediating ERα DNA binding (45). However, our results do not favour the mechanism by which FOXA1 would act by opening chromatin structure for a significant portion of EREs.

EREs motifs in RARα data set could be bound by RARα directly or via ERα (46). Small differences in binding observed could result from different nuclear receptor bound or from selection of a subset of ER-bound EREs by tethered TF. EREs have been suggested before to have allosteric effects on cofactor recruitment (44). It will be interesting to validate this observation for other TFs. However, this differences in binding could result from biases introduced by antibodies used, and would have to be confirmed using unrelated antibodies.

TFs DNA binding can be mediated by both tethering and chromatin loops. In both cases, could be affected by experimental conditions. For example, low formaldehyde content should favour direct interactions versus tethered or loops as two cross-link events are needed vs 1 for direct binding. This may explain the low enrichment of other TF binding sites in ERα bound regions compared to the enrichment of EREs in other data sets. In case of chromatin loops, reciprocal association of TFs with each other's sites would be expected. Globally, the simplest model that takes into account these results is that ERα binds to its sites with high affinity and attracts binding by other TFs either by tethering or formation of long range interactions with TFs bound to their own binding sites. If other TFs have a lower affinity to their own sites, they may partition significantly on ERα binding sites even though the second order cross-link is less-efficient. In addition, TF-TF

interactions may modulate binding of one partner more than the other through allosteric interactions. This may explain for instance the high enrichment of EREs in FOXA1 bound regions while the FOXA1 motif in not highly enriched in the ERα dataset, while both factors have a high affinity for their respective sites. It will be of interest in the future to characterize the TF-TF interactions between the different factors studied here and determine how modulation of a given TF affects binding sites by others. For instance, it would be expected that pioneer factors would be little affected in their binding patterns by the suppression of other TFs.

The strength of the present analysis lies in the survey of multiple whole-genome TFs binding in MCF7 breast cancer cells and should provide insight into the integration of the signalling pathways in ERα+ breast cancer cells.

References

1. Gilchrist, D.A., Fargo, D.C. and Adelman, K. (2009) Using ChIP-chip and ChIP-seq to study the regulation of gene expression: genome-wide localization studies reveal widespread regulation of transcription elongation. *Methods*, **48**, 398-408.
2. Hao, H. (2012) Genome-wide occupancy analysis by ChIP-chip and ChIP-Seq. *Adv Exp Med Biol*, **723**, 753-759.
3. Thudi, M., Li, Y., Jackson, S.A., May, G.D. and Varshney, R.K. (2012) Current state-of-art of sequencing technologies for plant genomics research. *Brief Funct Genomics*, **11**, 3-11.
4. Chen, H.F. and Wang, J.K. (2010) [The databases of transcription factors.]. *Yi Chuan*, **32**, 1009-1017.
5. Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, **34**, D108-110.

6. Zambelli, F., Pesole, G. and Pavesi, G. (2012) Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Brief Bioinform*, **19**, 19.
7. Lemon, B. and Tjian, R. (2000) Orchestrated response: a symphony of transcription factors for gene control. *Genes & development*, **14**, 2551-2569.
8. Zhang, J., Yuan, Z. and Zhou, T. (2009) Combinatorial regulation: characteristics of dynamic correlations. *IET Syst Biol*, **3**, 440-452.
9. Chen, L., Glover, J.N., Hogan, P.G., Rao, A. and Harrison, S.C. (1998) Structure of the DNA-binding domains from NFAT, Fos and Jun bound specifically to DNA. *Nature*, **392**, 42-48.
10. Bjornstrom, L. and Sjoberg, M. (2005) Mechanisms of estrogen receptor signaling: convergence of genomic and nongenomic actions on target genes. *Mol Endocrinol*, **19**, 833-842.
11. Karin, M., Liu, Z. and Zandi, E. (1997) AP-1 function and regulation. *Curr Opin Cell Biol*, **9**, 240-246.
12. Nair, S.K. and Burley, S.K. (2003) X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular bases of regulation by proto-oncogenic transcription factors. *Cell*, **112**, 193-205.
13. Wadman, I.A., Osada, H., Grutz, G.G., Agulnick, A.D., Westphal, H., Forster, A. and Rabbitts, T.H. (1997) The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *Embo J*, **16**, 3145-3157.
14. Wu, L. and Berk, A. (1988) Constraints on spacing between transcription factor binding sites in a simple adenovirus promoter. *Genes & development*, **2**, 403-411.
15. Wolberger, C. (1999) Multiprotein-DNA complexes in transcriptional regulation. *Annu Rev Biophys Biomol Struct*, **28**, 29-56.
16. Zhang, Z., Chang, C.W., Goh, W.L., Sung, W.K. and Cheung, E. (2011) CENTDIST: discovery of co-associated factors by motif distribution. *Nucleic Acids Res*, **39**, 20.

17. Zhou, Q. and Wong, W.H. (2004) CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci U S A*, **101**, 12114-12119.

18. Donaldson, I.J. and Gottgens, B. (2007) CoMoDis: composite motif discovery in mammalian genomes. *Nucleic Acids Res*, **35**, 27.

19. Aerts, S., Van Loo, P., Thijs, G., Moreau, Y. and De Moor, B. (2003) Computational detection of cis -regulatory modules. *Bioinformatics*, **19**, ii5-14.

20. Sandelin, A., Wasserman, W.W. and Lenhard, B. (2004) ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res*, **32**, 249-252.

21. Sharan, R., Ben-Hur, A., Loots, G.G. and Ovcharenko, I. (2004) CREME: Cis-Regulatory Module Explorer for the human genome. *Nucleic Acids Res*, **32**, W253-256.

22. Sinha, S., van Nimwegen, E. and Siggia, E.D. (2003) A probabilistic method to detect regulatory modules. *Bioinformatics*, **19**, i292-301.

23. Roven, C. and Bussemaker, H.J. (2003) REDUCE: An online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data. *Nucleic Acids Res*, **31**, 3487-3490.

24. Sinha, S., Liang, Y. and Siggia, E. (2006) Stubb: a program for discovery and analysis of cis-regulatory modules. *Nucleic Acids Res*, **34**, W555-559.

25. Su, J., Teichmann, S.A. and Down, T.A. (2010) Assessing computational methods of cis-regulatory module prediction. *PLoS Comput Biol*, **6**.

26. Klepper, K., Sandve, G.K., Abul, O., Johansen, J. and Drablos, F. (2008) Assessment of composite motif discovery methods. *BMC Bioinformatics*, **9**, 123.

27. Cheung, E. and Kraus, W.L. (2010) Genomic analyses of hormone signaling and gene regulation. *Annu Rev Physiol*, **72**, 191-218.

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

29. He, H.H., Meyer, C.A., Shin, H., Bailey, S.T., Wei, G., Wang, Q., Zhang, Y., Xu, K., Ni, M., Lupien, M. *et al.* (2010) Nucleosome dynamics define transcriptional enhancers. *Nat Genet*, **42**, 343-347.

30. Whitington, T., Frith, M.C., Johnson, J. and Bailey, T.L. (2011) Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res*, **39**, e98.

31. Mahony, S. and Benos, P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res*, **35**, 3.

32. Kuhn, R.M., Karolchik, D., Zweig, A.S., Wang, T., Smith, K.E., Rosenbloom, K.R., Rhead, B., Raney, B.J., Pohl, A., Pheasant, M. *et al.* (2009) The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res*, **37**, 7.

33. Bourdeau, V., Deschenes, J., Laperriere, D., Aid, M., White, J.H. and Mader, S. (2008) Mechanisms of primary and secondary estrogen target gene regulation in breast cancer cells. *Nucleic Acids Res*, **36**, 76-93.

34. Jiang, M., Anderson, J., Gillespie, J. and Mayne, M. (2008) uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, **9**, 192.

35. Fu, A.Q. and Adryan, B. (2009) Scoring overlapping and adjacent signals from genome-wide ChIP and DamID assays. *Molecular bioSystems*, **5**, 1429-1438.

36. Lupien, M. and Brown, M. (2009) Cistromics of hormone-dependent cancer. *Endocr Relat Cancer*, **16**, 381-389.

37. Eeckhoute, J., Keeton, E.K., Lupien, M., Krum, S.A., Carroll, J.S. and Brown, M. (2007) Positive cross-regulatory loop ties GATA-3 to estrogen receptor alpha expression in breast cancer. *Cancer Res*, **67**, 6477-6483.

38. Jozwik, K.M. and Carroll, J.S. (2012) Pioneer factors in hormone-dependent cancers. *Nat Rev Cancer*, **12**, 381-385.

39. Joseph, R., Orlov, Y.L., Huss, M., Sun, W., Kong, S.L., Ukil, L., Pan, Y.F., Li, G., Lim, M., Thomsen, J.S. *et al.* (2010) Integrative model of genomic factors for determining binding site selection by estrogen receptor-alpha. *Mol Syst Biol*, **6**, 456.

40. Serandour, A.A., Avner, S., Percevault, F., Demay, F., Bizot, M., Lucchetti-Miganeh, C., Barloy-Hubler, F., Brown, M., Lupien, M., Metivier, R. *et al.* (2011) Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. *Genome Res*, **21**, 555-565.

41. Ross-Innes, C.S., Stark, R., Holmes, K.A., Schmidt, D., Spyrou, C., Russell, R., Massie, C.E., Vowler, S.L., Eldridge, M. and Carroll, J.S. (2010) Cooperative interaction between retinoic acid receptor-alpha and estrogen receptor in breast cancer. *Genes & development*, **24**, 171-182.

42. Tan, S.K., Lin, Z.H., Chang, C.W., Varang, V., Chng, K.R., Pan, Y.F., Yong, E.L., Sung, W.K. and Cheung, E. (2011) AP-2gamma regulates oestrogen receptor-mediated long-range chromatin interaction and gene transcription. *Embo J*, **30**, 2569-2581.

43. Lupien, M., Eeckhoute, J., Meyer, C.A., Krum, S.A., Rhodes, D.R., Liu, X.S. and Brown, M. (2009) Coactivator function defines the active estrogen receptor alpha cistrome. *Mol Cell Biol*, **29**, 3413-3423.

44. Bourdeau, V., Deschenes, J., Metivier, R., Nagai, Y., Nguyen, D., Bretschneider, N., Gannon, F., White, J.H. and Mader, S. (2004) Genome-wide identification of high-affinity estrogen response elements in human and mouse. *Mol Endocrinol*, **18**, 1411-1427.

45. Lupien, M., Eeckhoute, J., Meyer, C.A., Wang, Q., Zhang, Y., Li, W., Carroll, J.S., Liu, X.S. and Brown, M. (2008) FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell*, **132**, 958-970.

46. Hua, S., Kittler, R. and White, K.P. (2009) Genomic antagonism between retinoic acid and estrogen signaling in breast cancer. *Cell*, **137**, 1259-1271.

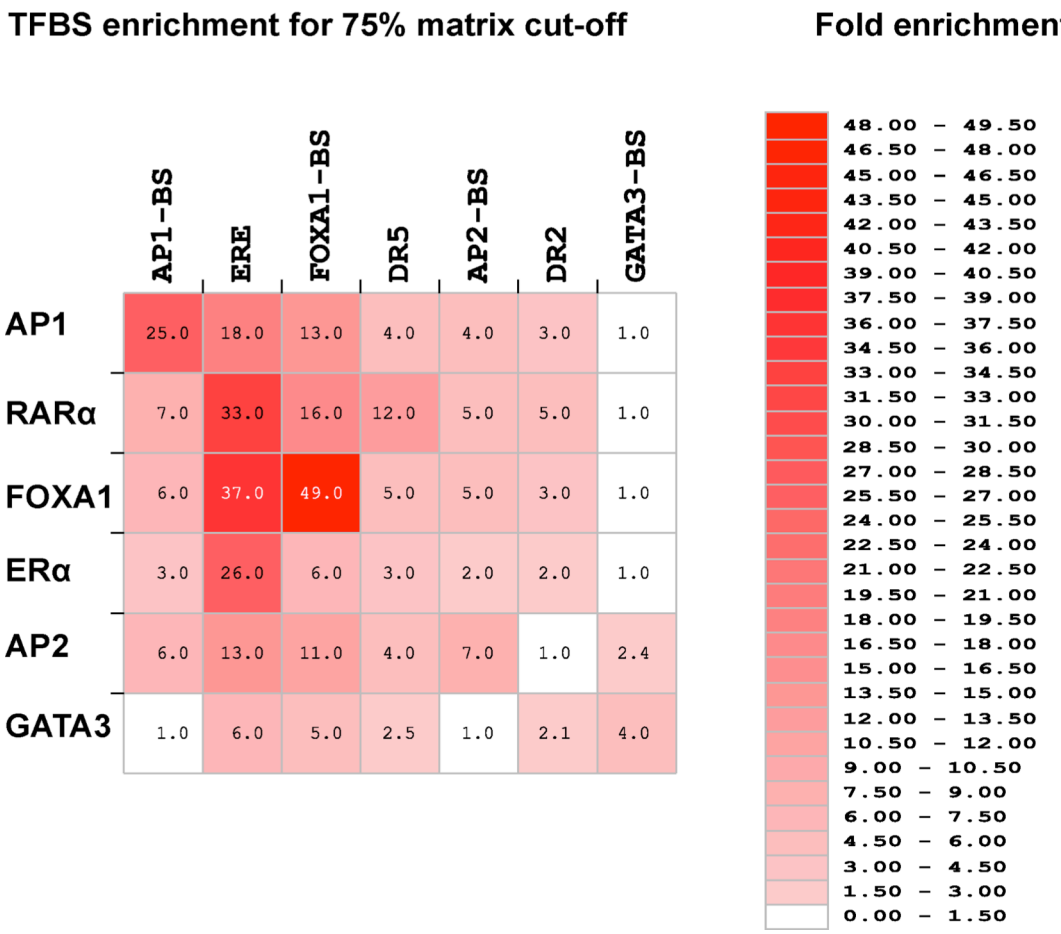


Figure 1. TFBS co-enrichment in different datasets

Heatmap rows represent TF datasets and columns represent TFBSs matrices. The fold enrichment is measured using a 75% matrix cut-off and a shuffle background.

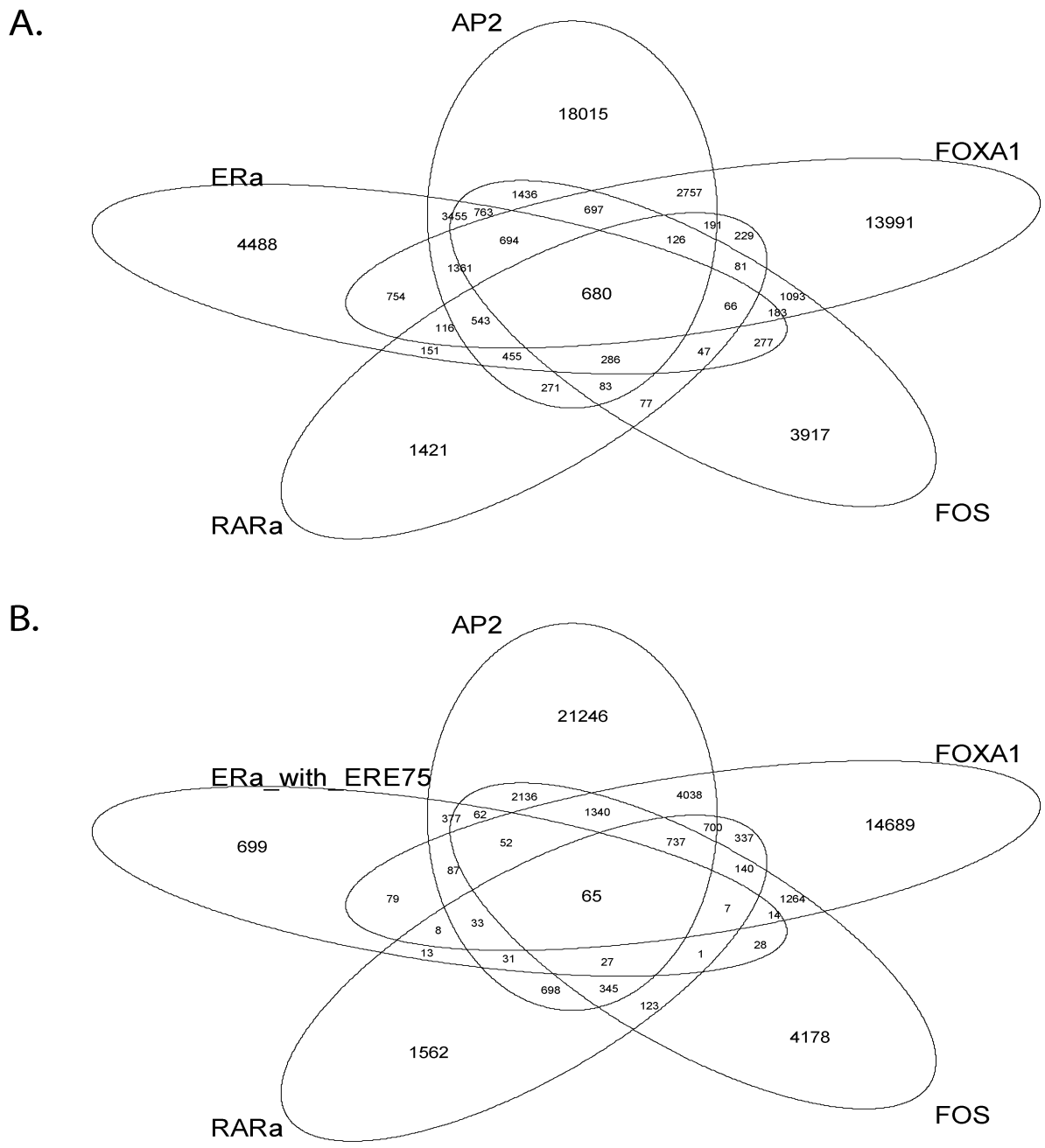
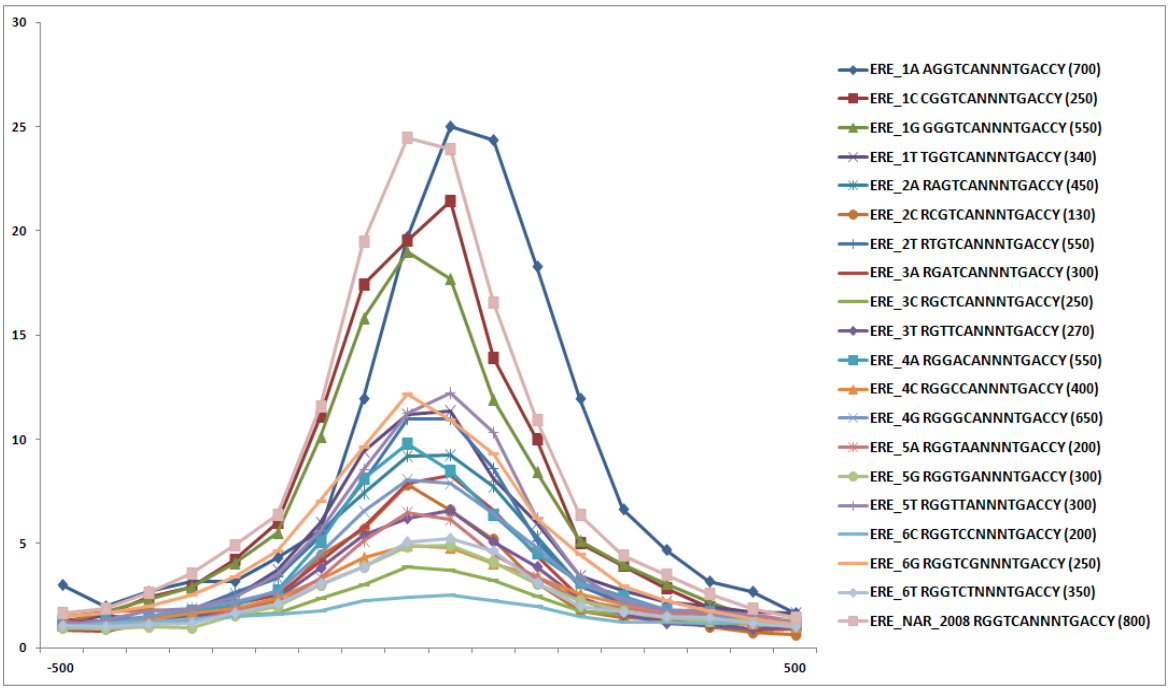


Figure 2. Venn diagram representing overlaps between different datasets in MCF7 cells.

(A) Overlap between all TFs bound regions (B) Overlap between ERα bound regions that contain at least one ERE prediction for 75% matrix cut-off and the other TFs bound regions.

Figure 3. EREs, AP1 BS and DR5 enrichment in regions bound by both ERα and RARα.

A.



B.

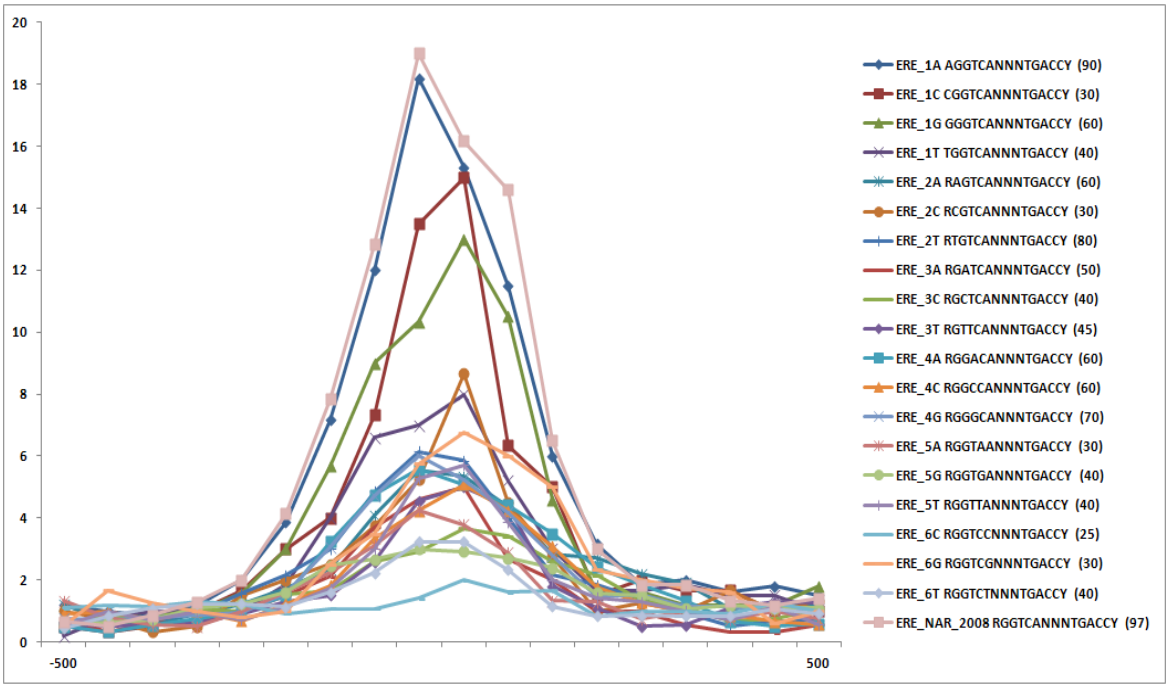


Figure 4. The enrichment of EREs variations in both ERα and RARα bound regions.

Fold enrichment is measured for 75% matrix cut-off. (A) in ERα and (B) in RARα.

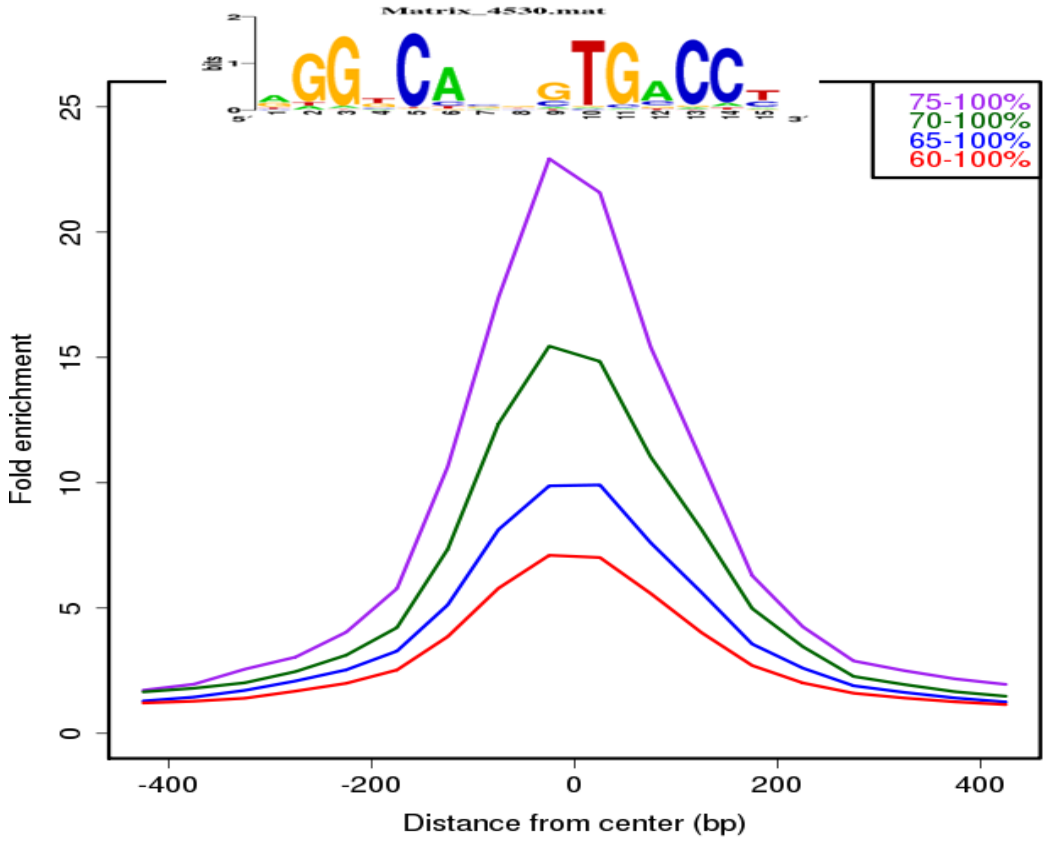


Figure 5. Enrichment of ERE matrix for different matrix cut-offs in ERα bound regions.

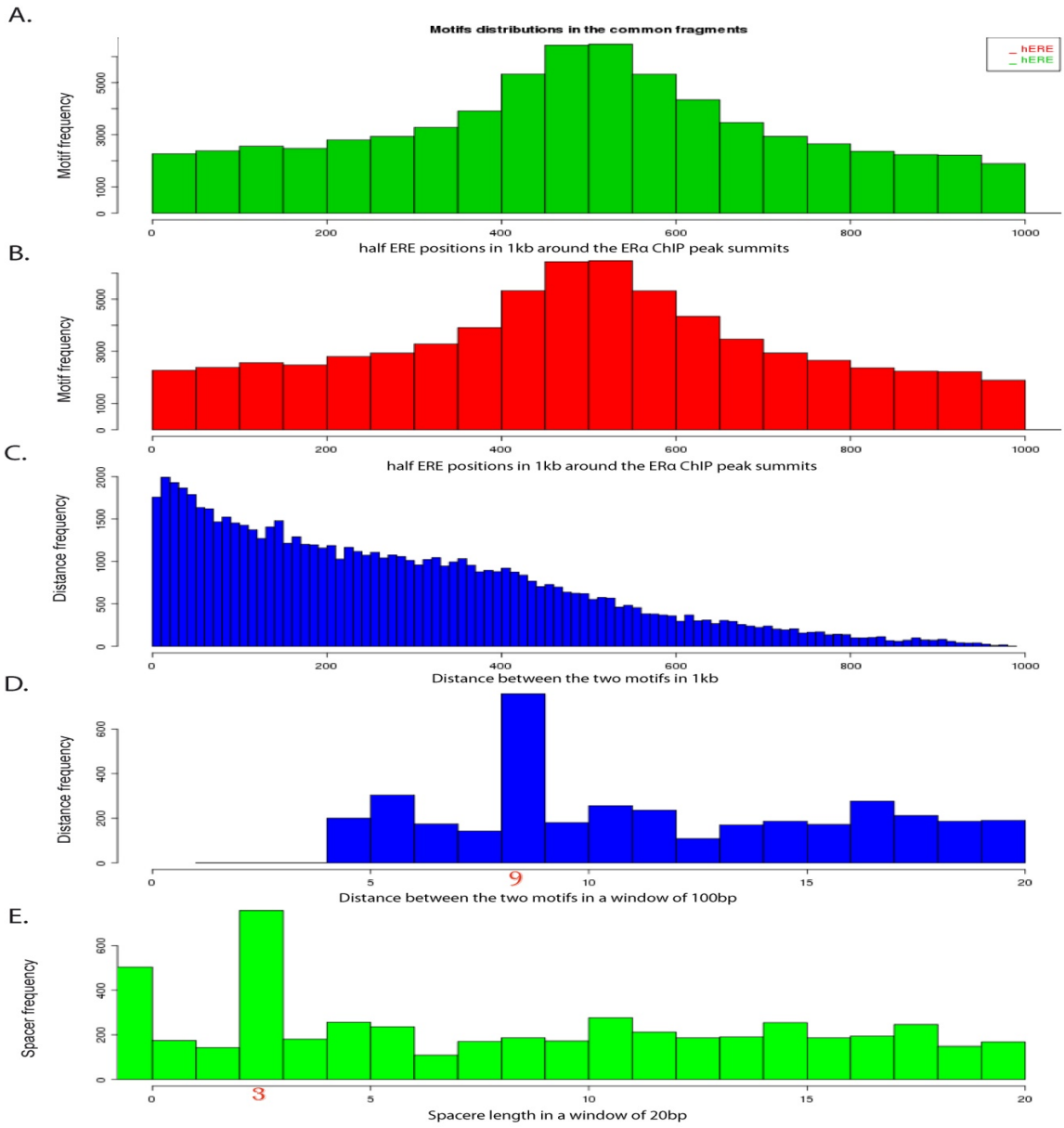


Figure 6. Distribution of half ERE motifs in ERα bound regions.

(A) and (B) represent the distribution of the half ERE positions in 1kb around the ERα ChIP peak summits using a 95% matrix cut-off. (C) Distribution of the distance between the two half EREs in a window of 1kb around the ChIP peak summits. (D) Distribution of the distance frequency between the two half EREs and (E) the distribution of the spacer length frequency between the two half EREs in a window of 20 bp.

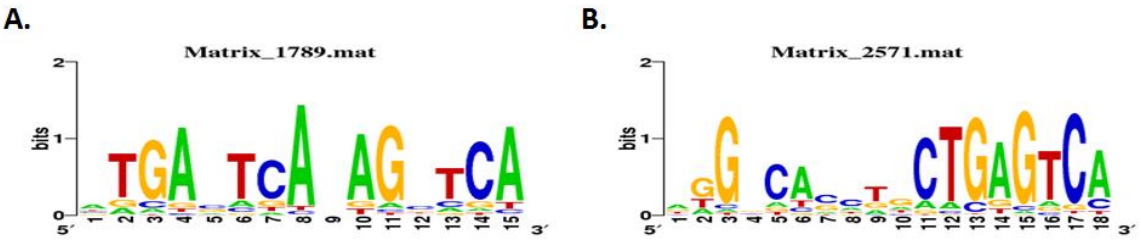


Figure 7. Composite half ERE (RGGTCA) and AP1 (TGANTCA) motifs identified in ERα and RARα bound regions.

(A) and (B) represent two composite motifs composed by and AP1 BS and a half ERE motif separated by 1 and 5 nucleotides respectively, predicted in ERα and RARα bound regions.

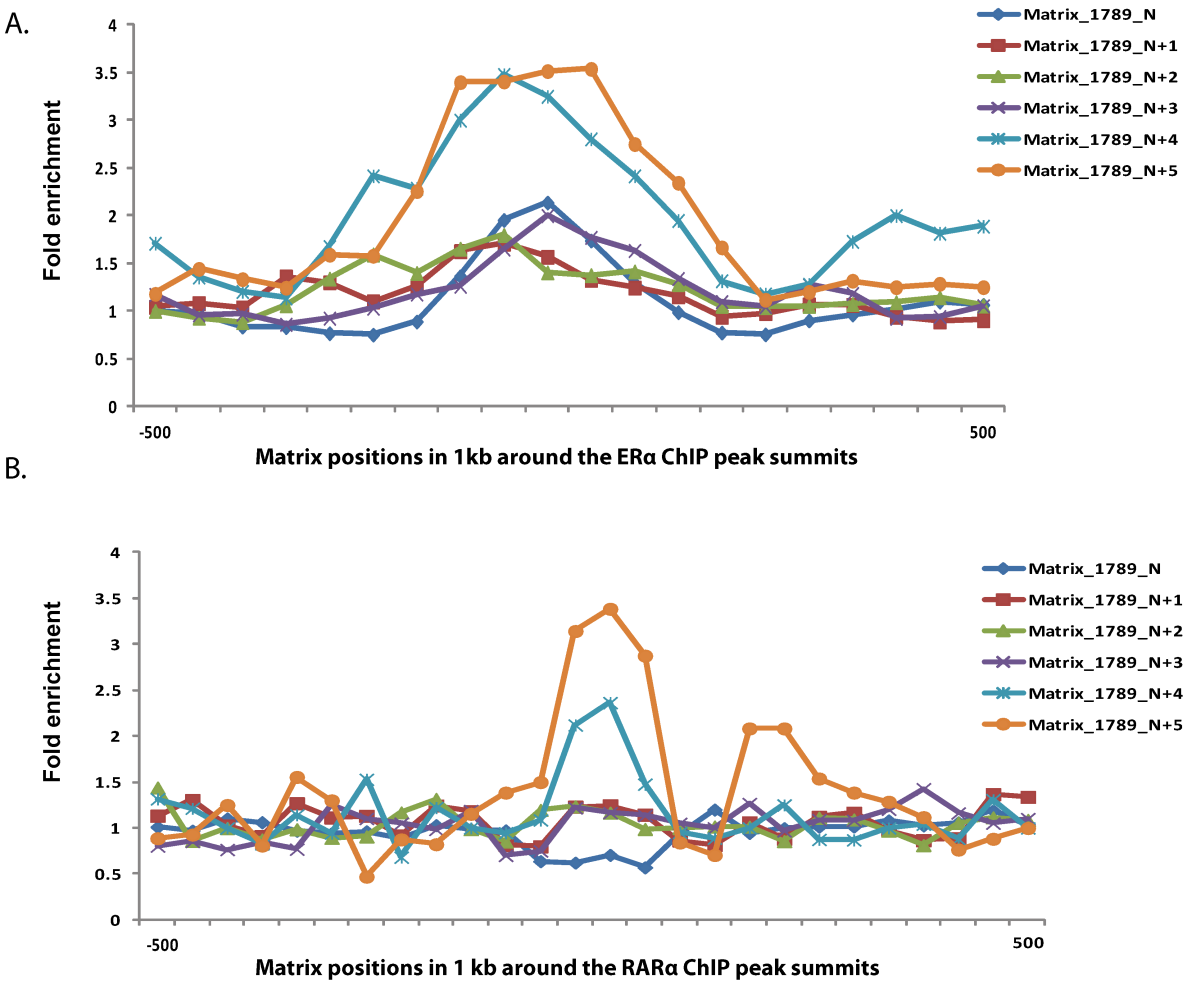


Figure 8. Fold enrichment of the first composite motif in ERα and RARα bound regions using 85% matrix cut-off.

(A) and (B) represent the fold enrichment distribution for different spacer length in ERα and RARα bound regions respectively. Spacer length varies from 1 bp (N) to 6 bp (N+5). Matrix_1789_N represents the original motif.

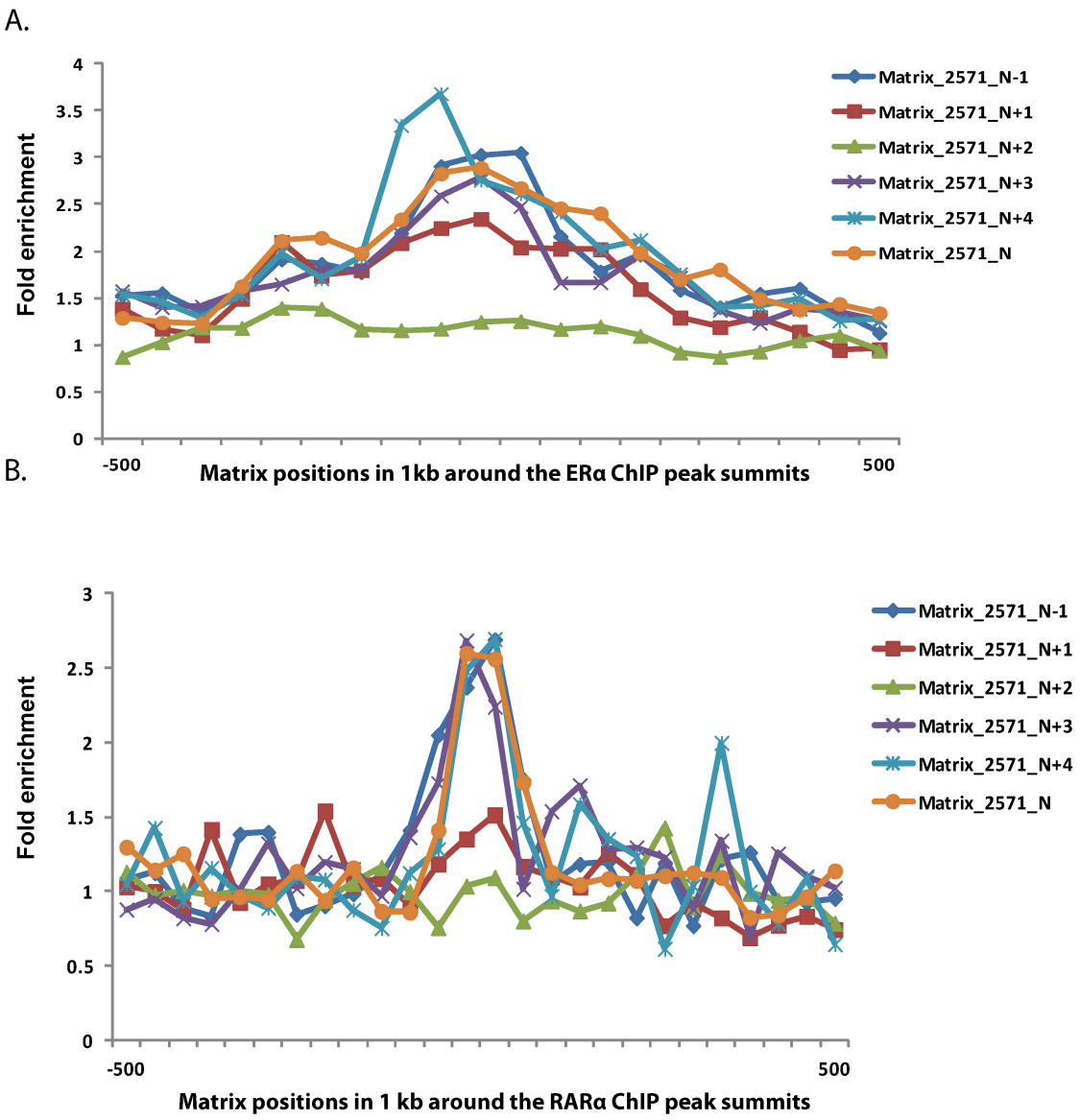


Figure 9. Fold enrichment of the second composite motif in ERα and RARα bound regions using 85% matrix cut-off.

(A) and (B) represent the fold enrichment distribution for different spacer length in ERα and RARα bound regions respectively. Spacer length varies from 4 bp (N-1) to 9 bp (N+4). Matrix_2571_N represents the original motif where we set the spacer to N.

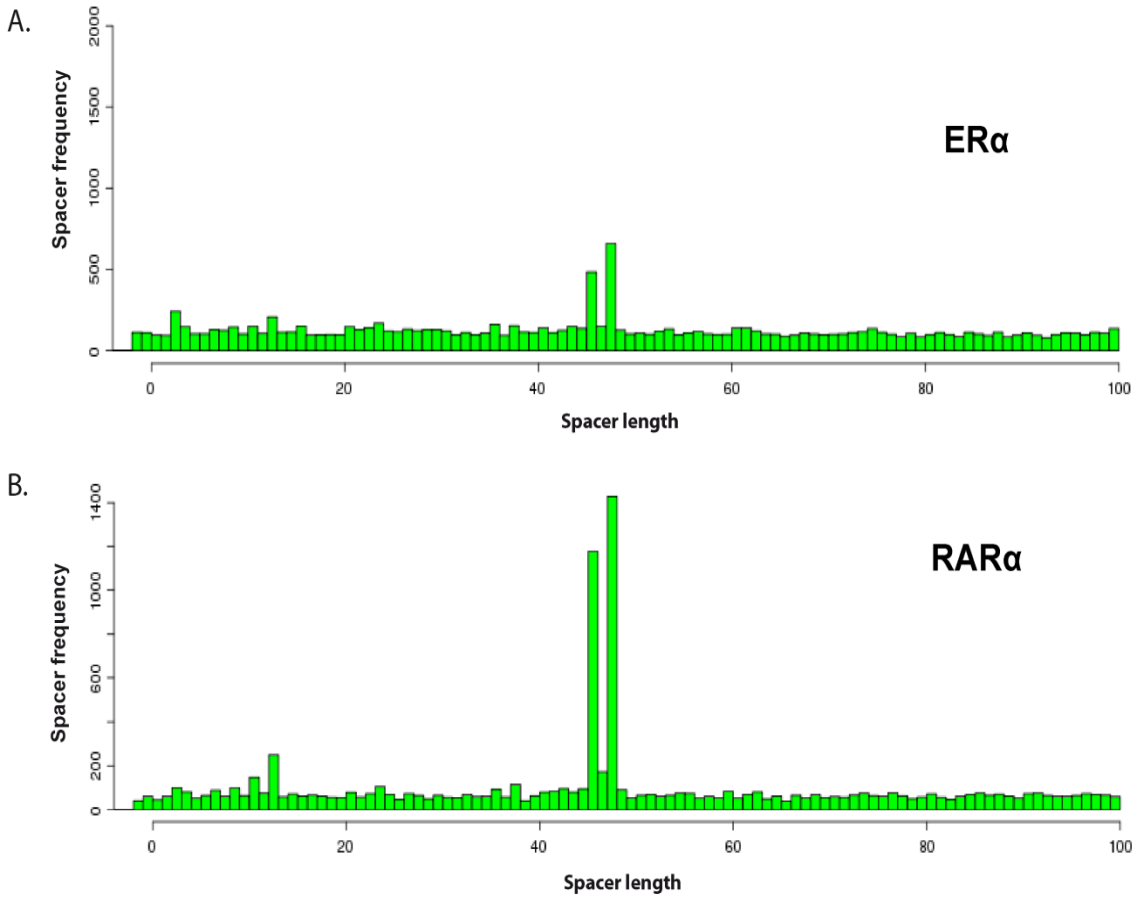


Figure 10. Distribution of spacer length between half ERE and AP1 BS in ERα and RARα bound regions

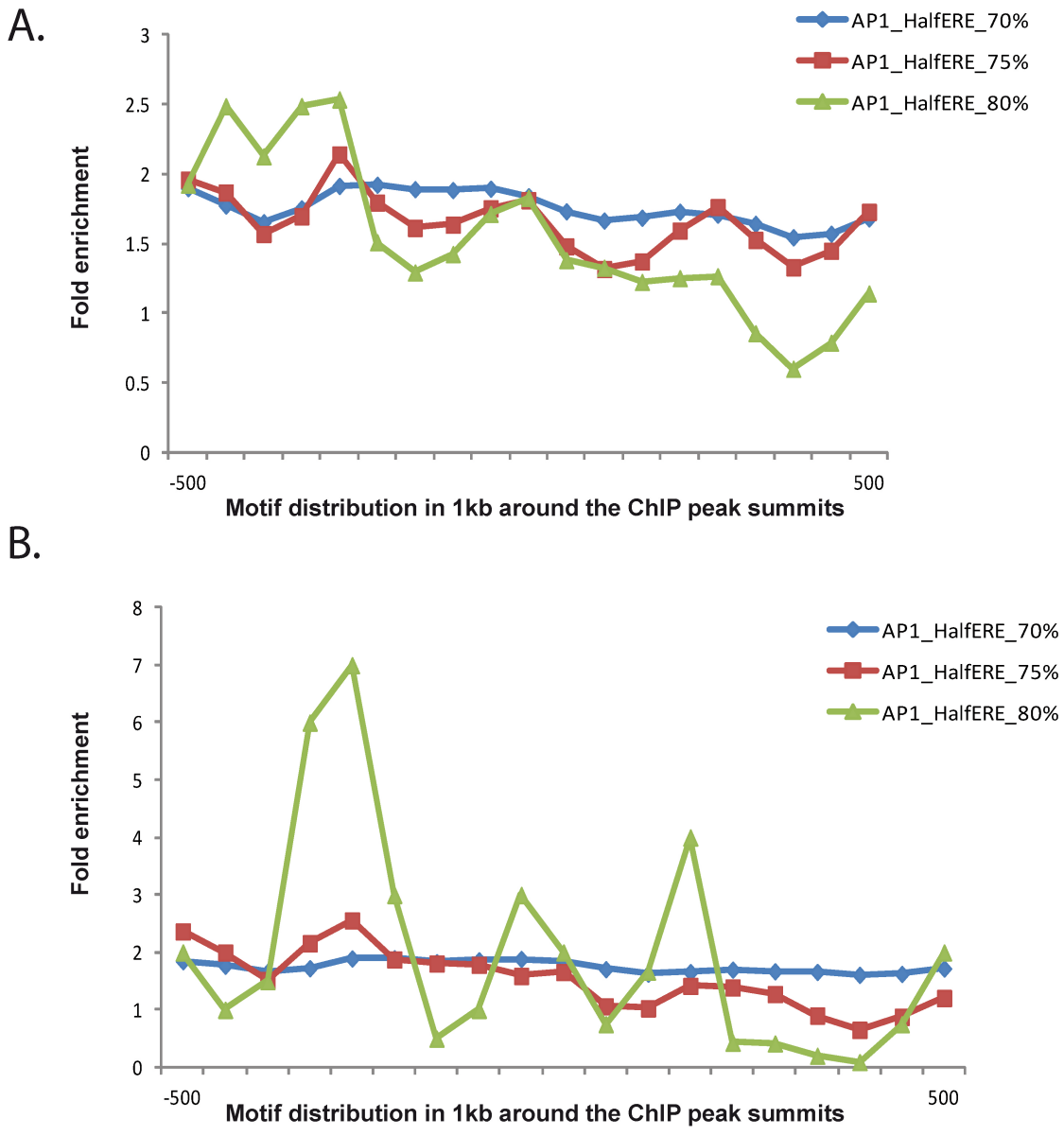


Figure 11. Enrichment of AP1BS and half ERE composite motifs for different matrix cut-off in ERα bound regions.

(A) Composite motif with 46 bp spacer length and (B) composite motif with 48 bp spacer length.

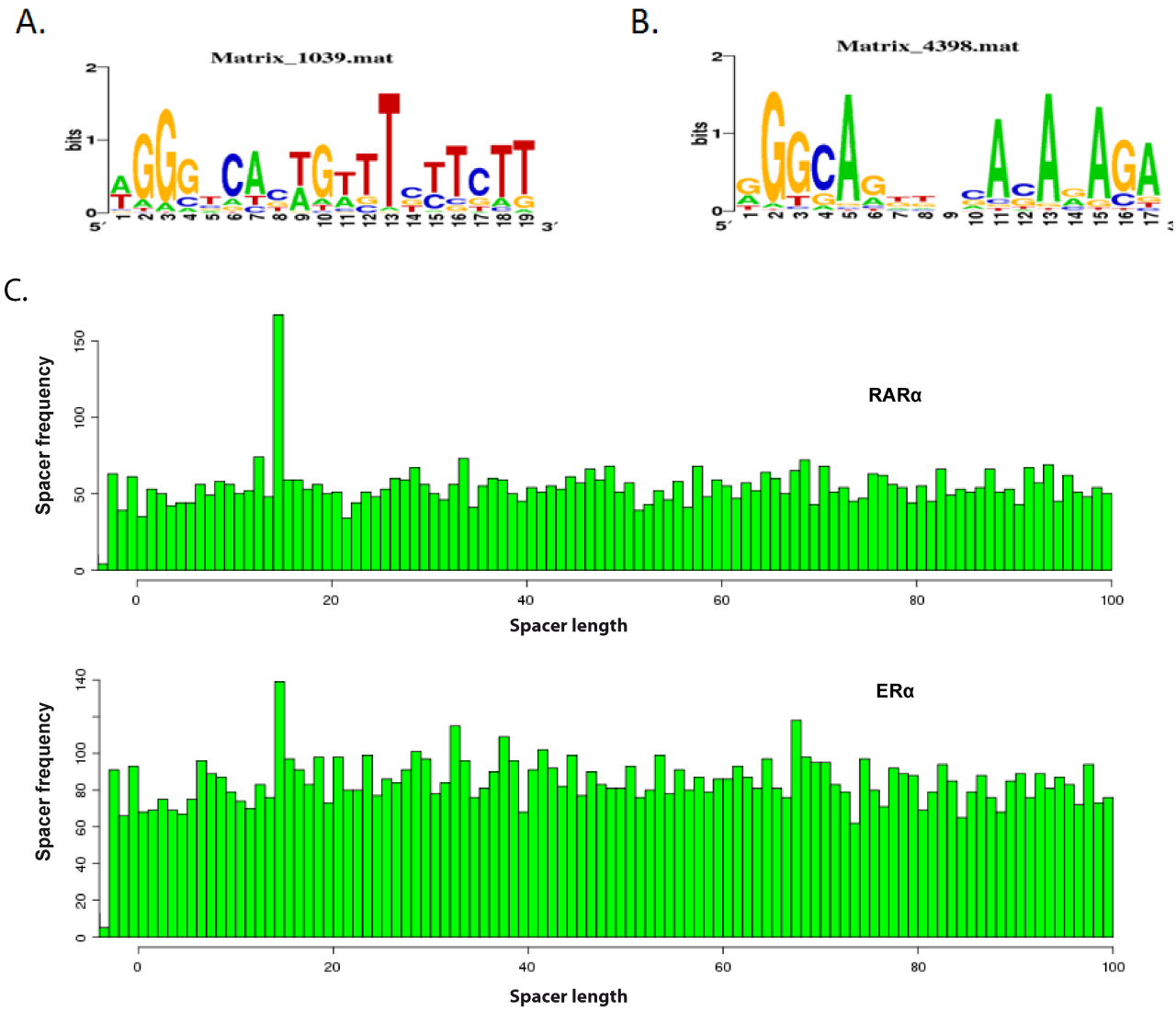


Figure 12. Half ERE and FOXA1 BS predicted in ERα and RARα

(A) and (B) composite motifs predicted in ERα bound regions. (C) Distribution of the spacer length between half ERE and an AT rich motif in a window of 100 bp in both ERα and RARα bound regions.

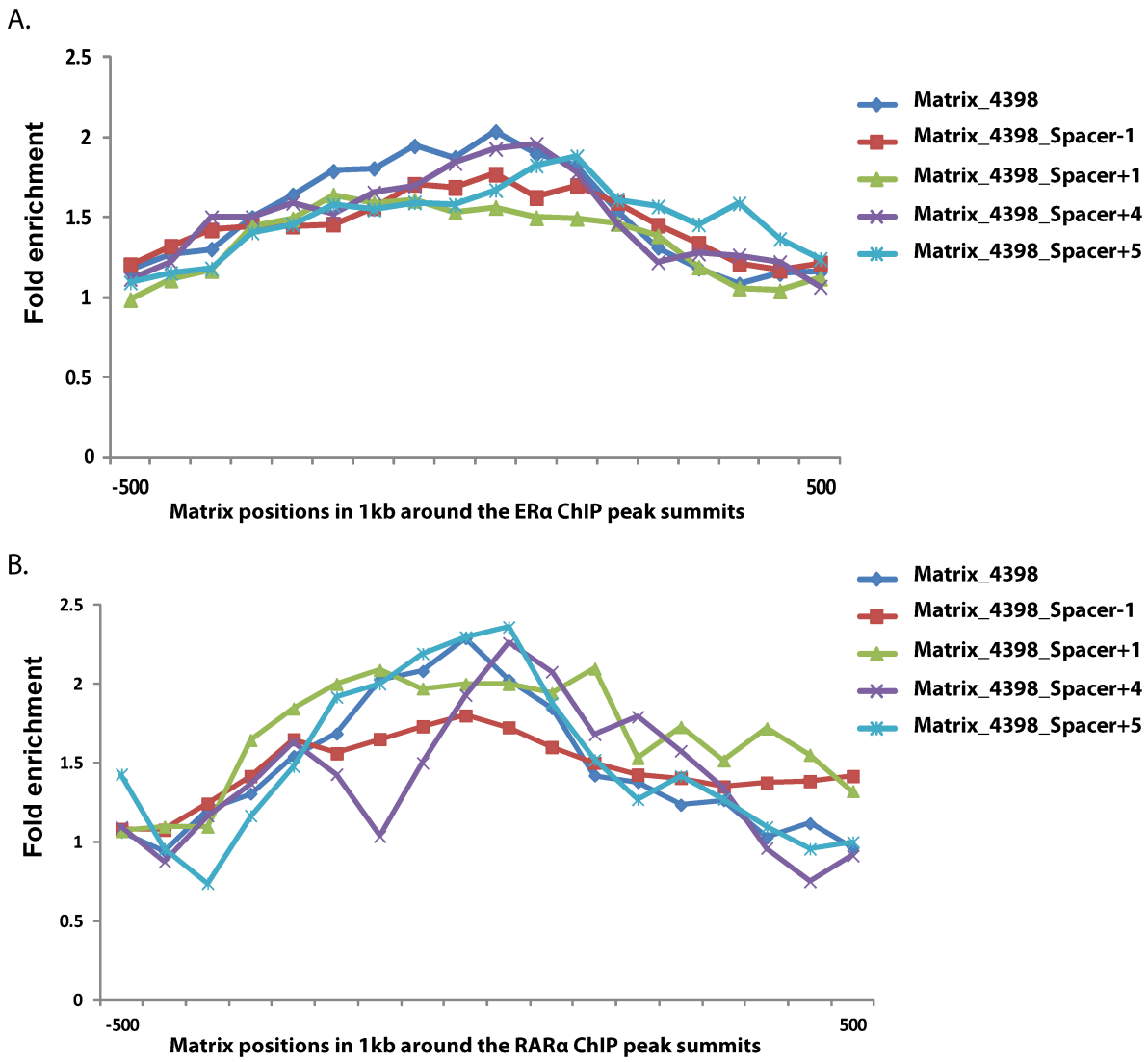


Figure 13. Distribution of a composite motif (half ERE, AT rich motif) and their variations in ERα (A) and RARα (B) bound regions.

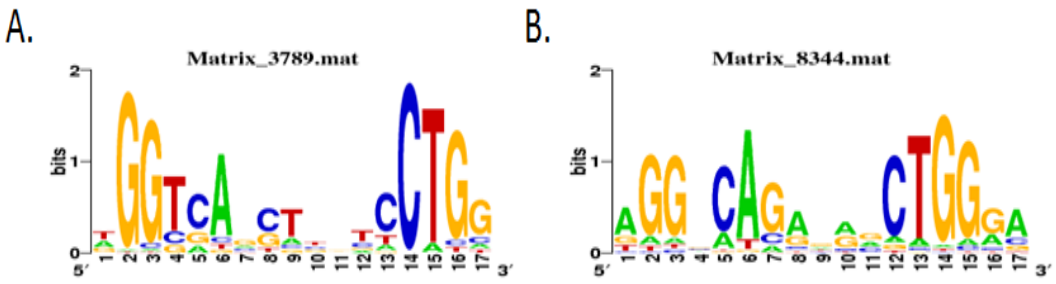
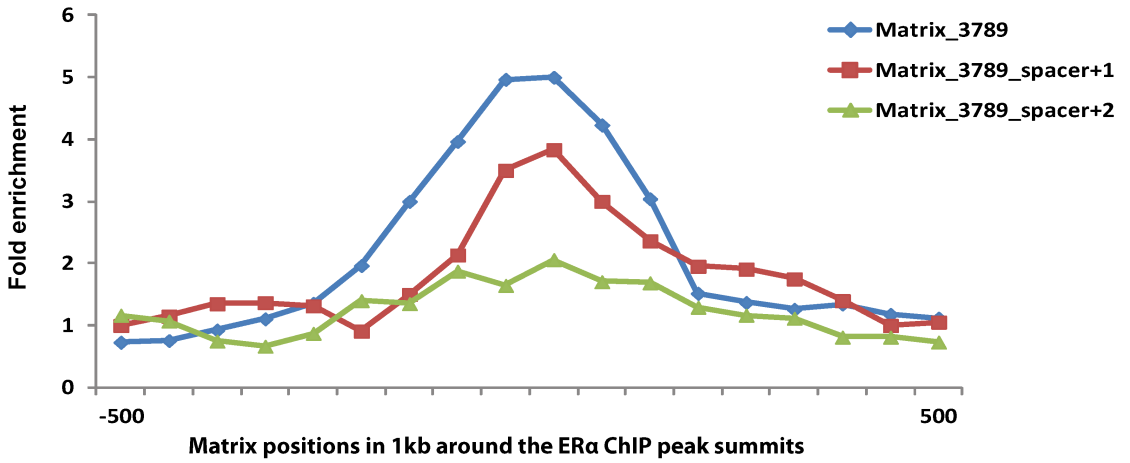


Figure 14. Composite motifs predicted in ERα bound regions.

The two motifs are composed by a half ERE and Ct/aG motif with different spacer length.

A.



B.

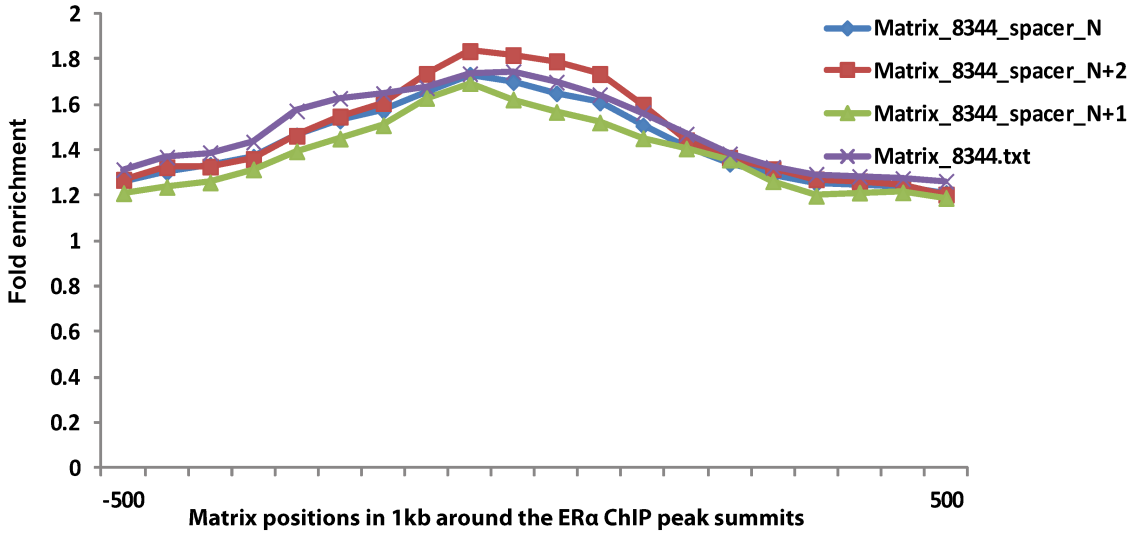


Figure 15. Enrichment of new composite motifs in ERα bound regions.

(A) and (B) represent the distribution of the two composite motifs and their variations (Figure 14) in ERα bound regions for 75% matrix cut-off.

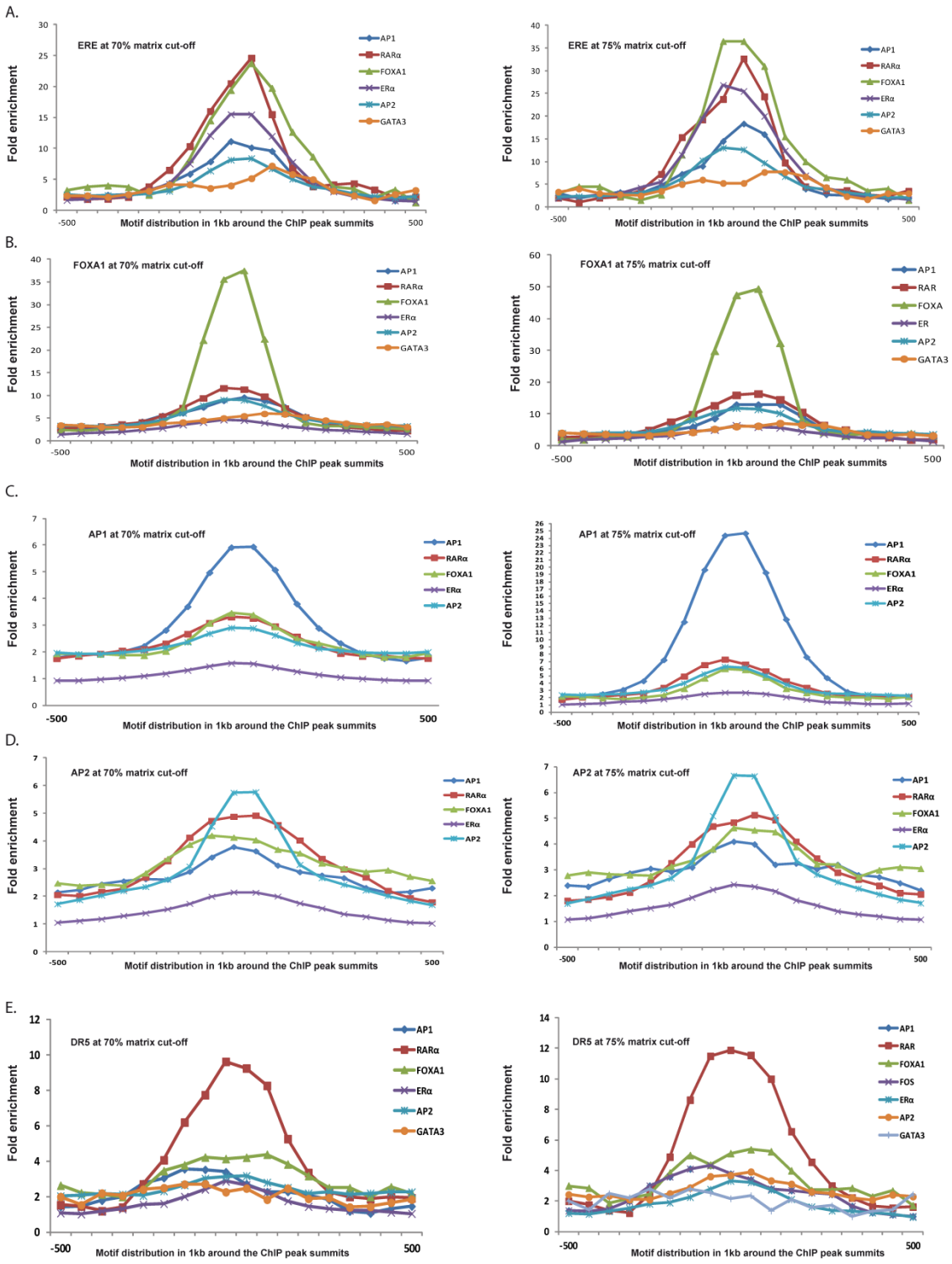


Figure S1. Fold enrichment of TFBS in different datasets for 70 and 75% matrix cut-offs.

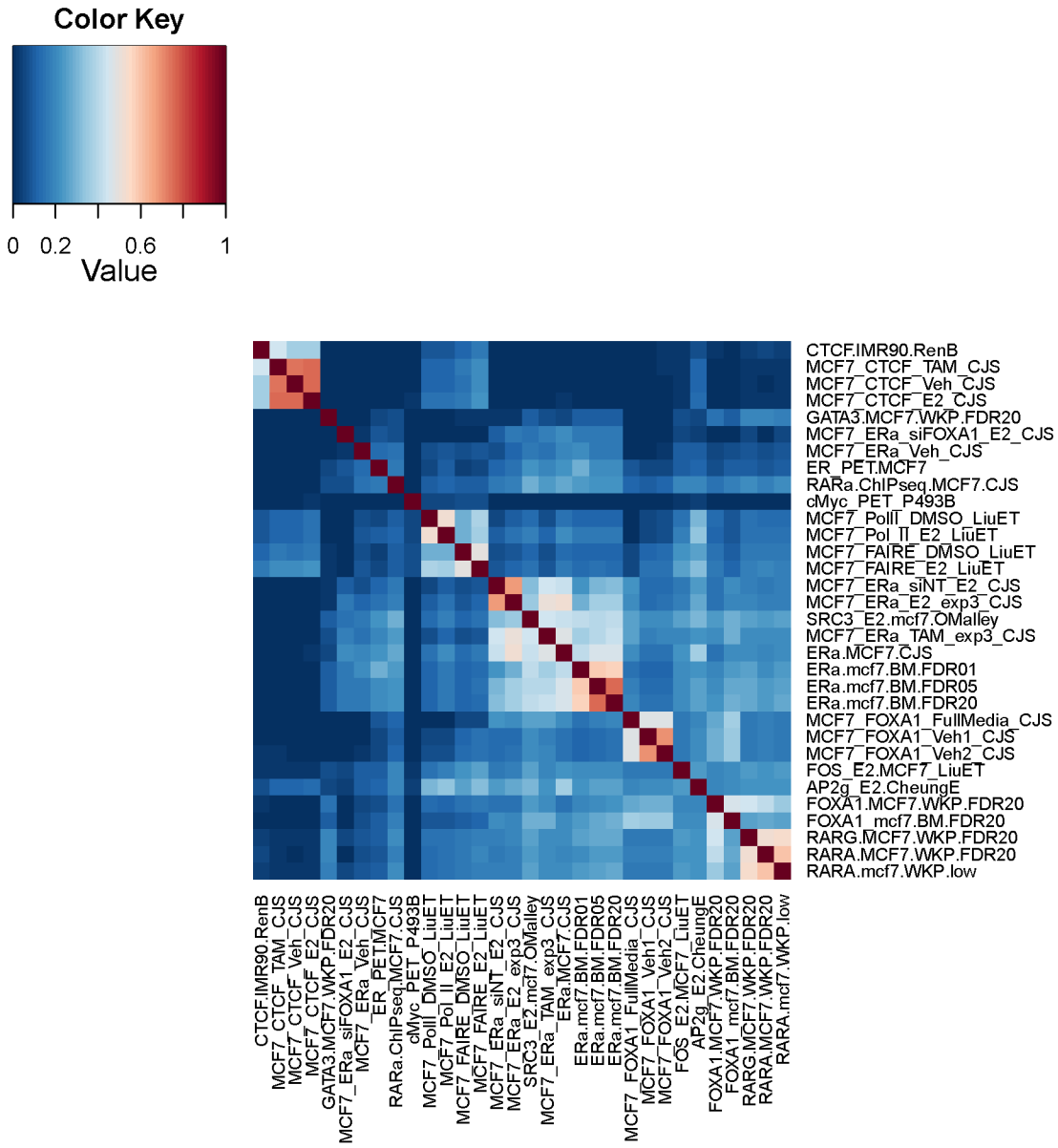


Figure S2. Heatmap representing the correlation between different TF bound regions in MCF7 cells.

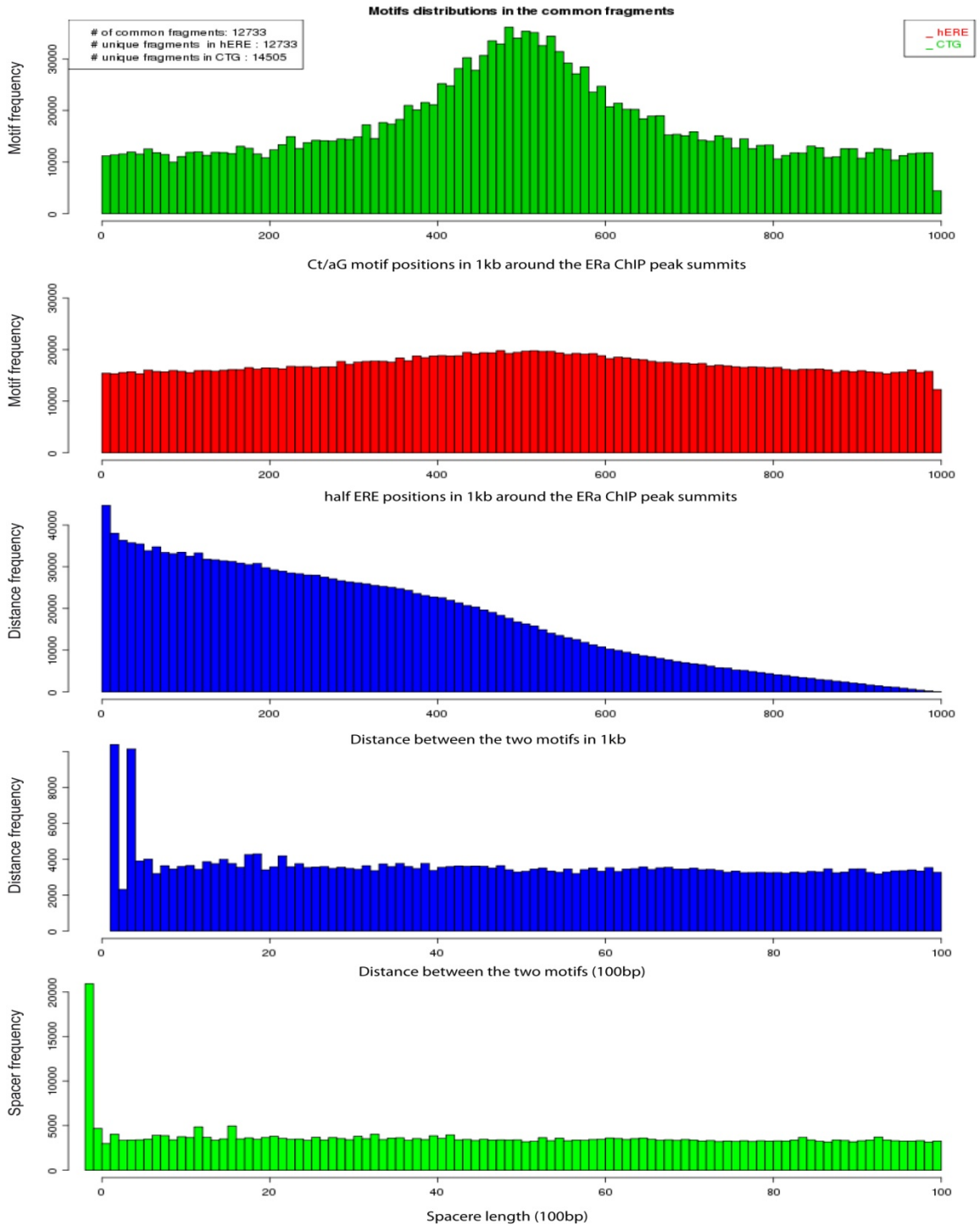


Figure S3. Distribution of half EREs motif and the Ct/aG motif in ERα bound regions

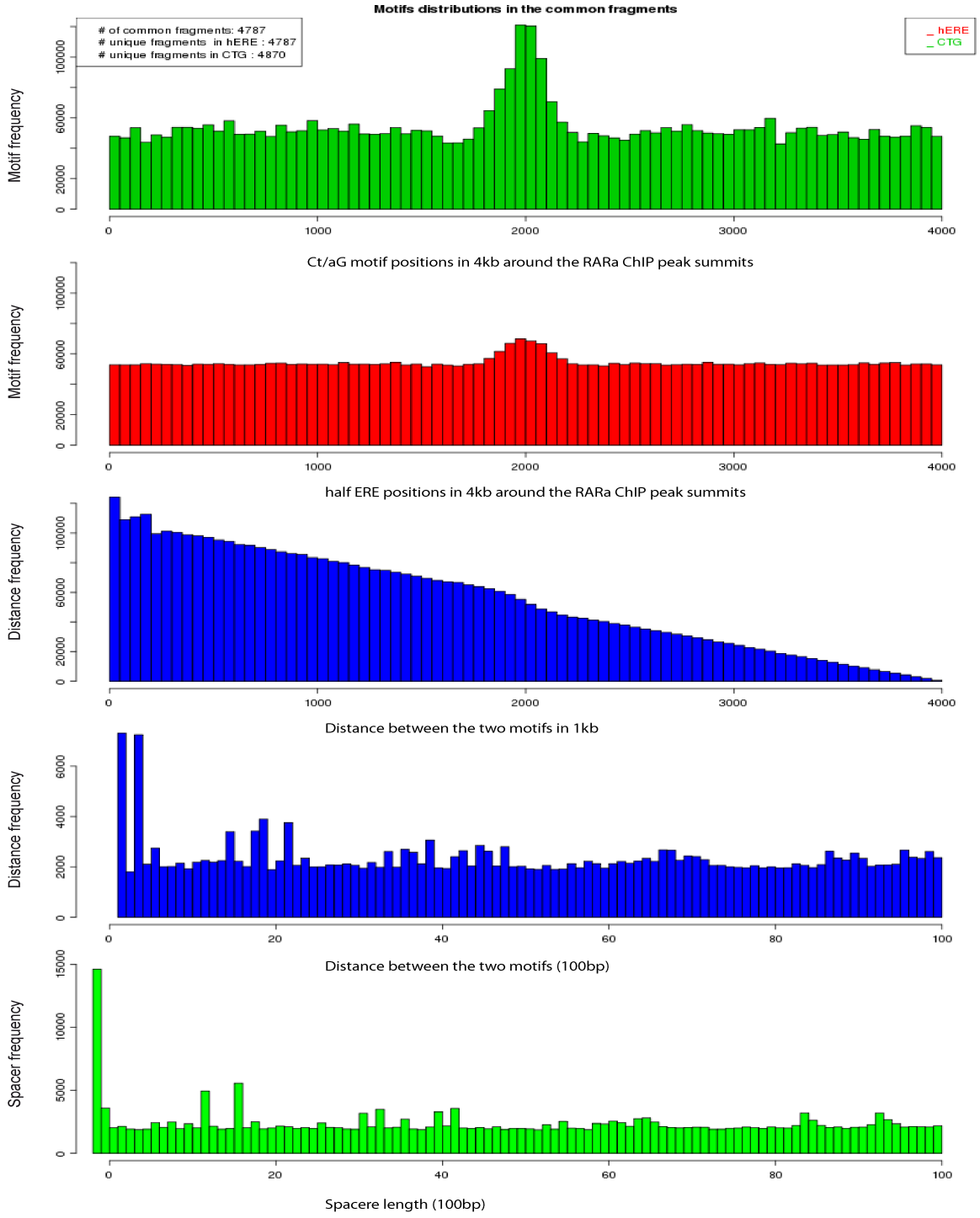


Figure S4. Distribution of half EREs and the Ct/aG motif in RARα bound regions.

Table 1. List of composite motifs predicted by SAMD-ChIP in ERα and RARα bound regions

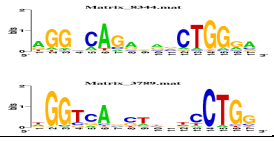
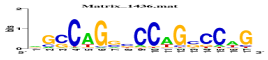
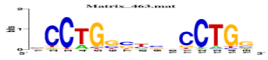
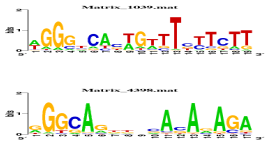
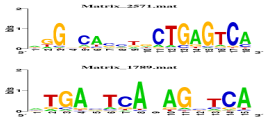
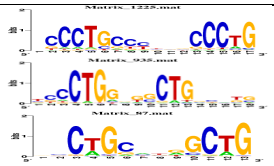
Motif logo	Predicted in
	ERα (half ERE, Ct/aG box)
	ERα (unknown motif)
	ERα (unknown motif)
	ERα (half ERE, Fkh motif)
	ERα, RARa (AP1, half ERE)
	RARα (new motifs)

Table S1. Biological data sets description

TF	# Regions	E2-treated	Experiment
ERα Ross Iness et al., Genes Dev. 2010)	14400	45mn/1h	ChIP-seq
RARα (Ross Iness et al., Genes Dev. 2010)	Top4000	45mn	ChIP-seq
FOS (Josheph R et al., Mol Syst Biol. 2010)	Top4000	3h	ChIP-seq
Foxa1/Gata3/AP2 (Tan et al., EMBO J. 2011)	Top4000	45mn	ChIP-seq

TROISIEME PARTIE : Discussion et perspectives

Chapitre 8

8.1 Questions posées par la caractérisation à grande échelle des sites de fixation de facteurs de transcription

L'accumulation rapide des séquences génomiques de différents organismes et des données d'expression rend l'utilisation des outils bioinformatiques nécessaire pour l'étude des mécanismes de régulation de la transcription. La caractérisation des réseaux de régulation transcriptionnelle par l'identification systématique des sites de fixation à l'ADN de FTs à l'échelle du génome est une étape incontournable afin de mieux comprendre les mécanismes de régulation des gènes par les FTs. Grâce au développement des techniques génomiques à large échelle comme la technique de ChIP-chip et de ChIP-seq, de nombreux FTs ont été ciblés par ces techniques dans différents tissus et sous différentes conditions physiologiques. L'analyse des régions liées nécessite le développement d'outils computationnels afin d'identifier les sites liés par le FT immuno-précipité ainsi que les sites liés par les FTs qui médient sa liaison à l'ADN via des interactions protéines-protéines.

8.1.1 Identification de motifs d'ADN liés par le FT immuno-précipité

Plusieurs algorithmes ont été développés pour répondre à cette question. Bien que certaines de ces approches, par exemple MEME [206], ont été utilisées avec succès, leur performances sont loin d'être satisfaisantes pour l'identification de tous les types de motifs liés par les FTs. En effet, plusieurs paramètres affectent les performances de ces approches à savoir : la taille des jeux de données analysés, la taille des motifs liés, leur composition, leur forme (motif simple ou composite) et leur orientation. Des études de comparaison de ces approches ont montré que la plupart des algorithmes existants montrent des performances satisfaisantes pour identifier les motifs courts (dont la taille est inferieure à 10 bp) et ayant une séquence non dégénérée [106]. Cependant, les performances de ces

outils décroissent rapidement pour les motifs longs, les motifs composites avec espacements, les motifs rares qui se trouvent dans une petite fraction des régions liées et les motifs dégénérés [104, 106]. Notre premier but a été de développer une nouvelle approche de découverte de sites de fixation à l'ADN des FT qui exploite les propriétés statistiques de ces sites dans les régions identifiées par les techniques de ChIP-chip et de ChIP-seq afin de contourner certaines limitations des approches existantes. Cette nouvelle stratégie a montré sa pertinence sur différents jeux de données simulés et des jeux de données de ChIP-chip et de ChIP-seq.

8.1.2 Mécanismes de coopérativité entre facteurs de transcription

L'utilisation de programmes de recherche ou de découverte de sites sur des jeux de données de ChIP à l'échelle génomique révèle souvent l'enrichissement d'une série de sites pour des facteurs de transcription autres que le facteur primaire contre lequel est dirigé l'anticorps utilisé dans la réaction de ChIP. Ces facteurs ont le potentiel de coopérer avec le FT primaire par des interactions protéine-protéine qui peuvent avoir lieu en absence ou présence d'interaction avec l'ADN par le FT primaire. Lorsque ces interactions associent deux facteurs de transcription liés sur l'ADN, elles peuvent prendre place sur de petites distances, à l'intérieur du même fragment de ChIP (par exemple dans le cas de facteurs homo- ou hétéro-dimériques) ou de longues distances par l'intermédiaire de boucles de chromatine. Les sites sont alors localisés sur des fragments de ChIP différents. Il est donc important de comprendre les mécanismes par lesquels ces facteurs agissent pour coopérer avec le FT primaire. La co-localisation de sites pour des FTs différents dans les fragments de ChIP peut dénoter leur association préférentielle à l'intérieur de modules cis-régulateurs [171]. Les interactions entre FTs peuvent stabiliser la liaison à l'ADN des TFs ou conduire à une coopérativité d'activation transcriptionnelle ; cependant une coopérativité peut avoir lieu en absence de liaison simultanée, par exemple dans le cas de facteurs de transcription pionniers dont le rôle est de faciliter la liaison d'autres facteurs de transcription par ouverture de la chromatine [83, 176]. Notre deuxième but a été de développer des

approches permettant d'étudier la co-localisation de motifs d'ADN dans les fragments de ChIP et de déterminer si une distance préférentielle entre les deux sites est observée, suggérant un mécanisme de coopérativité dans la liaison à l'ADN. De plus, il est attendu que les motifs composites résultants soient eux-mêmes enrichis dans le centre des régions de ChIP s'ils médient le recrutement du facteur de transcription primaire.

8.1.3 Biais dû à l'expérimentation et choix des références

Il est important cependant de noter que des artéfacts peuvent être générés dans ce type d'expériences, par exemple si l'anticorps utilisé réagit de manière croisée avec un autre facteur de transcription. Ce risque peut être éliminé par la génération de jeux de données avec des anticorps indépendants, mais aussi par le constat d'enrichissement réciproque de deux FT dans leurs jeux de données respectifs. Dans cette perspective, nous avons utilisés 6 jeux de données générés dans la même lignée cellulaire, les cellules de cancer du sein MCF7, en utilisant des anticorps contre différents TFs dont les sites sont co-enrichis dans leurs jeux de données respectifs afin d'étudier les mécanismes coopératifs dans les conditions introduisant le moins d'artéfacts possible. Cependant une limitation à l'interprétation de données provenant d'expériences réalisées dans des laboratoires indépendants est la variabilité dans les conditions expérimentales utilisées, qui inclut des cultivars différents de la même souche cellulaire, des conditions de culture différentes, des traitements hormonaux pour des temps différents, des conditions de réticulation et d'immuno-précipitation différentes et enfin l'utilisation de plateformes de micropuces ou de séquençage différentes.

8.2 SAMD-ChIP : Nouvelle approche pour la découverte des sites de fixation à l'ADN des facteurs de transcription adaptée aux données de ChIP-chip et de ChIP-séquençage

8.2.1 Performance de SAMD-ChIP sur des jeux de données simulés

L'analyse des données simulées montre la pertinence de notre approche comparée aux outils de découvertes de motifs existants. Les résultats montrent que SAMD-ChIP offre un meilleur compromis entre le taux des faux positifs et de faux négatifs comparé aux autres outils utilisés dans cette analyse (article 1 Figure 4). Ceci est dû à la fonction de score implémentée dans SAMD-ChIP qui permet de sélectionner uniquement les motifs enrichis dans une fenêtre autour du centre des fragments liés. Nous avons aussi remarqué que SAMD-ChIP a prédit plus d'occurrences des motifs implémentés comparé à MEME-ChIP, DREME et TRAWLER. Ceci pourrait être expliqué par la distance utilisée par notre approche pour retrouver les instances similaires, qui est dépendante de la longueur du motif recherché (article 1 Table S2). En effet, cette propriété est très importante en particulier pour les motifs avec espacements ou bien les motifs ayant des positions dégénérées.

8.2.2 Performance de SAMD-ChIP sur un jeu de données de ChIP-chip du facteur de transcription Mrr1p chez Candida albicans.

Cette analyse avait pour objectif d'identifier le site de fixation à l'ADN du FT Mrr1p à partir des données de ChIP-chip réalisé chez Candida albicans. Mrr1p est un FT de la famille des Zinc Cluster, qui lie des motifs d'ADN composés des triplets CGG avec différents espacements et orientations. Cette famille de FTs lie l'ADN de manière constitutive indépendamment des conditions et de l'environnement cellulaire [211]. Le

motif d'ADN lié par le FT Mrr1p n'est cependant pas connu. Pour caractériser le site d'ADN lié par Mrr1p, nous avons appliqué trois outils de découverte de motifs : MEME-ChIP [206], SCOPE [212] et notre approche SAMD-ChIP. Les trois outils ont identifié un motif commun composé d'un triplet CGG, enrichi dans une fenêtre de 200 bp autour du centre des régions liées (annexe 2 Figure 6) comparé à ce qui est attendu aux extrémités de ces mêmes fragments. SAMD-ChIP a prédit d'autres motifs composites dans les régions liées par Mrr1p. Ces motifs sont composés de deux triplets CGG avec différents espacements entre les deux demi sites CGG. Nous avons montré que ces motifs sont enrichis non seulement dans les régions liées par Mrr1p mais aussi dans un groupe de 30 gènes connus pour être liés et régulés par Mrr1p [48] (annexe 3 Figure 1, 2 et 3). Afin de valider statistiquement cet enrichissement nous avons utilisé tous les gènes de Candida albicans comme ensemble de référence. Nous avons observé que les niveaux d'enrichissent de ces motifs sont comparables dans les régions liées par Mrr1p. Ceci suggère que Mrr1p pourrait lier ces motifs avec la même affinité. Ces observations appuient les résultats obtenus à partir de l'analyse des données simulés à savoir l'efficacité de SAMD-ChIP dans l'identification des motifs composites comparés aux autres outils de découverte de motifs testés dans cet article. Une validation expérimentale de la liaison de Mrr1p aux sites prédits pourrait nous renseigner sur les mécanismes de liaison à l'ADN de Mrr1p (liaison directe ou indirecte) et par conséquent contribuer à élucider les mécanismes de régulation de ses gènes cibles. La liaison directe de Mrr1p à ces motifs pourrait être validée par une expérience de ChIP couplée à la PCR quantitative en utilisant des amorces spécifiques aux différents motifs ou régions qui contiennent ces motifs. Afin de vérifier si cette liaison est directe ou indirecte, nous proposons des essais de retard sur gel *in vitro* utilisant la protéine recombinée exprimée chez les bactéries ou bien tout simplement son domaine de liaison à L'ADN (DBD) et des fragments d'ADN contenant les motifs à valider.

8.2.3 Analyse des données de ChIP-chip/seq contre sept FT dans des cellules MCF7 du cancer du sein traitées aux estrogènes (E2)

Nous avons choisi de tester SAMD-ChIP sur un ensemble de jeux de données de ChIP-chip et de ChIP-seq réalisés dans des cellules MCF7 du cancer du sein traitées aux œstrogènes (E2) qui ciblent les FT suivants : ERα, FOXA1, GATA3, AP1, RARα, Myc et AP2.

ERα est surexprimé dans 2/3 des tumeurs mammaires et induit la prolifération cellulaire à travers divers mécanismes. Il interagit avec l'ADN de manière directe en liant son site de fixation ERE. Il peut aussi être recruté sur l'ADN par d'autres FT. Cependant, les mécanismes d'interactions entre ERα et ses partenaires transcriptionnels restent mal définis [212-215]. Le choix de ces FT a pour but de caractériser leurs mécanismes de liaison à l'ADN ainsi que leur rôle dans le recrutement de ERα à l'ADN.

Les motifs d'ADN liés par ces FT sont connus. La composition, la longueur et la structure de ces motifs sont différentes ce qui va nous permettre de comparer les performances de notre approche à celles des outils de découverte de motifs existants en utilisant différents profiles de motifs. Dans cette analyse, nous avons aussi inclus le jeu de données de ChIP-chip contre le FT Myc, qui est composé de 300 régions liées par Myc. Ce jeu de données nous a permis de comparer les performances des différents outils incluant SAMD-ChIP dans la prédiction des motifs rares (motifs présents dans une petite fraction des régions liées). Huas et ses collaborateurs ont utilisé la matrice de TRANSFAC représentant le site d'ADN lié par Myc pour scanner les 300 régions identifiées par ChIP-chip [215]. Ils ont identifié seulement 34 motifs dans une fenêtre de 500 bp autour du centre de ces régions. Ce résultat confirme d'une part la présence du motif lié par Myc dans ces régions et d'autre part sa rareté. Les sites liés par ERα et RARα sont composés de deux demi-sites RGGTCA séparés par différents espacements et organisés en différentes orientations. Dans le cas de ERα l'espacement est de trois nucléotides. Quant à RARα , peut lier deux répétitions directes de ce demi site avec 1, 2 ou bien 5 espacements [214]. Ces formes de motifs nous ont permis de comparer les performances des outils testés pour prédire les motifs avec espacements.

La première analyse a pour but de vérifier si les motifs primaires (motifs liés par les FT immuno-précipités) ont été identifiés par les différents outils testés. Les motifs liés par ERα, FOXA1, AP1, AP2 et GATA3 ont été identifiés par l'ensemble des outils testés dans leurs régions respectives. Cependant, le site lié par Myc ainsi que les sites liés par RARα ont été identifié uniquement par SAMD-ChIP (article 1 Figure 10 et 11). Ces résultats confirment l'efficacité de SAMD-ChIP à identifier les motifs rares et les motifs avec espacements. Le critère de similarité entre séquences implémenté dans SAMD-ChIP est dépendant de la longueur des séquences à comparer. Par conséquent, il offre une flexibilité lors de la recherche des instances similaires pour créer les motifs initiaux. Cette propriété présente l'avantage de ne pas pénaliser les motifs avec espacements contrairement aux fonctions qui implémentent un seuil de similarité fixe ou bien celles basées sur les alignements locaux à l'instar des fonctions implémentées respectivement dans Mdscan [128] et MEME-ChIP [131].

La deuxième analyse avait pour but de comparer les performances de notre approche par rapport à celle des autres outils pour la prédiction de motifs secondaires connus (motifs qui pourraient être liés par des partenaires d'interaction des FT immuno-précipités) ainsi que de nouveaux motifs qui pourraient être des cibles de FT non encore caractérisés. La Figure 11 de l'article 1 montre la liste des motifs prédits dans les régions liées par ERα. Plusieurs de ces motifs ont été identifiés uniquement par SAMD-ChIP. Certains de ces motifs ont été identifiés comme des cibles de FT connus dans la littérature comme des partenaires de ERα, par exemple FOXA1, AP2, AP1, RARα et GATA3. Ceci suggère que ces FT pourraient médier la liaison à l'ADN de ERα via des interactions protéines-protéine.

D'autres motifs ont été prédits par notre approche dans les régions liées par ERα et identifiés comme les sites liés par EBF (early B-cell factor 1), LMAF (v-maf musculoaponeurotic fibrosarcoma) et LBP9 (transcription factor CP2-like 1). Le site lié par EBF a été identifié par SAMD-ChIP et MEME-ChIP, alors que les sites liés par LMAF et LBP9 sont spécifiques à SAMD-ChIP. Ces FT ne sont pas des partenaires connus de ERα, cependant plusieurs études ont montré que ces FT sont surexprimés dans les cancers en

général et sont impliqués dans la prolifération cellulaire [213, 214, 216]. Des analyses supplémentaires seront nécessaires pour identifier les mécanismes par lesquels ces FT interagissent avec ERα.

Plusieurs nouveaux motifs ont été prédit par SAMD-ChIP dans les différents jeux de données analysés (article 1 Table 2 et Table S3 et S4). Parmi ces motifs, un motif palindromique (WGATnnnATCW) prédit dans les régions liées par ERα et GATA3. Nous n'avons pas pu assigner ce motif aux matrices connus dans la littérature. Le niveau d'enrichissement de ce motif dans les régions liées par ERα et GATA3 est deux fois plus élevé que ce qui est observé aléatoirement. Ce motif pourrait donc être la cible d'un FT qui médie la liaison à l'ADN de ERα et de GATA3 via des interactions protéine-protéine. Dans le futur, il serait intéressant d'identifier le(s) FT(s) qui lie(nt) ce motif et de caractériser les mécanismes de recrutement de ERα et GATA3 sur ces sites.

Nous avons aussi identifié un motif composé de deux demi-sites (CWG) avec différents espacements dans plusieurs jeux de données. Les formes avec 1 et 2 espacements ont été prédites aussi par MEME-ChIP (article 1 Figure 12) dans les régions liées par ERα. D'autres formes avec 3, 4 et 6 espacements sont spécifiques à SAMD-ChIP (article 1 Table2, S3 et S4) et ont été identifié dans les régions liées par ERα et RARα. La fraction des régions qui contiennent au moins une occurrence de ces motifs varie de 20 à 40% dans les deux jeux de données. Cependant, nous n'avons pas observé un enrichissement plus élevé pour une forme spécifique comparé aux autres. Ceci suggère que ce profile de motif pourrait être lié par un FT qui peut lier avec des affinités comparables, les deux demi sites (CWG) séparés par des espacements de longueur variable.

En utilisant notre approche, nous avons aussi identifié un profile de motif composé d'un demi ERE (RGGTCA) et du motif (CWG) dans les régions liées par ERα, RARα, GATA3 et FOS (article 1 Table2). Cependant, nous n'avons pas observé une longueur d'espacement préférée entre le demi-ERE et le motif CWG. La fraction des régions liées par RARα et ERα et qui contiennent au moins une occurrence de ces motifs varie de 5 à 10% suivant la longueur du motif. Les niveaux d'enrichissements sont autour de 2 fois comparés à ce qui

est attendu aléatoirement dans les régions liées par RARα et supérieur à 5 fois dans les régions liées par ERα. Ce résultat suggère l'association entre ERα et le FT liant les sites CWG sur ces motifs composites. Cependant leur faible enrichissement ne suggère pas une forte coopérativité de liaison.

8.2.4 SAMD-ChIP : améliorations et extensions …..

- Dans la version actuelle de SAMD-ChIP, l'instance ayant le meilleur score selon la matrice est sélectionnée (l'instance la plus proche de la forme consensus du motif représenté par cette matrice). Cependant, il arrive qu'un FT lie des formes dégénérées de son site de fixation à l'ADN. Par conséquent en choisissant la meilleure instance par séquence on pénalise les sites de faible affinité. Il serait donc intéressant d'explorer les avantages/inconvénients liés à la sélection de plus d'une instance par séquence lors des étapes d'initialisation et d'optimisation des matrices.

- Dans SAMD-ChIP, l'initialisation des matrices se fait soit en utilisant l'ensemble des séquences à analyser ou bien en choisissant un sous ensemble de séquences de manière aléatoire. Dans ce dernier cas, on présume que si on réalise cet échantillonnage deux fois de suite, on obtiendrait certaines différences entre les deux analyses. Nous pensons qu'il serait intéressant d'implémenter un module supplémentaire dans SAMD-ChIP qui permet de lancer l'étape d'initialisation plus d'une fois, ensuite de comparer les résultats de ces différentes analyses afin d'éliminer les redondances avant de procéder aux étapes suivantes.

- Une autre amélioration qui pourrait être implémentée dans SAMD-ChIP serait l'ajout d'un module intégrant les résultats de la découverte de motifs avec les données d'expression des gènes dans la même lignée et sous les mêmes conditions expérimentales. Ceci permettrait de caractériser le rôle des motifs enrichis dans la régulation des gènes cibles situés en cis (activation ou répression de la transcription).

- Dans la section introduction, nous avons discuté de l'importance d'intégrer d'autres sources d'informations comme les marques de chromatine, le positionnement des nucléosomes afin d'améliorer la qualité des prédictions des outils de découverte de motifs

liés par les FTs. Dans le cas où ces informations sont disponibles dans la même lignée cellulaire et sous les mêmes conditions expérimentales dans lesquelles un FT a été ciblé par ChIP-chip/seq, il serait intéressant de développer un modèle statistique qui permettrait d'intégrer les résultats de SAMD-ChIP avec les données sur les marques de chromatine active/inactive et le positionnement des nucléosomes pour éliminer certains motifs qui ont été prédit par cause des biais de l'approche utilisée.

- Notre approche, SAMD-ChIP, permet de prédire la liste des motifs enrichis autour du centre des régions de pics pour l'intervalle de longueur de motifs définis par l'utilisateur. Par conséquent, il arrive que le même motif soit prédit pour des longueurs différentes. L'identification de la longueur optimale d'un motif est une étape importante afin de choisir les motifs à valider expérimentalement et permet de nous renseigner sur la longueur et la composition des motifs liés par les FTs. Nous avons examiné l'impact de la longueur du motif sur les niveaux d'enrichissement observés en utilisant le site ERE (motif lié par ERα) comme contrôle. Ce site est un motif palindromique de 15 bp et est effectivement le motif de 15 bp présentant le niveau d'enrichissement le plus élevé. Cependant, l'analyse des motifs ERE de 6 bp à 20 bp montre que le niveau d'enrichissement augmente avec la longueur sans présenter un plateau pour le motif à 15 bp. Les motifs de 16 bp et plus, qui sont formés d'un ERE étendus de quelques nucléotides, présentent des niveaux d'enrichissement plus élevés que le motif à 15 bp. Ceci est possiblement dû à la rareté croissante des motifs de plus grande longueur dans l'ensemble de référence, et par conséquent son niveau d'enrichissement plus élevé. Pour l'instant, notre approche permet donc de prédire la liste des motifs enrichis pour toutes les longueurs de motifs sélectionnées par l'utilisateur mais ne nous renseigne pas sur la longueur optimale du motif lié par un FT.

- Nécessité de mieux comprendre les mécanismes de coopérativité entre FTs : module de co-localisation, voir ci-dessous.

8.3 Inférence des mécanismes de coopérativité entre FT à partir des données de ChIP-chip et de ChIP-séquençage

8.3.1 Limitations des programmes de découverte de motifs dans le cas des motifs composites ou contenant un fort pourcentage de positions sans contraintes

Dans le cadre du premier article, nous avons développé SAMD-ChIP, une nouvelle approche pour la prédiction des sites liés par les FT dans les régions de ChIP-chip et de ChIP-séquençage. Les résultats obtenus par analyses de plusieurs jeux de données ont montré que notre approche offre une bonne performance pour l'identification de motifs composites tels que les sites reconnus par des homo- ou hétéro-dimères de récepteurs nucléaires. Cependant notre programme ne permet pas de déterminer la longueur optimale d'un motif, et sa performance diminue pour les motifs dont la longueur est supérieure à 20 bp, et ce particulièrement pour les motifs composites dans lesquels l'espacement est composé de positions dégénérées. En effet, le taux de similitude devient alors plus faible en proportion de la longueur totale de la séquence. Cette limitation est commune d'ailleurs à tous les outils de découvertes de motifs.

Les motifs composites représentent un cas extrême de coopération entre FT, celui dans lequel la liaison à l'ADN se fait de manière coopérative. De manière plus générale, il est pertinent de se poser la question si les motifs enrichis dans les données de ChIP-chip ou ChIP-seq sont retrouvés de manière statistiquement significative sur les mêmes motifs, avec ou non un espacement fixe. En effet, une co-localisation des motifs pourrait signifier une coopérativité à un autre niveau, par exemple dans le cas ou un facteur permet par une activité d'ouverture de la chromatine de faciliter la liaison du FT primaire.

8.3.2 Module de co-localisation de motifs dans les jeux de données de ChIP

Etant donné l'importance de mieux comprendre les mécanismes de coopérativité entre FTs, nous avons développé un module qui permet de tester à partir des résultats de la découverte de motifs l'éventuelle co-localisation de motifs dans les fragments de ChIP. Ce module prend comme entrée deux motifs représentés par leurs matrices respectives, prédites par un outil de découverte de motifs ou bien sélectionnées à partir de banques de matrices, et un ensemble de séquences représentant les régions liées par le FT d'intérêt ; ces séquences d'une longueur prédéfinie par l'usager sont extraites de manière centrée autour du pic des régions de ChIP. La première étape consiste à scanner ces séquences en utilisant les matrices pour différents seuils (fixés par l'utilisateur). La deuxième étape consiste à estimer la distance ainsi que l'espacement entre chaque paire d'instances des matrices recherchées de manière à estimer si leur distribution est aléatoire ou fixe. La significativité statistique des maxima observés est estimée par un test de permutation sous le langage R qui vérifie si cette distance pourrait être observée aléatoirement. Cette recherche est effectuée en prenant en considération les trois arrangements possibles de ces séquences, i.e. en répétition directe, palindromique ou palindromique inverse.

Nous avons validé cette approche en estimant la longueur et l'espacement optimaux pour les motifs qui composent le site lié par le récepteur des estrogènes ERα. Nous avons effectivement identifié l'arrangement palindromique de 15 paires de bases, avec un espacement de 3 nucléotides, comme arrangement optimal dans les jeux de données de ERα. Le module vérifie ensuite que cet arrangement est effectivement enrichi dans le centre des régions de ChIP, tel qu'attendu pour une coopérativité au niveau de la liaison à l'ADN *in vivo*.

Une absence de co-localisation avec des motifs ½ERE a été aussi observée sauf dans le cas de motifs ½ ERE et AP1 (voir ci-dessous).

8.3.3 Analyse des mécanismes de coopérativité entre FTs dans les données de ChIP de ERα.

Contrairement aux ½EREs, nous n'avons pas observé de co-localisation entre deux séquences EREs, indiquant qu'un ERE est suffisant pour générer un fragment de ChIP. Ce résultat est à contraster avec le constat de l'enrichissement de sites EREs dans les promoteurs de gènes induits par les œstrogènes [217]. En effet, plusieurs EREs peuvent coopérer pour l'activation transcriptionnelle mais par l'intermédiaire de boucles de chromatine [218, 219]. Dans ce cas, chaque site se retrouve associé à une région de ChIP individuelle.

Nous n'avons pas non plus observé de co-localisation significative des sites pour les facteurs de transcription enrichis dans les régions de ChIP de ERα et les EREs. Il est possible également que les sites enrichis dans les données de ChIP de ERα coopèrent avec des EREs par la formation de boucles de chromatine. Il est en effet attendu que l'ensemble des sites des FTs avec lesquels ERα peut interagir de manière stable dans les conditions de réticulation utilisées soient détectés comme sites associés à ERα de manière indirecte. Il serait ainsi intéressant de déterminer si une co-localisation significative est observée entre un ERE et les sites enrichis dans les promoteurs des gènes cibles des œstrogènes, qu'il s'agisse de gènes induits ou réprimés. Il est cependant possible que ces interactions aient lieu entre promoteurs de gènes cibles des œstrogènes, ou que l'interaction avec le FT soit suffisamment forte pour recruter ERα sur son site en absence totale d'interaction avec un ERE. Dans tous les cas, ces interactions devraient être plus sensibles à la dose d'agents réticulant utilisés car elles nécessitent deux évènements de réticulation plutôt qu'un seul dans le cas de la liaison de ERα à un ERE.

8.3.4 Enrichissement réciproque entre FTs dans leurs données de ChIP respectives.

De manière frappante, les EREs sont retrouvés fortement enrichis dans les jeux de données de ChIP de plusieurs des FTs dont les sites sont enrichis, bien que proportionnellement plus faiblement, dans les jeux de données de ERα. Cette observation démontre que la liaison à l'ADN de ERα est possible dans le cadre d'interaction avec ces FTs et appuie la possibilité d'interactions réciproques dans lesquelles ERα lié sur son site interagit avec un FT lié sur le sien par la formation de boucles de chromatine. Le fort enrichissement d'EREs dans les séquences de ChIP des autres FTs comparativement à l'enrichissement relativement faible des sites liés par les autres FTs dans les régions de ChIP de ERα peut refléter un biais dû aux conditions expérimentales (réticulation plus faible dans le jeu de données de ERα par exemple), ou possiblement une affinité de ERα plus grande pour son site que celle des autres FTs pour le leur, résultant en une asymétrie des interactions avec l'ADN détectées. Ces hypothèses devraient être examinées en comparant l'interaction réciproque des FTs avec leurs sites respectifs dans des expériences de ChIP réalisées en parallèle avec des anticorps (si possible plusieurs) contre ERα et les autres facteurs de transcription.

L'analyse des régions liées par ERα et RARα a démontré la présence de plusieurs sites composites qui pourraient suggérer des coopérations entre ces FTs. Certains de ces derniers sont connus comme des partenaires transcriptionnels de ERα ou de RARα à l'instar de AP1 et de FOXA1. Mais les mécanismes avec lesquels ces FT interagissent avec ERα et RARα ne sont pas totalement élucidés. D'autres motifs composites font intervenir des motifs que nous n'avons pas pu identifier dans les banques de matrices de TRANSFAC et de JASPAR. Ceci suggère que ces motifs pourraient être liés conjointement par ERα/RARα et un nouveau partenaire qui reste à déterminer. En utilisant le module développé dans le cadre de cet article, nous avons vérifié la possibilité d'une coopération entre ERα/RARα avec AP1 et FOXA1 respectivement. En effet, plusieurs études ont démontré le rôle de FOXA1 dans le recrutement de ERα à l'ADN [77, 220]. Certaines de ces études ont suggéré que

FOXA1 est d'abord recruté sur l'ADN pour ouvrir la chromatine et faciliter le recrutement de ERα à ses sites ERE.

8.3.5 Rôle de séquences répétées dans la co-localisation de motifs.

Nous avons observé un enrichissement faible mais significatif de motifs composés d'un demi-ERE et d'un site AP1 dans les jeux de données de ERα et RARα. Notre module de co-localisation indique que ces motifs sont effectivement co-localisés dans les régions de ChIP de ces jeux de données. De manière inattendue, la distribution de la longueur de l'espacement à montré un pic à 46 et à 48 pb entre les deux sites. Afin de vérifier si les motifs composés d'un demi-ERE et d'un site AP1 avec un espacement de 46 et 48 respectivement, sont enrichis dans ces régions, nous avons construit deux matrices représentants ces deux motifs, et avons recherché ces matrices dans les régions liées par ERα et RARα. Cette analyse n'a pas révélé un enrichissement de ces deux matrices dans le centre des régions de ChIP des jeux de données analysés. La caractérisation des régions liées indique leur présence dans les séquences répétées de type Alu. Ceci explique l'absence d'enrichissement dans le centre des régions de ChIP car ces séquences sont évitées dans les oligonucléotides utilisés comme sondes pour les micropuces et également non cartographiées dans les produits de séquençage dans les expériences de ChIP-seq. Cette limitation ne permet donc pas de conclure en ce qui concerne le rôle potentiel de ces séquences dans le recrutement d'ERα ou de RARα, et indiquent simplement que ces séquences peuvent être retrouvées fréquemment à proximité des pics apparents de ChIP de ERα et RARα. Des oligonucléotides ciblant des régions spécifiques situées de part et d'autre des séquences Alu pourraient permettre de clarifier leur rôle potentiel dans la liaison de ERα, RARα et AP1.

8.4 Interprétation du rôle biologique des motifs prédits

8.4.1 Perspectives

Les résultats obtenus dans le cadre de nos analyses montrent que les FTs testés présentent un potentiel de coopération pour médier leur liaison à l'ADN. Dans le futur, il serait intéressant de tester expérimentalement le rôle des FTs secondaires prédits pour coopérer avec le FT primaire afin de caractériser les mécanismes d'interaction entre eux. Pour vérifier la co-liaison de ces TFs sur sites on suggère des expériences de re-ChIP. Nous proposons aussi de déterminer l'association réciproque de sites dans les promoteurs de gènes cibles de ERα ainsi que de vérifier la formation de boucles de chromatine entre les sites distants. Nous proposons aussi de vérifier l'impact de la suppression ou de la surexpression d'un facteur de transcription sur le recrutement de l'autre : modèle de coopérativité de liaison à l'ADN.

Finalement, on propose de tester l'impact de la présence de ces sites sur l'activation ou la répression transcriptionnelles. En présence des données d'expression des gènes dans les mêmes conditions expérimentales, il est possible d'utiliser des modèles de régression linéaire [221] pour associer le profile spatial et temporel d'expression des gènes à la présence de certains sites fixés par les FTs dans leur promoteurs.

8.5 Conclusion

Dans le cadre cette thèse, nous avons proposé SAMD-ChIP une nouvelle approche pour la découverte des sites de fixation à l'ADN des FTs. Notre approche est spécifique aux données issues des expériences de ChIP-chip et de ChIP-séquençage. Le processus de découverte de motifs selon SAMD-ChIP se fait en plusieurs étapes (article 1 Figure 1), qui sont : soumission des données à analyser, extraction des motifs enrichis dans une fenêtre autour du centre des fragments de ChIP-hip/seq, création des modèles initiaux à partir de ces motifs en utilisant une représentation par matrices de poids de positions (PWM), optimisation des motifs initiaux, regroupement des motifs similaires pour éliminer les

redondances et enfin création d'un fichier de sortie sous un format de page web qui contient les résultats détaillés de l'analyse. Le résultat final est présenté sous forme liste composée des motifs enrichis autour du centre des régions liées pour chaque longueur de motif. Dans chaque liste les motifs sont triés selon leur signal d'enrichissement.

L'algorithme d'extraction des motifs, proposé dans le cadre de cet article implémente une approche mixtes qui combine les stratégies énumérative et probabiliste [106, 129, 146]. Ce choix a été motivé par les limites des algorithmes qui ont implémenté l'une ou l'autre de ces stratégies séparément. En effet, les algorithmes basés sur une approche énumérative présente le désavantage d'être gourmand en temps de calculs et génèrent souvent des résultats redondants à cause de la stratégie exhaustive utilisée [106]. Les algorithmes probabilistes, quant à eux adoptent l'une des deux stratégies EM [106, 146] ou Gibbs sampling [129] pour échantillonner un sous ensemble de données à partir des données à analyser mais souffrent de la convergence prématurée des motifs prédits vers des optimums locaux. Par conséquent, la solution optimale est rarement atteinte.

L'approche proposée dans le cadre du premier article présente l'avantage de prédire les motifs simples liés par les FTs et des motifs composites qui pourraient suggérer des coopérations entre FTs pour lier l'ADN. Notre approche se limite à tester les motifs présents dans une fenêtre autour du centre des régions liées par les FTs. Ceci réduit considérablement son temps de calculs ainsi que le bruit, introduit en général, par les séquences avoisinantes. L'évaluation statistique de l'enrichissement des motifs prédits présente un vrai défit aux approches de découverte de motifs. Dans SAMD-ChIP, nous avons implémenté deux niveaux d'évaluation. Le premier consiste à mesurer le niveau d'enrichissement des motifs prédits dans la fenêtre autour du centre des régions liées. Seuls les motifs ayant un niveau d'enrichissement supérieur au seuil fixé par l'utilisateur seront considérés. Le deuxième teste consiste à vérifier si la distribution des occurrences de chaque motif est différente de la distribution uniforme (ce qui est attendu aléatoirement).

Ici, il est important de s'attarder sur le choix de l'ensemble de référence utilisé pour mesurer cet enrichissement. En effet dans le cadre de ce travail nous avons testé différents ensembles de référence et montré que les résultats obtenus sont dépendants de l'ensemble de référence utilisé (article 1 Table 1). Suite aux différentes analyses effectuées dans le cadre de ce travail, nous suggérons l'utilisation des régions randomisées comme ensemble de référence. Ce dernier est obtenu par permutation des positions des nucléotides dans chacun des fragments à analyser tout en permettant la conservation des blocs de nucléotides d'un certain ordre (par exemple un ordre de trois permet de sauvegarder la composition des blocs de trois nucléotides).

L'analyse de la co-localisation de motifs dans les jeux de données de ChIP est essentielle pour mieux comprendre la régulation de l'expression des gènes par les FTs. Cet objectif nécessite le développement de nouveaux algorithmes et méthodologies d'analyses prenant en compte des données complexes comme le nombre et la distance entre les sites liés par les FTs. Le module de co-localisation proposé dans le cadre du second article permettra d'inférer d'éventuelles co-localisations et interactions entre FTs dans les régions liées. Il est également applicable à l'analyse de sites composites liés par des complexes de FTs tels que les homo- ou hétéro-dimères de récepteurs nucléaires.

L'ensemble de ces travaux contribuera à mieux comprendre les mécanismes de liaison à l'ADN des FTs et par conséquent aider à la construction des réseaux de régulation des gènes, étape cruciale pour la modélisation des systèmes biologiques. Dans les cas des maladies comme le cancer, une meilleure connaissance de ces réseaux pourrait permettre d'identifier de nouvelles cibles contribuant aux signaux mitogènes sur lesquelles il convient d'agir afin de proposer des traitements plus efficaces.

Bibliographie

1. Cole, C.G., et al., *Finishing the finished human chromosome 22 sequence.* Genome Biol, 2008. **9**(5): p. 13.
2. Arents, G., et al., *The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left-handed superhelix.* Proc Natl Acad Sci U S A, 1991. **88**(22): p. 10148-52.
3. Richmond, T.J. and C.A. Davey, *The structure of DNA in the nucleosome core.* Nature, 2003. **423**(6936): p. 145-50.
4. Strahl, B.D. and C.D. Allis, *The language of covalent histone modifications.* Nature, 2000. **403**(6765): p. 41-5.
5. Sun, F.L., M.H. Cuaycong, and S.C. Elgin, *Long-range nucleosome ordering is associated with gene silencing in Drosophila melanogaster pericentric heterochromatin.* Mol Cell Biol, 2001. **21**(8): p. 2867-79.
6. Grigoryev, S.A., et al., *Evidence for heteromorphic chromatin fibers from analysis of nucleosome interactions.* Proc Natl Acad Sci U S A, 2009. **106**(32): p. 13317-22.
7. Levine, M. and R. Tjian, *Transcription regulation and animal diversity.* Nature, 2003. **424**(6945): p. 147-51.
8. Lemon, B. and R. Tjian, *Orchestrated response: a symphony of transcription factors for gene control.* Genes Dev, 2000. **14**(20): p. 2551-69.
9. Thomas, M., J. Lieberman, and A. Lal, *Desperately seeking microRNA targets.* Nat Struct Mol Biol, 2010. **17**(10): p. 1169-74.
10. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.
11. Lockhart, D.J., et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nat Biotechnol, 1996. **14**(13): p. 1675-80.
12. Orphanides, G., T. Lagrange, and D. Reinberg, *The general transcription factors of RNA polymerase II.* Genes Dev, 1996. **10**(21): p. 2657-83.
13. Stormo, G.D., *DNA binding sites: representation and discovery.* Bioinformatics, 2000. **16**(1): p. 16-23.
14. Maston, G.A., S.K. Evans, and M.R. Green, *Transcriptional regulatory elements in the human genome.* Annu Rev Genomics Hum Genet, 2006. **7**: p. 29-59.
15. Blackwell, T.K. and A.K. Walker, *Transcription mechanisms.* WormBook, 2006. **5**: p. 1-16.
16. Luscombe, N.M., et al., *An overview of the structures of protein-DNA complexes.* Genome Biol, 2000. **1**(1): p. 9.
17. Hannenhalli, S., *Eukaryotic transcription factor binding sites--modeling and integrative search methods.* Bioinformatics, 2008. **24**(11): p. 1325-31.
18. Kummerfeld, S.K. and S.A. Teichmann, *DBD: a transcription factor prediction database.* Nucleic Acids Res, 2006. **34**(Database issue): p. D74-81.
19. Gray, P.A., et al., *Mouse brain organization revealed through direct genome-scale TF expression analysis.* Science, 2004. **306**(5705): p. 2255-7.
20. Vaquerizas, J.M., et al., *A census of human transcription factors: function, expression and evolution.* Nat Rev Genet, 2009. **10**(4): p. 252-63.
21. Germain, P., et al., *Overview of nomenclature of nuclear receptors.* Pharmacol Rev, 2006. **58**(4): p. 685-704.

22. Glass, C.K., *Differential recognition of target genes by nuclear receptor monomers, dimers, and heterodimers.* Endocr Rev, 1994. **15**(3): p. 391-407.
23. Olefsky, J.M., *Nuclear Receptor Minireview Series.* J. Biol. Chem., 2001. **276**(40): p. 36863-36864.
24. Nguyen, D., et al., *A G577R mutation in the human AR P box results in selective decreases in DNA binding and in partial androgen insensitivity syndrome.* Mol Endocrinol, 2001. **15**(10): p. 1790-802.
25. Carroll, J.S., et al., *Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1.* Cell, 2005. **122**(1): p. 33-43.
26. Bai, Z. and R. Gust, *Breast cancer, estrogen receptor and ligands.* Arch Pharm, 2009. **342**(3): p. 133-49.
27. Villard, J., *Transcription regulation and human diseases.* Swiss Med Wkly, 2004. **134**(39-40): p. 571-9.
28. Reece-Hoyes, J.S., et al., *A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks.* Genome Biol, 2005. **6**(13): p. 30.
29. Miele, A. and J. Dekker, *Long-range chromosomal interactions and gene regulation.* Mol Biosyst, 2008. **4**(11): p. 1046-57.
30. Gill, G., *Regulation of the initiation of eukaryotic transcription.* Essays Biochem, 2001. **37**: p. 33-43.
31. Thomas, M.C. and C.M. Chiang, *The general transcription machinery and general cofactors.* Crit Rev Biochem Mol Biol, 2006. **41**(3): p. 105-78.
32. Gill, G., et al., *A glutamine-rich hydrophobic patch in transcription factor Sp1 contacts the dTAFII110 component of the Drosophila TFIID complex and mediates transcriptional activation.* Proc Natl Acad Sci U S A, 1994. **91**(1): p. 192-6.
33. Doetzlhofer, A., et al., *Histone deacetylase 1 can repress transcription by binding to Sp1.* Mol Cell Biol, 1999. **19**(8): p. 5504-11.
34. Bulger, M. and M. Groudine, *Looping versus linking: toward a model for long-distance gene activation.* Genes Dev, 1999. **13**(19): p. 2465-77.
35. Dorn, E.S. and J.G. Cook, *Nucleosomes in the neighborhood: new roles for chromatin modifications in replication origin control.* Epigenetics, 2011. **6**(5): p. 552-9.
36. Jiang, C. and B.F. Pugh, *Nucleosome positioning and gene regulation: advances through genomics.* Nat Rev Genet, 2009. **10**(3): p. 161-72.
37. Qi, H.Y., et al., *[Role of chromatin conformation in eukaryotic gene regulation].* Yi Chuan, 2011. **33**(12): p. 1291-9.
38. Ogbourne, S. and T.M. Antalis, *Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes.* Biochem J, 1998. **331**(Pt 1): p. 1-14.
39. Valenzuela, L. and R.T. Kamakaka, *Chromatin insulators.* Annu Rev Genet, 2006. **40**: p. 107-38.
40. Gurudatta, B.V. and V.G. Corces, *Chromatin insulators: lessons from the fly.* Brief Funct Genomic Proteomic, 2009. **8**(4): p. 276-82.
41. Yang, J. and V.G. Corces, *Insulators, long-range interactions, and genome function.* Curr Opin Genet Dev, 2012. **22**(2): p. 86-92.

42. Yusufzai, T.M., et al., *CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species.* Mol Cell, 2004. **13**(2): p. 291-8.
43. Vostrov, A.A. and W.W. Quitschke, *The zinc finger protein CTCF binds to the APBbeta domain of the amyloid beta-protein precursor promoter. Evidence for a role in transcriptional activation.* J Biol Chem, 1997. **272**(52): p. 33353-9.
44. Schubert, S., et al., *Regulation of efflux pump expression and drug resistance by the transcription factors Mrr1, Upc2, and Cap1 in Candida albicans.* Antimicrob Agents Chemother, 2011. **55**(5): p. 2212-23.
45. Matys, V., et al., *TRANSFAC: transcriptional regulation, from patterns to profiles.* Nucleic Acids Res, 2003. **31**(1): p. 374-8.
46. Portales-Casamar, E., et al., *JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles.* Nucleic Acids Res, 2010. **38**(Database issue): p. 11.
47. Kunert, N. and A. Brehm, *Novel Mi-2 related ATP-dependent chromatin remodelers.* Epigenetics, 2009. **4**(4): p. 209-11.
48. Kouzarides, T., *Chromatin modifications and their function.* Cell, 2007. **128**(4): p. 693-705.
49. Campos, E.I. and D. Reinberg, *Histones: annotating chromatin.* Annu Rev Genet, 2009. **43**: p. 559-99.
50. Adam, M., et al., *H2A.Z is required for global chromatin integrity and for recruitment of RNA polymerase II under specific conditions.* Mol Cell Biol, 2001. **21**(18): p. 6270-9.
51. Hublitz, P., M. Albert, and A.H. Peters, *Mechanisms of transcriptional repression by histone lysine methylation.* Int J Dev Biol, 2009. **53**(2-3): p. 335-54.
52. Shilatifard, A., *Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression.* Annu Rev Biochem, 2006. **75**: p. 243-69.
53. van Holde, K.E., D.E. Lohr, and C. Robert, *What happens to nucleosomes during transcription?* J Biol Chem, 1992. **267**(5): p. 2837-40.
54. Johnson, C.A. and B.M. Turner, *Histone deacetylases: complex transducers of nuclear signals.* Semin Cell Dev Biol, 1999. **10**(2): p. 179-88.
55. Lo, W.S., et al., *Phosphorylation of serine 10 in histone H3 is functionally linked in vitro and in vivo to Gcn5-mediated acetylation at lysine 14.* Mol Cell, 2000. **5**(6): p. 917-26.
56. Izzo, A. and R. Schneider, *Chatting histone modifications in mammals.* Brief Funct Genomics, 2010. **9**(5-6): p. 429-43.
57. Kornberg, R.D. and Y. Lorch, *Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome.* Cell, 1999. **98**(3): p. 285-94.
58. Schwabish, M.A. and K. Struhl, *The Swi/Snf complex is important for histone eviction during transcriptional activation and RNA polymerase II elongation in vivo.* Mol Cell Biol, 2007. **27**(20): p. 6987-95.
59. Schwabish, M.A. and K. Struhl, *Evidence for eviction and rapid deposition of histones upon transcriptional elongation by RNA polymerase II.* Mol Cell Biol, 2004. **24**(23): p. 10111-7.
60. Field, Y., et al., *Distinct modes of regulation by chromatin encoded through nucleosome positioning signals.* PLoS Comput Biol, 2008. **4**(11): p. 7.
61. Yuan, G.C., et al., *Genome-scale identification of nucleosome positions in S. cerevisiae.* Science, 2005. **309**(5734): p. 626-30.
62. Nathan, D. and D.M. Crothers, *Bending and flexibility of methylated and unmethylated EcoRI DNA.* J Mol Biol, 2002. **316**(1): p. 7-17.

63. Slotkin, R.K. and R. Martienssen, *Transposable elements and the epigenetic regulation of the genome.* Nat Rev Genet, 2007. **8**(4): p. 272-85.
64. Li, E., T.H. Bestor, and R. Jaenisch, *Targeted mutation of the DNA methyltransferase gene results in embryonic lethality.* Cell, 1992. **69**(6): p. 915-26.
65. Yang, A., et al., *Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells.* Mol Cell, 2006. **24**(4): p. 593-602.
66. Lidor Nili, E., et al., *p53 binds preferentially to genomic regions with high DNA-encoded nucleosome occupancy.* Genome Res, 2010. **20**(10): p. 1361-8.
67. Guertin, M.J. and J.T. Lis, *Chromatin landscape dictates HSF binding to target DNA elements.* PLoS Genet, 2010. **6**(9).
68. Anderson, J.D. and J. Widom, *Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites.* J Mol Biol, 2000. **296**(4): p. 979-87.
69. Polach, K.J. and J. Widom, *Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation.* J Mol Biol, 1995. **254**(2): p. 130-49.
70. Giniger, E. and M. Ptashne, *Cooperative DNA binding of the yeast transcriptional activator GAL4.* Proc Natl Acad Sci U S A, 1988. **85**(2): p. 382-6.
71. Merika, M. and S.H. Orkin, *Functional synergy and physical interactions of the erythroid transcription factor GATA-1 with the Kruppel family proteins Sp1 and EKLF.* Mol Cell Biol, 1995. **15**(5): p. 2437-47.
72. Lupien, M. and M. Brown, *Cistromics of hormone-dependent cancer.* Endocr Relat Cancer, 2009. **16**(2): p. 381-9.
73. Zhang, Z. and G.M. Fuller, *The competitive binding of STAT3 and NF-kappaB on an overlapping DNA binding site.* Biochem Biophys Res Commun, 1997. **237**(1): p. 90-4.
74. Holmes, K.A., et al., *Nkx3-1 and LEF-1 function as transcriptional inhibitors of estrogen receptor activity.* Cancer Res, 2008. **68**(18): p. 7380-5.
75. Darnell, J.E., Jr., *Transcription factors as targets for cancer therapy.* Nat Rev Cancer, 2002. **2**(10): p. 740-9.
76. Birney, E., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.
77. Xie, Z., et al., *Systematic characterization of protein-DNA interactions.* Cell Mol Life Sci, 2011. **68**(10): p. 1657-68.
78. Badis, G., et al., *Diversity and complexity in DNA recognition by transcription factors.* Science, 2009. **324**(5935): p. 1720-3.
79. Kim, T.H., L.O. Barrera, and B. Ren, *ChIP-chip for genome-wide analysis of protein binding in mammalian cells.* Curr Protoc Mol Biol, 2007. **21**(21): p. 13.
80. Kharchenko, P.V., M.Y. Tolstorukov, and P.J. Park, *Design and analysis of ChIP-seq experiments for DNA-binding proteins.* Nat Biotechnol, 2008. **26**(12): p. 1351-9.
81. MacQuarrie, K.L., et al., *Genome-wide transcription factor binding: beyond direct target regulation.* Trends Genet, 2011. **27**(4): p. 141-8.
82. Park, P.J., *ChIP-seq: advantages and challenges of a maturing technology.* Nat Rev Genet, 2009. **10**(10): p. 669-80.
83. Ren, B., et al., *Genome-wide location and function of DNA binding proteins.* Science, 2000. **290**(5500): p. 2306-9.

84. Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1135-45.
85. Ho, J.W., et al., *ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis.* BMC Genomics, 2011. **12**: p. 134.
86. Johnson, D.S., et al., *Genome-wide mapping of in vivo protein-DNA interactions.* Science, 2007. **316**(5830): p. 1497-502.
87. Rozowsky, J., et al., *PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.* Nat Biotechnol, 2009. **27**(1): p. 66-75.
88. Heintzman, N.D., et al., *Histone modifications at human enhancers reflect global cell-type-specific gene expression.* Nature, 2009. **459**(7243): p. 108-12.
89. Mikkelsen, T.S., et al., *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.* Nature, 2007. **448**(7153): p. 553-60.
90. Barski, A., et al., *High-resolution profiling of histone methylations in the human genome.* Cell, 2007. **129**(4): p. 823-37.
91. Pique-Regi, R., et al., *Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data.* Genome Res, 2011. **21**(3): p. 447-55.
92. Pan, Y., et al., *How do transcription factors select specific binding sites in the genome?* Nat Struct Mol Biol, 2009. **16**(11): p. 1118-20.
93. Lunter, G. and M. Goodson, *Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads.* Genome Res, 2011. **21**(6): p. 936-9.
94. Ji, Y., et al., *A new strategy for better genome assembly from very short reads.* BMC Bioinformatics, 2011. **12**: p. 493.
95. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
96. Pepke, S., B. Wold, and A. Mortazavi, *Computation for ChIP-seq and RNA-seq studies.* Nat Methods, 2009. **6**(11 Suppl): p. S22-32.
97. Schmid, C.D. and P. Bucher, *ChIP-Seq data reveal nucleosome architecture of human promoters*. Cell. 2007 Nov 30;131(5):831-2; author reply 832-3.
98. Baugh, L.R., J. Demodena, and P.W. Sternberg, *RNA Pol II accumulates at promoters of growth genes during developmental arrest.* Science, 2009. **324**(5923): p. 92-4.
99. Sandve, G.K. and F. Drablos, *A survey of motif discovery methods in an integrated framework.* Biol Direct, 2006. **1**: p. 11.
100. Su, J., S.A. Teichmann, and T.A. Down, *Assessing computational methods of cis-regulatory module prediction.* PLoS Comput Biol, 2010. **6**(12).
101. Das, M.K. and H.K. Dai, *A survey of DNA motif finding algorithms.* BMC Bioinformatics, 2007. **1**(8).
102. Dreszer, T.R., et al., *The UCSC Genome Browser database: extensions and updates 2011.* Nucleic Acids Res, 2012. **40**(Database issue): p. 15.
103. Kersey, P.J., et al., *Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species.* Nucleic Acids Res, 2012. **40**(Database issue): p. 8.
104. Zhu, L.J., et al., *ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data.* BMC Bioinformatics, 2010. **11**: p. 237.
105. Brosius, J., *The contribution of RNAs and retroposition to evolutionary novelties.* Genetica, 2003. **118**(2-3): p. 99-116.

106. Davidson, E.H. and R.J. Britten, *Regulation of gene expression: possible role of repetitive sequences.* Science, 1979. **204**(4397): p. 1052-9.
107. McClintock, B., *The significance of responses of the genome to challenge.* Science, 1984. **226**(4676): p. 792-801.
108. Bourque, G., et al., *Evolution of the mammalian transcription factor binding repertoire via transposable elements.* Genome Res, 2008. **18**(11): p. 1752-62.
109. Roberts, R.J., *Restriction enzymes and their isoschizomers*. Nucleic Acids Res. 1989;17 Suppl:r347-87.
110. Nair, S.K. and S.K. Burley, *X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular bases of regulation by proto-oncogenic transcription factors.* Cell, 2003. **112**(2): p. 193-205.
111. Frech, K., K. Quandt, and T. Werner, *Software for the analysis of DNA sequence elements of transcription.* Comput Appl Biosci, 1997. **13**(1): p. 89-97.
112. Crooks, G.E., et al., *WebLogo: a sequence logo generator.* Genome Res, 2004. **14**(6): p. 1188-90.
113. Wasserman, W.W. and A. Sandelin, *Applied bioinformatics for the identification of regulatory elements.* Nat Rev Genet, 2004. **5**(4): p. 276-87.
114. Fickett, J.W., *Quantitative discrimination of MEF2 sites.* Mol Cell Biol, 1996. **16**(1): p. 437-41.
115. Tronche, F., et al., *Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome.* J Mol Biol, 1997. **266**(2): p. 231-45.
116. Wolffe, A.P. and D. Guschin, *Review: chromatin structural features and targets that regulate transcription.* J Struct Biol, 2000. **129**(2-3): p. 102-22.
117. Ludwig, M.Z., et al., *Functional evolution of a cis-regulatory module.* PLoS Biol, 2005. **3**(4): p. 15.
118. Zambelli, F., G. Pesole, and G. Pavesi, *Motif discovery and transcription factor binding sites before and after the next-generation sequencing era.* Brief Bioinform, 2012. **19**: p. 19.
119. Chen, X. and M. Blanchette, *Prediction of tissue-specific cis-regulatory modules using Bayesian networks and regression trees.* BMC Bioinformatics, 2007. **8**(10).
120. Robinson, M., et al., *Improving computational predictions of cis-regulatory binding sites.* Pac Symp Biocomput, 2006: p. 391-402.
121. Hu, J., B. Li, and D. Kihara, *Limitations and potentials of current motif discovery algorithms.* Nucleic Acids Res, 2005. **33**(15): p. 4899-913.
122. Leung, M.Y., G.M. Marsh, and T.P. Speed, *Over- and underrepresentation of short DNA words in herpesvirus genomes.* J Comput Biol, 1996. **3**(3): p. 345-60.
123. Apostolico, A., M.E. Bock, and S. Lonardi, *Monotony of surprise and large-scale quest for unusual words.* J Comput Biol, 2003. **10**(3-4): p. 283-311.
124. Schneider, T.D. and R.M. Stephens, *Sequence logos: a new way to display consensus sequences.* Nucleic Acids Res, 1990. **18**(20): p. 6097-100.
125. Liu, X.S., D.L. Brutlag, and J.S. Liu, *An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments.* Nat Biotechnol, 2002. **20**(8): p. 835-9.

126. Chen, X., et al., *W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data.* Bioinformatics, 2008. **24**(9): p. 1121-8.
127. Wang, T. and G.D. Stormo, *Combining phylogenetic data with co-regulated genes to identify regulatory motifs.* Bioinformatics, 2003. **19**(18): p. 2369-80.
128. Bailey, T.L., et al., *MEME: discovering and analyzing DNA and protein sequence motifs.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W369-73.
129. Bailey, T.L. and P. Machanick, *Inferring direct DNA binding from ChIP-seq.* Nucleic Acids Res, 2012. **18**: p. 18.
130. Valouev, A., et al., *Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data.* Nat Methods, 2008. **5**(9): p. 829-34.
131. Kel-Margoulis, O.V., et al., *TRANSCompel: a database on composite regulatory elements in eukaryotic genes.* Nucleic Acids Res, 2002. **30**(1): p. 332-4.
132. Waterman, M.S., R. Arratia, and D.J. Galas, *Pattern recognition in several sequences: consensus and alignment.* Bull Math Biol, 1984. **46**(4): p. 515-27.
133. Sadler, J.R., M.S. Waterman, and T.F. Smith, *Regulatory pattern identification in nucleic acid sequences.* Nucleic Acids Res, 1983. **11**(7): p. 2221-31.
134. Galas, D.J., M. Eggert, and M.S. Waterman, *Rigorous pattern-recognition methods for DNA sequences. Analysis of promoter sequences from Escherichia coli.* J Mol Biol, 1985. **186**(1): p. 117-28.
135. Marsan, L. and M.F. Sagot, *Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification.* J Comput Biol, 2000. **7**(3-4): p. 345-62.
136. Pavesi, G. and G. Pesole, *Using Weeder for the discovery of conserved transcription factor binding sites.* Curr Protoc Bioinformatics, 2006. **2**(2): p. 11.
137. Thomas-Chollier, M., et al., *RSAT 2011: regulatory sequence analysis tools.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W86-91.
138. Sinha, S. and M. Tompa, *YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation.* Nucleic Acids Res, 2003. **31**(13): p. 3586-8.
139. Ecker, K. and L. Welch, *A Concept for ab initio Prediction of cis-regulatory Modules.* In Silico Biol, 2009. **9**(5): p. 285-306.
140. Hertz, G.Z., G.W. Hartzell, 3rd, and G.D. Stormo, *Identification of consensus patterns in unaligned DNA sequences known to be functionally related.* Comput Appl Biosci, 1990. **6**(2): p. 81-92.
141. Hertz, G.Z. and G.D. Stormo, *Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.* Bioinformatics, 1999. **15**(7-8): p. 563-77.
142. Tompa, M., *An exact method for finding short motifs in sequences, with application to the ribosome binding site problem.* Proc Int Conf Intell Syst Mol Biol, 1999: p. 262-71.
143. Bailey, T.L. and C. Elkan, *Fitting a mixture model by expectation maximization to discover motifs in biopolymers.* Proc Int Conf Intell Syst Mol Biol, 1994. **2**: p. 28-36.
144. Lawrence, C.E. and A.A. Reilly, *An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences.* Proteins, 1990. **7**(1): p. 41-51.

145. Hughes, J.D., et al., *Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae.* J Mol Biol, 2000. **296**(5): p. 1205-14.
146. Lawrence, C.E., et al., *Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.* Science, 1993. **262**(5131): p. 208-14.
147. Dermitzakis, E.T. and A.G. Clark, *Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover.* Mol Biol Evol, 2002. **19**(7): p. 1114-21.
148. Bustamante, C.D., et al., *Natural selection on protein-coding genes in the human genome.* Nature, 2005. **437**(7062): p. 1153-7.
149. De, S., N. Lopez-Bigas, and S.A. Teichmann, *Patterns of evolutionary constraints on genes in humans.* BMC Evol Biol, 2008. **8**: p. 275.
150. Lai, C.S., et al., *A forkhead-domain gene is mutated in a severe speech and language disorder.* Nature, 2001. **413**(6855): p. 519-23.
151. Enard, W., et al., *Molecular evolution of FOXP2, a gene involved in speech and language.* Nature, 2002. **418**(6900): p. 869-72.
152. Haygood, R., et al., *Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution.* Nat Genet, 2007. **39**(9): p. 1140-4.
153. Odom, D.T., et al., *Tissue-specific transcriptional regulation has diverged significantly between human and mouse.* Nat Genet, 2007. **39**(6): p. 730-2.
154. Gilad, Y., et al., *Expression profiling in primates reveals a rapid evolution of human transcription factors.* Nature, 2006. **440**(7081): p. 242-5.
155. Khaitovich, P., et al., *Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees.* Science, 2005. **309**(5742): p. 1850-4.
156. Chin, C.S., J.H. Chuang, and H. Li, *Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence.* Genome Res, 2005. **15**(2): p. 205-13.
157. Kellis, M., et al., *Sequencing and comparison of yeast species to identify genes and regulatory elements.* Nature, 2003. **423**(6937): p. 241-54.
158. Lenhard, B., et al., *Identification of conserved regulatory elements by comparative genome analysis.* J Biol, 2003. **2**(2): p. 22.
159. Moses, A.M., et al., *Position specific variation in the rate of evolution in transcription factor binding sites.* BMC Evol Biol, 2003. **3**(19): p. 28.
160. Wasserman, W.W., et al., *Human-mouse genome comparisons to locate regulatory sites.* Nat Genet, 2000. **26**(2): p. 225-8.
161. Xie, X., et al., *Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.* Nature, 2005. **434**(7031): p. 338-45.
162. Jensen, S.T., L. Shen, and J.S. Liu, *Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes.* Bioinformatics, 2005. **21**(20): p. 3832-9.
163. Qin, Z.S., et al., *Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites.* Nat Biotechnol, 2003. **21**(4): p. 435-9.
164. Chiang, D.Y., et al., *Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts.* Genome Biol, 2003. **4**(7): p. 26.

165. Hardison, R.C., *Conserved noncoding sequences are reliable guides to regulatory elements.* Trends Genet, 2000. **16**(9): p. 369-72.
166. Woolfe, A., et al., *Highly conserved non-coding sequences are associated with vertebrate development.* PLoS Biol, 2005. **3**(1): p. 11.
167. Prud'homme, B., N. Gompel, and S.B. Carroll, *Emerging principles of regulatory evolution.* Proc Natl Acad Sci U S A, 2007. **1**: p. 8605-12.
168. Kato, M., et al., *Identifying combinatorial regulation of transcription factors and binding motifs.* Genome Biol, 2004. **5**(8): p. 28.
169. Berman, B.P., et al., *Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura.* Genome Biol, 2004. **5**(9): p. 20.
170. Levine, M. and E.H. Davidson, *Gene regulatory networks for development.* Proc Natl Acad Sci U S A, 2005. **102**(14): p. 4936-42.
171. Smith, A.D., et al., *Mining ChIP-chip data for transcription factor and cofactor binding sites.* Bioinformatics, 2005. **21**(1): p. i403-12.
172. GuhaThakurta, D. and G.D. Stormo, *Identifying target sites for cooperatively binding factors.* Bioinformatics, 2001. **17**(7): p. 608-21.
173. Makeev, V.J., et al., *Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information.* Nucleic Acids Res, 2003. **31**(20): p. 6016-26.
174. Erives, A. and M. Levine, *Coordinate enhancers share common organizational features in the Drosophila genome.* Proc Natl Acad Sci U S A, 2004. **101**(11): p. 3851-6.
175. Johansson, O., et al., *Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm.* Bioinformatics, 2003. **19**(1): p. i169-76.
176. Bailey, T.L. and W.S. Noble, *Searching for statistically significant regulatory modules.* Bioinformatics, 2003. **19**(2): p. ii16-25.
177. Pierstorff, N., C.M. Bergman, and T. Wiehe, *Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA.* Bioinformatics, 2006. **22**(23): p. 2858-64.
178. Frith, M.C., M.C. Li, and Z. Weng, *Cluster-Buster: Finding dense clusters of motifs in DNA sequences.* Nucleic Acids Res, 2003. **31**(13): p. 3666-8.
179. Sinha, S., E. van Nimwegen, and E.D. Siggia, *A probabilistic method to detect regulatory modules.* Bioinformatics, 2003. **19**(1): p. i292-301.
180. Zhou, Q. and W.H. Wong, *CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling.* Proc Natl Acad Sci U S A, 2004. **101**(33): p. 12114-9.
181. Matys, V., et al., *TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.* Nucleic Acids Res, 2006. **34**(Database issue): p. D108-10.
182. Sandelin, A. and W.W. Wasserman, *Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics.* J Mol Biol, 2004. **338**(2): p. 207-15.
183. Seldeen, K.L., et al., *Evidence that the bZIP domains of the Jun transcription factor bind to DNA as monomers prior to folding and homodimerization.* Arch Biochem Biophys, 2008. **480**(2): p. 75-84.

184. Steingrimsson, E., et al., *The semidominant Mi(b) mutation identifies a role for the HLH domain in DNA binding in addition to its role in protein dimerization.* Embo J, 1996. **15**(22): p. 6280-9.
185. Schwabe, J.W., et al., *The crystal structure of the estrogen receptor DNA-binding domain bound to DNA: how receptors discriminate between their response elements.* Cell, 1993. **75**(3): p. 567-78.
186. Favorov, A.V., et al., *A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length.* Bioinformatics, 2005. **21**(10): p. 2240-5.
187. Bi, C. and P.K. Rogan, *Bipartite pattern discovery by entropy minimization-based multiple local alignment.* Nucleic Acids Res, 2004. **32**(17): p. 4979-91.
188. Macisaac, K.D., et al., *A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data.* Bioinformatics, 2006. **22**(4): p. 423-9.
189. Tabach, Y., et al., *Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site.* PLoS One, 2007. **2**(8).
190. Elemento, O., N. Slonim, and S. Tavazoie, *A universal framework for regulatory element discovery across all genomes and data types.* Mol Cell, 2007. **28**(2): p. 337-50.
191. Espinosa, J.M., *Mechanisms of regulatory diversity within the p53 transcriptional network.* Oncogene, 2008. **27**(29): p. 4013-23.
192. Liu, X., et al., *Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection.* Genome Res, 2006. **16**(12): p. 1517-28.
193. Kaplan, N., et al., *The DNA-encoded nucleosome organization of a eukaryotic genome.* Nature, 2009. **458**(7236): p. 362-6.
194. Harbison, C.T., et al., *Transcriptional regulatory code of a eukaryotic genome.* Nature, 2004. **431**(7004): p. 99-104.
195. Won, K.J., B. Ren, and W. Wang, *Genome-wide prediction of transcription factor binding sites using an integrated model.* Genome Biol, 2010. **11**(1).
196. Ernst, J., et al., *Integrating multiple evidence sources to predict transcription factor binding in the human genome.* Genome Res, 2010. **20**(4): p. 526-36.
197. Cuellar-Partida, G., et al., *Epigenetic priors for identifying active transcription factor binding sites.* Bioinformatics, 2012. **28**(1): p. 56-62.
198. van Heeringen, S.J. and G.J. Veenstra, *GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments.* Bioinformatics, 2011. **27**(2): p. 270-1.
199. Romer, K.A., G.R. Kayombya, and E. Fraenkel, *WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and Bayesian approaches.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. 21.
200. Kuttippurathu, L., et al., *CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments.* Bioinformatics, 2011. **27**(5): p. 715-7.
201. Mahony, S. and P.V. Benos, *STAMP: a web tool for exploring DNA-binding motif similarities.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. 3.
202. Gupta, S., et al., *Quantifying similarity between motifs.* Genome Biol, 2007. **8**(2).

203. Machanick, P. and T.L. Bailey, *MEME-ChIP: motif analysis of large DNA datasets.* Bioinformatics, 2011. **27**(12): p. 1696-7.
204. Wilbanks, E.G. and M.T. Facciotti, *Evaluation of algorithm performance in ChIP-seq peak detection.* PLoS One, 2010. **5**(7).
205. Jothi, R., et al., *Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data.* Nucleic Acids Res, 2008. **36**(16): p. 5221-31.
206. Hu, M., et al., *On the detection and refinement of transcription factor binding sites using ChIP-Seq data.* Nucleic Acids Res, 2010. **38**(7): p. 2154-67.
207. Rhee, H.S. and B.F. Pugh, *Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution.* Cell, 2011. **147**(6): p. 1408-19.
208. Liu, T.T., et al., *Genome-wide expression and location analyses of the Candida albicans Tac1p regulon.* Eukaryot Cell, 2007. **6**(11): p. 2122-38.
209. Carlson, J.M., et al., *SCOPE: a web server for practical de novo motif discovery.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. 7.
210. Liao, D., *Emerging roles of the EBF family of transcription factors in tumor suppression.* Mol Cancer Res, 2009. **7**(12): p. 1893-901.
211. Eychene, A., N. Rocques, and C. Pouponnot, *A new MAFia in cancer.* Nat Rev Cancer, 2008. **8**(9): p. 683-93.
212. Hua, S., et al., *Genomic analysis of estrogen cascade reveals histone variant H2A.Z associated with breast cancer progression.* Mol Syst Biol, 2008. **4**(188): p. 15.
213. Culhane, A.C. and J. Quackenbush, *Confounding effects in "A six-gene signature predicting breast cancer lung metastasis"*. Cancer Res. 2009 Sep 15;69(18):7480-5. Epub 2009 Sep 1.
214. Bussemaker, H.J., H. Li, and E.D. Siggia, *Regulatory element detection using correlation with expression.* Nat Genet, 2001. **27**(2): p. 167-71.
215. Lupien, M., et al., *Coactivator function defines the active estrogen receptor alpha cistrome.* Mol Cell Biol, 2009. **29**(12): p. 3413-23.

ANNEXE 1

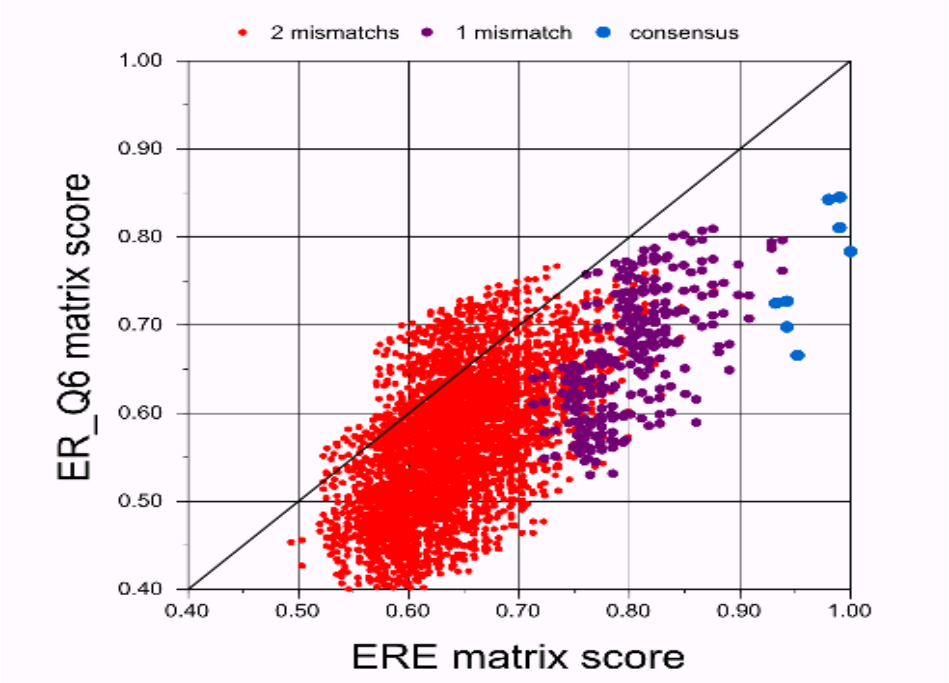


Figure 1. Comparaison des scores des matrices ERE

La matrice ERE prédite dans [220] et la matrice ERE définie dans TRANSFAC (ER_Q6). Chaque point correspond aux scores assignés par les deux matrices à un ERE comptant de 0 à 2 nucléotides différent d'un ERE de haute affinité (RGGTCAnnnTGACCY).