Etude des éléments cis-régulateurs : identification et caractérisation

Jeanne Chèneby

UNIVERSITE D'AIX-MARSEILLE

ECOLE DOCTORALE SCIENCES DE LA VIE ET DE LA SANTE

INSERM U1090 TAGC

Theories and Approaches of Genomic Complexity

Thèse présentée pour obtenir le grade universitaire de docteur

Discipline : Biologie

Spécialité : Génomique et Bioinformatique

Jeanne CHÈNEBY

Etude des éléments cis-régulateurs : identification et caractérisation

Soutenue le 19/12/2019 devant le jury :

Philipp BUCHER	Rapporteur
Morgane THOMAS-CHOLLIER	Rapporteur
François PARCY	Examinateur
Delphine POTIER	Invité
Andrew SAURIN	Invité
Christine BRUN	Directrice de thèse
Benoît BALLESTER	Co-directeur de thèse

Résumé

Le processus de régulation de la transcription des gènes repose très largement sur l'existence de séquences d'ADN non codantes dans le génome. Ces séquences d'ADN, appelées "éléments cis-régulateurs", ont la particularité de recruter de nombreuses protéines capables de réguler le niveau de transcription des gènes. Parmi ces protéines, les facteurs de transcription sont capables de se fixer directement sur l'ADN. Les facteurs de transcription coopèrent avec d'autres protéines régulatrices, les cofacteurs, afin de réguler la transcription. Les protéines régulatrices de la transcription permettent la fixation et la régulation de l'enzyme d'ARN polymérase II qui transcrit les gènes en ARN messager. Leurs fixations sur les éléments cis-régulateurs permettent une régulation des gènes dans l'espace et dans le temps. Pour mieux comprendre la régulation de l'expression des gènes, il est nécessaire d'identifier les éléments cis-régulateurs dans le génome afin de caractériser et d'identifier les mécanismes d'action des éléments régulateurs et des protéines qui leur sont liés. Le développement rapide des méthodes de séquençage à haut débit a permis l'identification des interactions ADN/protéines à grande échelle. L'accumulation massive des données de séquençage dans les banques de données publiques permet l'intégration de nombreuses expériences capturant les interactions entre les facteurs de transcription et l'ADN par des moyens bioinformatiques. Le but de mon doctorat a été d'annoter et traiter de façon uniforme les données brutes issues d'expériences de séquençage ayant pour objectif d'identifier les régions de fixation des protéines régulatrices pour l'Homme puis chez *Arabidopsis Thaliana*. Nous avons traité des données de ChIP-seq, ChIP-exo et DAP-seq afin d'élaborer plusieurs catalogues de régions régulatrices chez l'homme et chez Arabidopsis *Thaliana.* Toutes ces données sont disponibles au sein du projet ReMap. Pour *Arabidopsis Thaliana, n*ous avons complété ces données par une analyse de toutes les marques d'histones. Pour effectuer ces analyses, nous avons développé des workflows reproductibles, scalables et portables sur des architectures différentes. Cette analyse intégrative à haut débit nous a permis d'identifier de nombreux nouveaux éléments cis-régulateurs. Ces données ont également été utilisées pour identifier les sites de fixations reconnus par les facteurs de transcription et pour consolider la base de données JASPAR pour l'Homme et pour *Arabidopsis Thaliana*. Enfin, ce catalogue a été utilisé dans le développement d'une nouvelle méthode appliquant un algorithme basé sur l'entropie. Cet algorithme permet de différencier les événements de fixations directes et indirectes par les protéines dans les résultats de ChIP-seq.

Abstract

The regulation of gene transcription is largely based on the existence of non-coding DNA sequences in the genome. These DNA sequences, called "cis-regulatory elements", have the particularity of recruiting many proteins capable of regulating the level of gene transcription. Among these proteins, transcription factors are capable of directly binding to DNA. Transcription factors cooperate with other regulatory proteins, called cofactors, to regulate transcription. Transcription regulatory proteins allow the binding and regulation of the RNA polymerase II enzyme that transcribes genes into messenger RNA. The fixation of transcription factors on the cis-regulatory elements allows the regulation of genes in space and time. To better understand the regulation of gene expression, it is necessary to identify cis-regulatory elements in the genome in order to characterize and identify the mechanisms of action of regulatory elements and the proteins that are linked to them. The rapid development of high throughput sequencing methods has made it possible to identify DNA/protein interactions on a large scale. The massive accumulation of sequencing data in public databases allows the integration of many experiments that capture the interactions between transcription factors and DNA through bioinformatics. The purpose of my PhD was to annotate and process in a uniform way the raw data from sequencing experiments whose objective is to identify the binding regions of regulatory proteins for humans and then for Arabidopsis Thaliana. We processed data from ChIP-seq, ChIP-exo and DAP-seq to develop several catalogues of regulatory regions in humans and Arabidopsis Thaliana. All this data is available within the ReMap project. We completed them with an analysis of all histone brands for Arabidopsis Thaliana. To carry out these analyses we have developed reproducible, scalable and portable workflows working on different architectures. This high throughput integrative analysis has allowed us to identify many new cis-regulatory elements. These data were also used to identify the attachment sites recognized by the transcription factors and to consolidate the JASPAR database for humans and Arabidopsis Thaliana. Finally, this catalogue was used in the development of a new method applying an entropy-based algorithm to differentiate between direct and indirect protein binding events in ChIP-seq results.

Remerciements

Tout d'abord, je tiens à remercier les membres Philipp Bucher et Morgane Thomas-Chollier d'avoir accèpté de lire ce manuscrit. Je remercie aussi François Parcy et Delphine Potier de participer à ce jury. Je remercie particulièrement Andrew Saurin de m'avoir suivie pendant ces trois années en participant à mes commités de suivi de thèse, ainsi qu'à ce jury.

Je remercie également Catherine Nguyen et Pascal Rihiet de m'avoir accueilli au sein du laboratoire, et ce, pour mon stage de Master ainsi que pour ma thèse. Merci à Christine Brun d'avoir accèpté d'être ma directice de thèse. Ses nombreuses signatures m'ont permis de mener à bien (ou mal) cette thèse.

Je tiens également à exprimer ma gratitude à Benoît Ballester. De m'avoir fait confiance pour un projet d'une telle envergure et permis d'explorer le monde merveilleux du Big data et de la régulation de la transcription. De m'avoir laissé la liberté de mener des collaborations scientifiques. Merci, en particulier, de ne pas m'avoir fait coder en Perl, ni fait utiliser de Mac. Je te remercie également pour tes conseils sur ce que représente le métier de chercheur, ses valeurs et le soutien que tu m'as fourni pendant ces trois années de thèse et au cours de mon stage.

Un grand merci à Lionel Spinelli pour ses nombreux conseils technique et scientifique concernant la bioinformatique. Merci également au service informatique Fafa et Aurélie pour leur soutien technique (désolée de ne pas être passée par Astec pour ce message).

Merci à l'équipe pédagogique de m'avoir fait découvrir l'envers du décor et que, oui, les enseignants sont bien des êtres humains normaux, fait de chair et de café. Je remercie particulièrement Chirstophe Bordi pour ton investissement à tous les niveaux de l'enseignement, pour tes discussions toujours interessantes et pour le co-voiturage. La qualité de ton enseignement, ainsi que celle de Pascal Hingamp, ont fait des TPs à Aix des souvenirs inoubliables.

Je remercie tout particulièrement les étudiants du labo pour les repas toujours dans la bonne humeur. A mes amis au labo et à la ville, Lucie, Marie, Sabrina, pour leur soutien moral et les soirées filles, à Flo, Laurent et Quentin pour le soutien, les discussions « geek » et partie de cartes illégales (Flo on sait tous que tu vampirises notre chance). Quentin, merci de m'avoir initiée au Deep learning et pour les discussions politiques et historiques improbables (je t'aurais nommé Chancelier si Glitterhoof n'occupait pas déjà le poste). Merci également à Adrien et Justine Long de m'avoir permis d'appliquer les enseignements du grand gourou Alex, à savoir : « Il n'y a pas de vie à côté de la thèse, mais une thèse à côté de la vie ». Je remercie également l'équipe d'Hippo'thèse de m'avoir fait découvrir la vie associative étudiante et fait sortir de ma zone de confort.

Finalement, je remercie ma mère de m'avoir soutenu tout au long de ma scolarité chaotique et pour les nombreuses corrections orthographiques et synthaxiques de ce manuscrit. Je ne serai pas arrivé là où je suis sans ton aide.

Table des matières

Abréviations

A. Thaliana : *Arabidopsis Thaliana*

BRE : TFIIB Recognition Element

CRE : éléments cis-régulateurs

CRF : Facteur remodelant de la chromatine

FRiP : Fraction of Reads in Peaks

GTF : Facteur général de transcription

HAT : Histone AcetyleTransferase

HSAC : Histone DeACetylase

IgG : Immunoglobulines de type G

Inr : Séquence initiatrice, *initiator*

ML : Machine learning

MPP : Matrice poids-position

NSC : Normalized Strand Coefficient

nt : nucléotide

pb : paire de bases

PCR: *Polymerase Chain Reaction*

PIC : Complexe Protéique de pré-initation

Pol II : ARN polymérase II

RSC : Relative Strand Correlation

TAD : Topologically Associated Domain

TF : Facteur de transcription

TFBS : Site de fixation des facteurs de transcription

TSS : Site d'initiation de la transcription

TTS : Site de terminaison de la transcription

WCE : *Whole cell extract*, control

Introduction

I Régulation chez l'homme

Bien que l'homme soit un organisme pluricellulaire complexe, la plupart des cellules du corps humain sont composées du même génome. Cela implique des mécanismes fins qui permettent le contrôle de l'expression des gènes allant de la transcription à la traduction. Cette régulation est à l'origine, non seulement de la grande diversité des types cellulaires chez l'Homme, mais aussi de l'adaptativité des cellules à leur environnement.

Chez les eucaryotes, la transcription de l'ADN en ARN est assurée par l'ARN polymérase répartie en trois classes de protéines : I, II et III pouvant transcrire différents types d'ARN. L'ARN polymérase I transcrit les grandes sous-unités du Ribosome[1]. L'ARN polymérase III transcrit principalement les ARN de transfert et la petite sous-unité du ribosome[2]. Pour les eucaryotes, la transcription des gènes est assurée par l'ARN polymérase II (Pol II). Le positionnement et l'activation de cette enzyme sont régulés par les protéines régulatrices de la transcription qui sont les facteurs généraux de la transcription (GTF), les activateurs et les répresseurs (facteur de transcription) et les cofacteurs.

A. Protéines régulatrices de la transcription

Les protéines régulatrices de la transcription jouent un rôle important dans l'initiation et le contrôle de la transcription des gènes. Une protéine régulatrice peut activer ou réprimer le niveau de transcription des gènes environnants. Dans certains cas, la même protéine peut se comporter en activateur ou répresseur en fonction des conditions environnantes[3]. Les protéines régulatrices sont les éléments trans-régulateurs de la transcription.

Les facteurs de transcription (TF) se fixent sur des séquences particulières de la chromatine appelées sites de fixation des TFs (*Trasncription Factor Binding Site*, TFBS). Ces TFBS sont des séquences dégénérées de 6 à 12 paires de bases (pb). Malgré la faible spécificité de ces séquences, les TFs reconnaissent les motifs des TFBS grâce à leur domaine de liaison à l'ADN. Pour un même TF, la séquence (motif) de fixation du TFBS peut varier légèrement. C'est pour cette raison que les motifs sont décrits par des séquences consensus. Ces variations sont plus ou moins conservées dans le génome et au cours de l'évolution[4]. Une mutation sur un nucléotide conservée du motif entraîne un impact plus prononcé sur la fixation du TF. Pour plus de précision, les motifs des TFBS sont aussi représentés sous forme de matrices de poids-position (MPP) et de logo (Figure I.1).

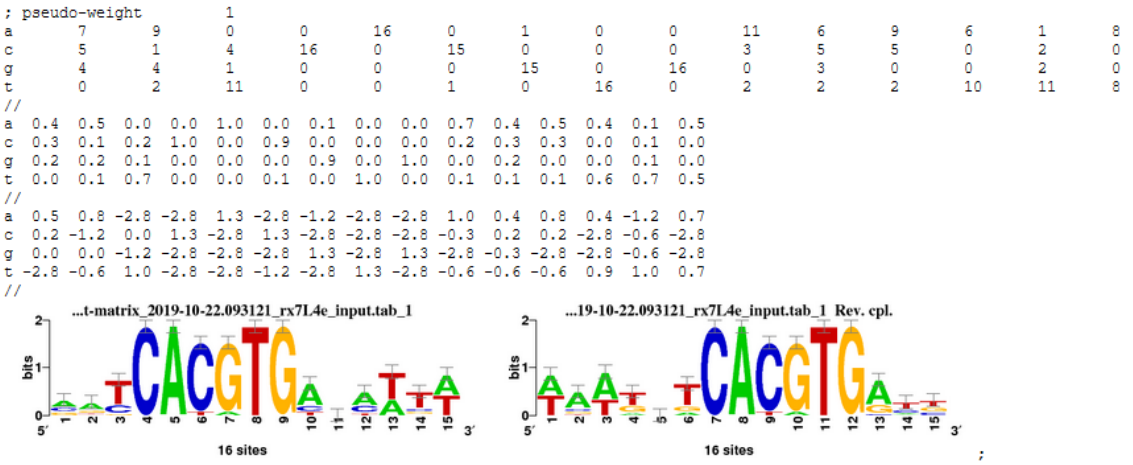


Figure I.1 ***: Exemple de construction de MPP et logo (démo RSAT convert-matrix[5]).*** *Les profils de fixations des TFs sont construits à partir de séquences obtenues par des expériences biologiques capturant les interactions protéine/ADN. Dans ce cas, des données ChIP-seq pour le TF MET4[6]. Le premier tableau est la matrice de comptage des nucléotides (nt) par position. Il est construit à partir de 16 séquences de 15 nt. Ce tableau de comptage est converti en tableau de fréquence, puis en MPP. La MPP est intéressante car elle prend en compte la fréquence théorique de chaque nucléotide dans le génome. Les logos permettent de représenter visuellement l'éventualité de chaque nucléotide pour toutes les positions du motif.*

Ces variations et la faible taille des régions reconnues, couplées à la plasticité du génome, implique qu'il y a de grandes chances qu'un TFBS apparaisse ou disparaisse au cours d'événements de mutation de l'ADN[7]. La régulation de la transcription peut considérablement varier entre les individus[8].

Les TFs jouent un rôle durant tout le processus de transcription des gènes. Pour qu'un gène soit transcrit il est nécessaire que : la Pol II se fixe sur la chromatine en amont des gènes et qu'elle soit activée par des protéines afin de transcrire l'ADN en ARN. La fixation de la Pol II ne peut se faire qu'à deux conditions : la première, que la chromatine soit ouverte et accessible au complexe de transcription, la deuxième, que la Pol II soit recrutée par d'autres protéines au niveau des gènes à transcrire. L'état de la chromatine est donc un élément important dans la régulation de la transcription.

1. Conformation de la chromatine

La chromatine est composée d'un assemblage d'ADN et de protéines. C'est une structure permettant le compactage du génome humain, ce dernier étant composé de 3 milliards de nucléotides et mesurant entre 2 et 3 mètres bout à bout[9,10] dans un noyau de cellule ne mesurant que quelques micromètres. Pour assurer ce niveau de compactage, l'ADN est enroulé autour du nucléosome, complexe formé de protéines histones. Le niveau de compactage de la chromatine est un facteur important dans la transcription car il permet de rendre accessibles ou non les gènes aux protéine régulatrices.

a. Le nucléosome

Le nucléosome est l'unité de base de compaction de l'ADN chez les eucaryotes. C'est un complexe formé d'ADN et d'un "cœur" de protéines. Il mesure 11 nanomètres de diamètre sur 5,5 nanomètres de haut. Environ 146 paires de bases d'ADN peuvent s'enrouler autour du "cœur" du nucléosome. L'ADN enroulé autour de ces nucléosomes sont six fois plus compactes qu'un fragment d'ADN nu comportant le même nombre de nucléotides[11]. Les nucléosomes sont ensuite regroupés par six pour former un solénoïde. Les nucléosomes sont liés entre eux par une protéine (histone H1). Les solénoïdes sont à leur tour enroulés sur d'une protéine d'échafaudage (*Scaffold proteins)*, lui-même enroulé pour former la matrice chromosomique (Figure I.2).

C'est la distance entre les nucléosomes qui décide de l'accessibilité de l'ADN. (Il est intéressant de noter que l'ADN est non seulement accessible à la Pol II, mais aussi à d'autres protéines pouvant se fixer à l'ADN). La distance entre nucléosomes n'est pas figée. Elle peut considérablement varier au cours de la vie de la cellule. Cette variation est, en grande partie, due aux histones (protéines composant le nucléosome).

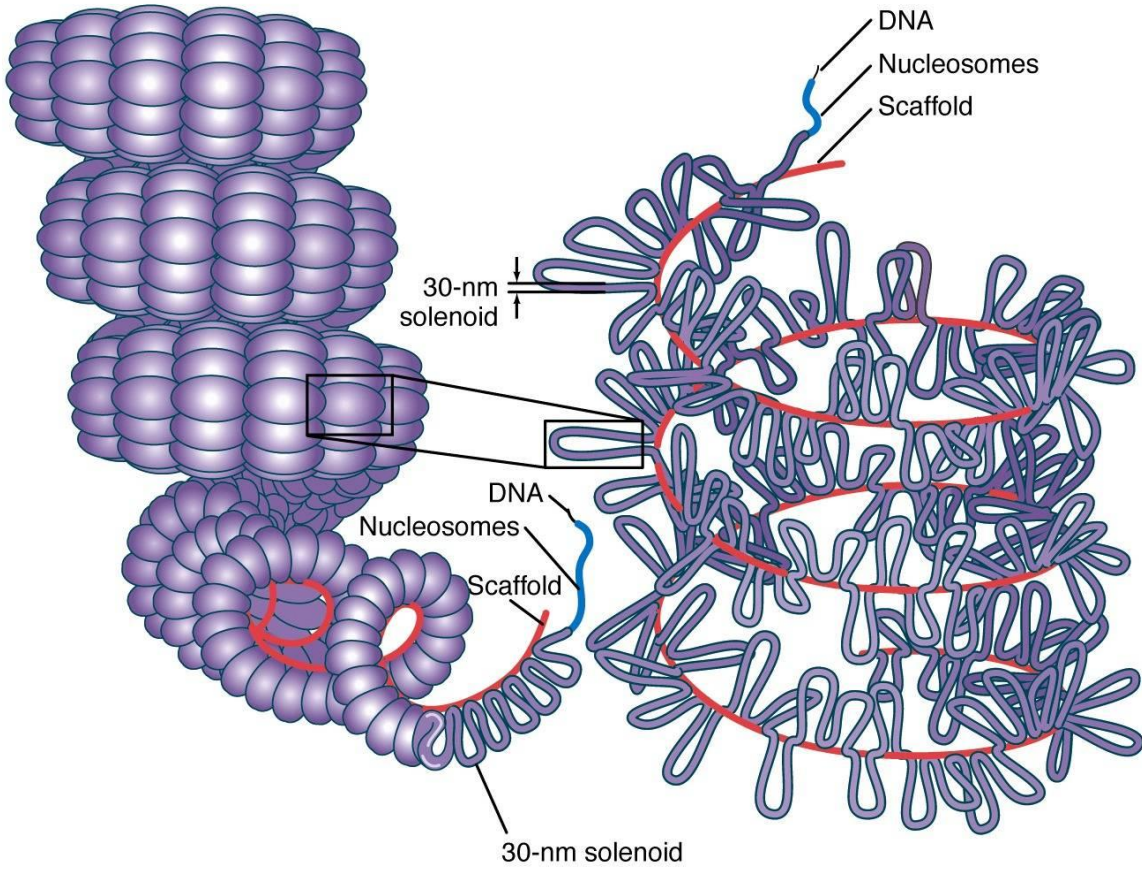


Figure I.2 ***: Structure de la chromatine.***
L'ADN est enroulé autour de nucléosomes qui forment des brins compacts de chromatine appelés solénoïdes. Les solénoïdes sont maintenus entre eux tous les 30 nanomètres par une structure protéique appelée Scaffold.

Figure I.2 : Griffiths, A. J. F. *An introduction to genetic analysis.* (W.H. Freeman, 2000).

b. Les histones

Les histones sont des protéines qui forment, avec l'ADN, le nucléosome. Les principales histones appartiennent aux classes : H2A, H2B, H3, H4 et constituent le "cœur" du nucléosome. Le nucléosome est composé de deux histones de ces classes formant un octamère (Figure I.3). Une autre classe importante d'histone est la H1. Elle "referme" le nucléosome et le scelle sur place. Les histones sont des protéines alcalines (charge positive des acides aminés qui la composent), ce qui permet l'interaction avec les molécules d'ADN (le phosphate portant des charges négatives).

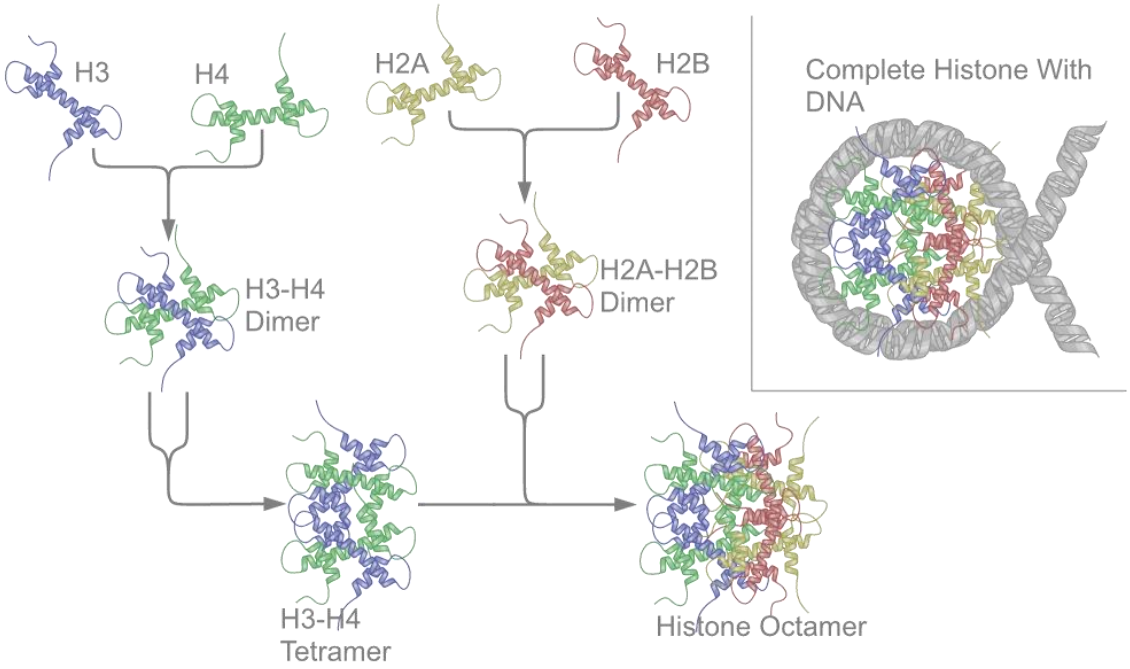


Figure I.3 ***: Assemblage des histones.***
Les histones sont formées de 2 hétérotétramères composés chacun de deux dimères de protéines (H3-H4 et H2A-H2B). L'enroulement des molécules d'ADN autour d'un octamère d'histone forme le nucléosome.

c. Modification de la conformation de la chromatine

La modification de la conformation de la chromatine joue un rôle majeur dans la régulation de la transcription. Une chromatine ouverte permet la fixation de la machinerie transcriptionnelle et des TFs, et donc la transcription des gènes[12]. De plus, les charges électrostatiques portées par les histones influencent les interactions entre les protéines et l'ADN[13]. Les mécanismes de modification de la transcription sont les suivants.

Figure I.3 : Richard Wheeler (Zephyris) -wikimedia Commons

Les histones peuvent agir dynamiquement sur le degré de compaction de l'ADN grâce à des modifications post-traductionnelles. Ces modifications se font sur les lysines (K) des extrémités amino-terminales (N-terminale) des histones. Les extrémités N-terminales sont accessibles aux protéines modificatrices car elle se projettent à l'extérieur du nucléosome (Figure I.4). Ces modifications agissent sur l'état transcriptionnel de la chromatine et ont été caractérisées par le Code des histones[14]. Le Code des histones décrit deux grands types de modification, l'acétylation (ajout d'un groupement acétyle) et la méthylation (ajout de groupements méthyl). Ces modifications ont généralement un rôle antagoniste. L'acétylation est une marque d'ADN accessible (euchromatine), alors que la méthylation marque l'ADN compacté (hétérochromatine).

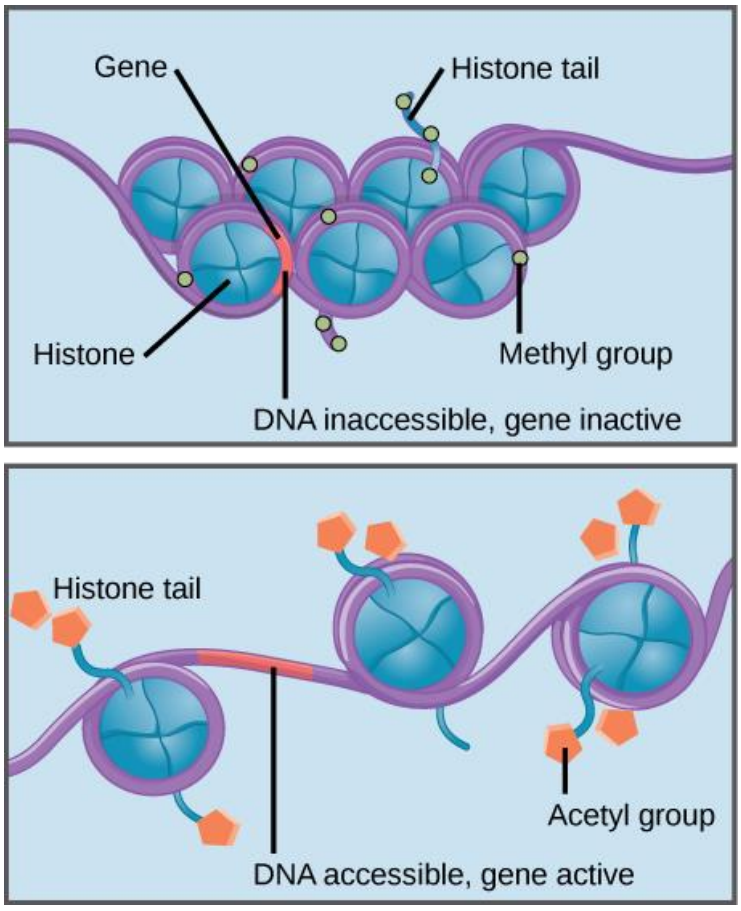


Figure I.4 ***: Accessibilité de l'ADN.***
La modification des queues d'histones influe sur le niveau de compaction de la chromatine. La présence de groupes méthyles sur les extrémités N-terminal des histones est généralement un signe d'ADN inaccessible. Par conséquent, les gènes enroulés ne peuvent pas être transcrits. Les groupes acétyles sont, quant à eux, un signe de chromatine ouverte. Cela implique que l'ADN situé entre les nucléosomes est accessible, permettant ainsi la fixation des TFs et de la machinerie transcriptionnelle.

Figure I.4 : http://cnx.org/contents/GFy_h8cu@10.53:rZudN6XP@2/Introduction

La nomenclature du Code d'histone permet d'identifier l'histone (H), le fragment de lysine (K), le type de modification (ac : acétylation, me : méthylation) et le nombre de modifications (1, 2 3, etc.). Chez l'humain, la modification se fait généralement sur l'histone 3 (H3)[15]. De plus, certaines modifications sont associées à des régions régulatrices particulières[13]. Ces modifications sont opérées par des protéines particulières appelées Facteurs de Remodelage de la Chromatine (*Chromatin Remodeling Factor*, CRF). Elles font partie de la classe des protéines régulatrices de la transcription. Parmi les CRFs, les corégulateurs de la transcription ont la particularité d'être recrutés par les TFs. Ces protéines peuvent, soit activer la transcription en acétylant la chromatine (*Histone AcetyleTransferase*, HAT), soit la réprimer en désacétylant la chromatine (*Histone DeACetylase*, HSAC). Dans le cas où la chromatine est fermée, les corépresseurs sont recrutés par des TFs particuliers, les facteurs pionniers. Ces TFs ont la particularité de pouvoir se fixer sur la chromatine enroulée autour du nucléosome. Chez les mammifères, le facteur OCT4[16], les familles FoxA[17] et PBX1[18] sont des candidats pour le rôle de facteurs pionniers. Le Tableau I.1 présente des exemples communs de modifications d'histones et leurs conséquences sur la transcription.

Tableau I.1 ***: Exemples de modifications d'histones et de leurs conséquences sur la transcription.***

Histone	Lysine	Modification	Facteur de remodelage	Effet sur la transcription
H3	K4	Triméthylation (me3)	SET7	Activation de la transcription[19]
H3	K4	Méthylation (me1)	ALL-1	Activation de la transcription[20]
H3	K9	Triméthylation (me3)	SUV39H1	Répression de la transcription[21]
H3	K14	Acétylation (ac)	TAF1, EP300	Transcription pour la Pol II/activation de la transcription[22,23]
H3	K27	Acétylation (ac)	EP300	Activation de la transcription[24]
H1	K26	Méthylation	Ezh2	Répression de la transcription[25]

2. Initiation de la transcription

Au niveau de la chromatine ouverte, la Pol II est recrutée au niveau des gènes codants par le Complexe Protéique de Pre-Initation (*Pre-Initation Complex*, PIC). Le PIC est composé de Facteurs Généraux de Transcription (GTF) qui sont des protéines capables de se fixer sur des régions spécifiques de l'ADN en amont des gènes. Ces régions peuvent être la TATA-*box*[26] (boite TATA). Ces TATA-*box* font partie des séquences régulatrices du promoteur situées en amont de certains gènes (environ ~24% des promoteurs humain possèdent une séquence TATA-*like*[27]). En l'absence de TATA-*box* le PIC se fixe, malgré tout, sur le promoteur au niveau de TFIIB Recognition Element[28] (*B Recognition Element*, BRE). Le promoteur contient le Site d'Initiation de la Transcription (TSS) qui est considéré comme la position +1 paire de bases (pb) du gène.

L'initiation de la transcription se fait par le recrutement successif des GTFs formant le PIC. Dans un premier temps, le promoteur est reconnu au niveau de la TATA-*box* ou du BRE par une protéine du PIC (*TATA-box Binding Protein*, TBP), le Facteur de Transcription II D (TFII D). Si le promoteur contient une TATA-*box* et des BREs, le TFII A stabilise le complexe TFII D/ADN. Le TFII B se fixe à la séquence BRE de l'ADN de part et d'autre de la TATA-box. La TFII F recrute ensuite la Pol II sur le site. Les protéines TFII E et la TFII H (*helicase ATP-dependante*) se fixent une fois le complexe formé (Figure I.5).

Le niveau de méthylation de l'ADN peut aussi influer sur la fixation du PIC. Chez l'homme, cette méthylation intervient au niveau des cytosines. Elle se produit à la position 5 du cycle de pyrimidine formant le 5-méthylcytosine. Chez les vertébrés, cette méthylation intervient dans les séquences CG (CpG) de l'ADN. Chez les mammifères, environ 70% des cytosines, dans des configurations CpG, sont méthylés[29]. Les CpG non méthylés sont regroupés dans des îlots de régions riches en CpG. Une région est considérée comme un îlot CpG si elle fait plus de 200pb, a une concentration de C/G supérieure à 50% et si le ratio nombre de CpG observés sur nombre de CpG théoriquement attendus est supérieur à 0,6[30]. Les îlots CpG se concentrent autour de régions fixant les TFs[31]. Environ 70% des promoteurs proches du TSS se trouvent dans des îlots CpG. En général, les promoteurs possédant des îlots CpG ne possèdent pas de TATA-box et inversement[32]. La méthylation de l'ADN influence la fixation de TFs et du complexe PIC [33,34]. Les régions situées en amont du TSS hypométhylé, comme les îlots CpG, favorisent la transcription des gènes[35]. La méthylation des îlots CpG peut entraîner la répression des gènes proches[36]. Chez l'homme, la méthylation de l'ADN est gérée par les DNA méthyltransférases (DNMT).

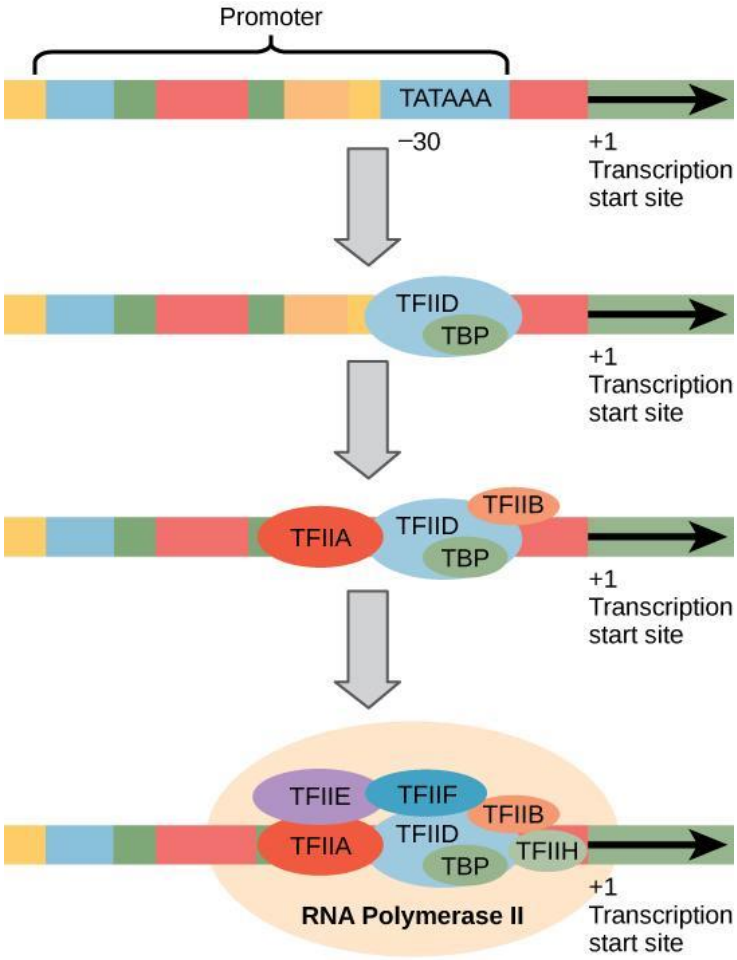


Figure I.5 ***: Fixations séquentielles du PIC en amont des gènes.***
Les promoteurs sont des régions situées en amont des gènes et capables de fixer des protéines régulatrices de la transcription. La position représente la distance, en nucléotides, entre le promoteur et le TSS. En amont elle est négative. La TATA-box se situe à la position -30 et peut être proche de BREs. Ceux-ci recrutent le TFIID, premier élément du PIC. Le positionnement de TFIID permet le recrutement de TFIIA et TFIIB. La fixation des autres protéines forme ensuite le PIC qui permet le recrutement de la Pol II.

3. L'ARN polymérase II

L'ARN polymérase de classe II (Pol II) est une protéine complexe de 550 Kilo Dalton. Chez l'homme, elle est formée de 12 sous-unités et a pour particularités de pouvoir se fixer à la chromatine et de transcrire les molécules d'ADN en ARN. La Pol II code pour l'ARN messager, les petits ARN nucléaires, les petits ARN interférents et des micro-ARN (ARNm, ARNsn, pARNi, miARN).

Les 12 sous-unités sont nommées de RPB1 à RPB12 pour RNA Polymérase B. Les plus importantes sont décrites dans le Tableau I.2 :

Figure I.5 : http://cnx.org/content/m44524/latest/

Tableau I.2 ***: Description des sous-unités du PIC.***

Sous-unité	Description
RPB1	Plus grande sous-unité Contient les domaines C-terminal (CTD) Se lie à la chromatine par une crevasse qui transcrit l'ADN en ARN
RPB2	Deuxième plus grande sous-unité Appartient à la structure qui maintient le contact entre le site actif de l'enzyme et la chromatine et l'ARN transcrit
RPB3	Troisième plus grande sous-unité Hétérodimère avec une autre sous-unité Constitue un sous assemblage avec la chaîne POLR2J Interagit avec RPB1 à 5, 7 et 10 à 12
RPB4	Est essentiel lors de stress thermique et de carence en nutriment
RPB5	Présente en deux exemplaires dans la Pol II Interagit avec RPB1, 3 et 6
RPB6	Appartient à la structure qui stabilise la protéine sur la chromatine
RPB7	Appartient à la structure qui stabilise la protéine sur la chromatine Interagit avec RPB1 et 5
RPB8	Interagit avec les sous-unités RPB1 à 3, 5 et 7
RPB9	Constitue, avec la sous-unité RPB1, la crevasse dans laquelle la chromatine se loge lors de la transcription
RPB10	Interagit avec RPB1 à 3 et 5
RPB11	Est constitué de 3 sous-unités POLR2J, POLR2J2 et POL2J3
RPB12	Interagit avec RPB3

L'activité et le niveau de transcription de la Pol II est contrôlée par les TFs et cofacteurs. Cette régulation intervient une fois la Pol II fixée sur le promoteur.

La Pol II peut être décrite selon trois états différents. En pause, la Pol II est en attente au niveau du promoteur proximal[37]. Cet état intervient après l'initiation de la transcription de quelques nucléotides (~30-50pb). Dans cet état, la transcription ne peut être déclenchée qu'en présence de TFs. Cette pause au niveau du promoteur proximal se produit sur plus de 70 % des gènes métazoaires. Elle permet le contrôle de la transcription des gènes régulés par l'environnement. Dans le cas de certains gènes intervenant dans le développement, la Pol II en pause adopte un état particulier appelé *poised*[38] (en position). La Pol II active est en cours de transcription. La Pol II active transcrit les gènes en ARN.

L'état d'activité de la Pol II est, en grande partie, dû à sa sous-unité RPB1 et plus précisément à son domaine C-terminal (CTD). Ce domaine est composé de 52 répétitions d'un motif de sept résidus d'acide aminé, Tyr–Ser–Pro–Thr–Ser–Pro–Ser. Il est projeté à l'extérieur du complexe transcriptionnel[39]. Cette accessibilité permet des modifications des acides aminés du CTD par des phosphorylases et déphosphorylases. Quand la Pol II est recrutée par le complexe d'initiation de la transcription (CID) il est hypo-phosphorylé. Il est mis en pause par le facteur de transcription générale (GTF) TFIIH qui phosphoryle le cinquième résidu sérine du CTD. Au niveau des gènes actifs, le deuxième résidu sérine est phosphorylé par P-TEFb ce qui permet l'élongation de l'ARN. La Pol II se met en pause si le deuxième résidu de sérine est déphosphorylé (Figure I.6).

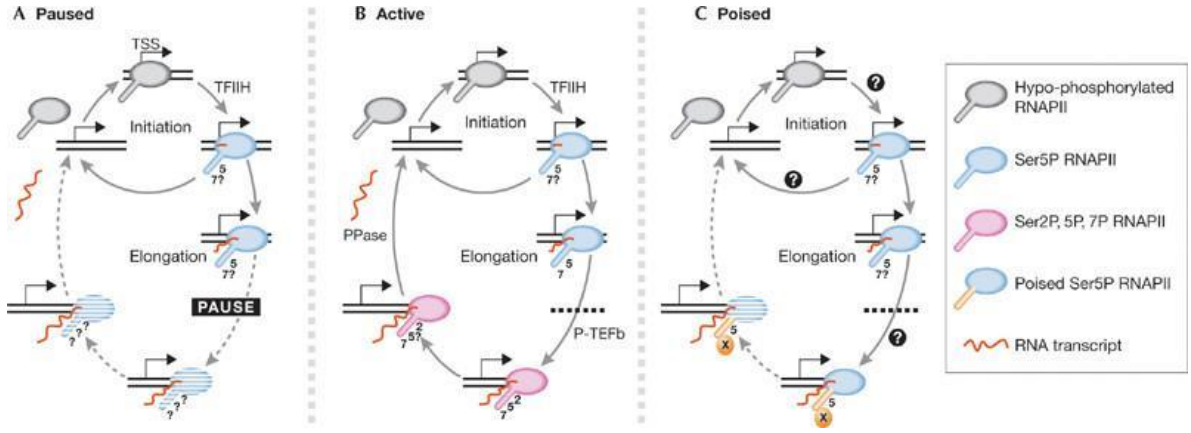


Figure I.6 ***: Les différents états de la Pol II.***
L'extrémité N-ter de la Pol II est modifiée par des protéines permettant d'activer ou de réprimer la transcription. ***A****. La Pol II est mise en pause par la phosphorylation de son 5ème résidu de sérine.* ***B****. La Pol II est activée par la phosphorylation de sa sérine 2, 5 et 7.* ***C****. La Pol II est arrêtée "en position" par la phosphorylation de son 5ème résidu de sérine.*

Figure I.6 : Brookes, E. & Pombo, A. Modifications of RNA polymerase II are pivotal in regulating gene expression states. *EMBO Rep.* **10**, 1213–1219 (2009).

4. Coopération entre les protéines régulatrices

Les TFs possèdent également un domaine d'activation leur permettant de fixer d'autres protéines appelées cofacteurs. Ces protéines sont des corépresseurs, coactivateurs de la transcription. Ces cofacteurs ne se fixent pas directement sur la chromatine. Ils n'interviennent donc qu'après la fixation des TFs. Le recrutement de protéines par les TFs permet la création de complexes protéiques comme le PIC. Ce recrutement joue, de plus, un rôle important dans la modification de la conformation de la chromatine (partie I.A.1.c. Modification de la conformation de la chromatine). La colocalisation des protéines peut se faire sur des TFs éloignés de plusieurs milliers de paires de bases, voire sur des chromosomes différents grâce à la formation de boucles d'ADN[40].

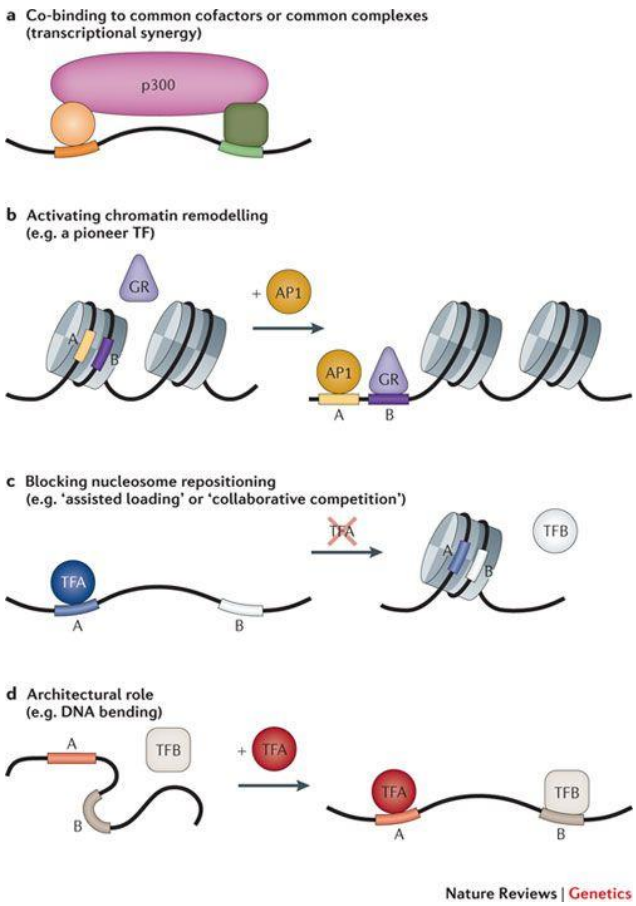


Figure I.7 ***: Différents types de collaboration entre protéines régulatrices de la transcription.***
a*- Plusieurs TFs peuvent coopérer afin de recruter un cofacteur. C'est le cas pour P300.* ***b****- Les protéines peuvent remodeler la chromatine permettant ainsi la fixation d'autres TFs, comme les facteurs pionniers.* ***c****- Les TFs peuvent empêcher la compaction de la chromatide par le déplacement des nucléosomes.* ***d****- Les TFs peuvent induire des modifications locales de conformation de la chromatine. Ces modifications peuvent aider ou prévenir la fixation de TFs proches.*

Figure I.7 : Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).

En général, la colocalisation est le résultat de la coopération au sein des protéines régulatrices de la transcription[41]. Plusieurs modèles de coopération ont été décrits. Dans le modèle additif, le niveau de transcription du gène est proportionnel à la concentration d'une protéine. On peut citer, entre autres, le facteur nucléaire κB (NF-κB) dont la concentration est proportionnelle à la transcription des gènes cibles[42]. Cette coopérativité peut être non linéaire. Elle peut agir comme un bouton on/off pour la transcription comme, par exemple, lors du développement de l'embryon de *D. melanogaster*[43,44].

La coopération entre TFs ne nécessite pas toujours d'interaction directe protéine/protéine. Les TFs pionniers, comme AP1 et FOXA1, peuvent modifier l'accessibilité de l'ADN libérant ainsi les TFBS[45,46]. Ces mécanismes sont listés dans la Figure I.7.

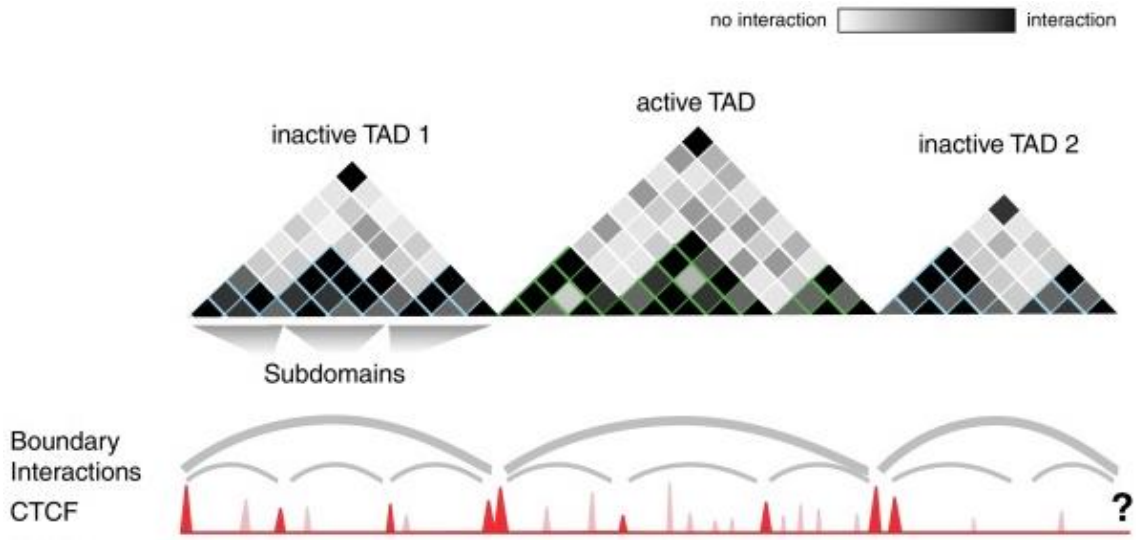


Figure I.8 ***: Visualisation du résultat d'une expérience de Hi-C mettant en avant les boucles maintenues par CTCF (cohésine) et les interactions entre loci.***
Le Hi-C est une expérience capable de capturer toutes les interactions chromatine/chromatine le long du génome à une résolution de base de 1Mb[47]. Les TAD forment des superstructures limitées par CTCF permettant d'isoler des régions d'ADN. Les régions situées au niveau des anneaux de cohésine interagissent entres elles. Les régions dans la boucle d'ADN ont une grande proximité ce qui facilite les interactions entres protéines régulatrices. L'interaction entre deux TAD est limitée.

Figure I.8 : Ali, T., Renkawitz, R. & Bartkuhn, M. Insulators and domains of gene expression. *Curr. Opin. Genet. Dev.* **37**, (2016).

Ces interactions sont également responsables de la formation de boucles de chromatine reliant plusieurs loci du génome. Chez l'homme, ces structures 3D sont maintenues par le TF CCCTC-binding[48] (CTCF) et la cohésine. CTCF peut fixer deux brins de chromatine entres eux et former une boucle. En collaboration avec RAD21 (cohésine) CTCF, il forme un complexe délimitant les territoires génomiques ainsi que les domaines topologiquement associés[49] (*Topologically Associated Domain*, TAD, Figure I.8). Ces TAD jouent un rôle majeur dans la régulation de la transcription. Des observations ont montré que la chromatine et les éléments régulateurs interagissent préférentiellement avec la chromatine située dans le même TAD[50]. Les gènes situés dans un même TAD sont généralement régulés par les mêmes TFs et sont donc actifs dans le même contexte. De plus, CTCF permet de faciliter les interactions entre TFs situés sur des loci distants situés au niveau des anneaux de cohésine (Figure I.10).

B. Les éléments régulateurs de la transcription

Les éléments régulateurs sont des régions du génome qui jouent un rôle dans la régulation de la transcription. Les éléments cis-régulateurs (CRE) sont des régions d'ADN non codant qui modulent la transcription des gènes environnants par la fixation de TFs.

Les CREs sont hétérogènes dans leur structure et dans leur fonction. Suivant leur fonction, ils sont classés en différents types :

- Les promoteurs
- Les enhancers/silencers
- Les insulateurs

1. Promoteur

Les promoteurs sont des régions situées en amont des gènes module la fixation du complexe de transcription. Leur structure permet la fixation de la Pol II et peut être décrite en deux régions : Le promoteur de base et le promoteur proximal. La composition des promoteurs peut varier d'un promoteur à l'autre, mais leur sens est important et dicte la direction de la transcription[51].

a. Promoteur de base (*core promoter*)

Le promoteur de base se situe en amont du TSS et inclut ce dernier. Il occupe les positions de -40 à +1 par rapport au TSS. Il contient les éléments minimaux nécessaires à la transcription des gènes[26]. Ces éléments minimaux, du plus proche au plus éloigné du TSS sont : La séquence Inr, ou initiator, comprenant le TSS, le TFIIB Recognition Element (BRE) et la TATA-*box* (Figure I.9). Ils permettent la fixation des GTFs. D'autres éléments peuvent être présents mais ne sont pas nécessaires à la transcription[52]. Le promoteur de base peut aussi contenir des îlots CpG (partie I.A.2. Initiation de la transcription).

b. Promoteur proximal

Le promoteur proximal comprend les régions situées en amont des gènes et du promoteur de base. Il se trouve en position -250pb par rapport au TSS. Le promoteur proximal est composé de sites capables de fixer les TFs. Ces sites de fixations se retrouvent majoritairement dans deux régions différentes. La première région est la boîte CCAAT (CAAT-*box*) à la position -100 à -60. Cette boîte est principalement présente chez les promoteurs de gènes codant pour des protéines spécifiques à chaque type cellulaire. La composition de ces régions est donc spécifique aux lignées cellulaires. La deuxième région est la boite CG (GC-*box*), Région riche en nucléotides C et G. Elle joue un rôle dans l'activation de la transcription (paragraphe I.A.2 Initiation de la transcription sur les îlots CpG).

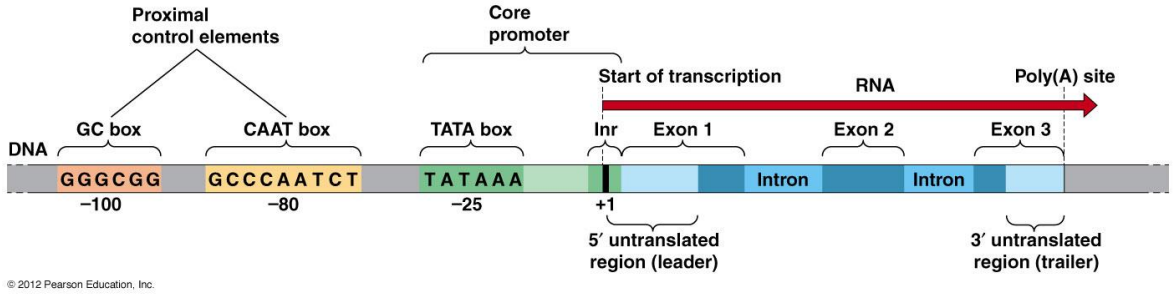


Figure I.9 ***: Exemple de promoteur de base et de promoteur proximal minimal.***
Les gènes sont transcrits à partir du TSS jusqu'au site de terminaison de la transcription (Transcription Termination Site, TTS). Le pré-ARNm contient les introns et exons, ainsi que les régions non traduites. Au niveau du TSS on trouve la région Inr (Initiator). Vers la position -25 on trouve la TATA box. Ces deux éléments forment le promoteur de base (core promoter). A la position -80 on trouve la CAAT box e à -100, la CG box. Ces deux boîtes forment le promoteur proximal.

2. Les enhancers et silencers

Les enhancers et silencers sont des régions régulatrices présentes en amont ou en aval des gènes. Ces éléments distaux sont capables d'influencer le niveau de transcription des gènes qui leur sont liés. Les enhancers activent la transcription des gènes, et les silencers la répriment. Un élément distal peut influencer la transcription de plusieurs gènes. Les enhancers sont définis comme des régions comprises entre 200 et 500 pb à des positions et distances variables par rapport à leur gène cible. Ils peuvent influer la transcription de gènes jusqu'à 1 Mbp[53]. Les enhancers et silencer sont spécifiques aux types cellulaires[54], à l'étape du développement de l'organisme[55], aux types de gènes régulés[56].

Figure I.9 : Chetry, L. B. Molecular Cell Biology 5th ed - Lodish et al.

L'activation des enhancers et silencers est le résultat de l'ouverture de la chromatine, ce qui permet la fixation des protéines régulatrices. L'activité de ces régions distales est liée au type cellulaire auquel elles appartiennent[57]. De plus, l'activation d'un enhancer ou d'un silencer peut varier au cours du temps, en fonction de stimuli extérieurs[58]. Prenons l'exemple du récepteur nucléaire NR3C4 (nuclear receptor subfamily 3, group C, member 4). C'est une protéine activée par la fixation d'hormones androgéniques dans le cytoplasme. L'activation de NR3C4 provoque son transport dans le noyau et sa dimérisation. Une fois dans le noyau, NR3C4 agit comme un TF. Il peut donc se fixer sur des enhancers spécifiques et provoquer l'activation de la transcription des gènes liés[59]. En conclusion, l'activation des enhancers et silencers permet une régulation spatiale et temporelle de la transcription.

Les enhancers et silencers n'ont pas de séquences ou éléments consensus connus mais possèdent de nombreux TFBS[60]. La majorité de ces éléments distaux possède un environnement épigénétique particulier. Les enhancers se trouvent sur de la chromatine ouverte possédant des marques d'histones H3K27ac[61]. Ils possèdent souvent des îlots CpG[62] et fixent le co-activateur EP300-CPB[63].

Les enhancers ont été décrits pour la première fois en 1981 par *Benerji et al*. et *Moreau et al.* Ils ont observé que les cellules dont une β-globine était recombinée par le plasmide du virus *Simian vacuolating virus 40* (SV40) avaient des niveaux d'expression de ce gène beaucoup plus forts que les cellules à l'état sauvage. Ils en ont déduit qu'une séquence introduite de SV40 amplifie l'expression de la β-globine.

Les enhancers et silencers peuvent réguler la transcription de gènes éloignés. Ils peuvent se situer, en amont ou en aval des gènes, dans leurs introns[64], voire sur un chromosome différent[40]. La distance entre les promoteurs et les éléments distaux régulateurs de la transcription ne semble pas jouer de rôle majeur dans l'efficacité des enhancers et des silencers. Cette distance est réduite par la formation de boucles d'ADN. Il a été observé que l'ADN forme des boucles entre les enhancers/silencers et les promoteurs cibles[65]. Ces boucles permettent l'interaction entre des loci distants. Il est même possible que ces boucles se regroupent en "usine à transcription"[66] maintenue par les interactions entre protéines régulatrices (Figure I.10). Ces structures regroupent de nombreuses boucles d'ADN ce qui augmente la concentration des protéines régulatrices de la transcription, facilitant ainsi leurs interactions. Ces usines à transcription permettraient également le contact entre plusieurs loci, ce qui expliquerait qu'un seul enhancer ou silencer puisse réguler plusieurs gènes[67].

Silencers et enhancers sont similaires et partagent la plupart de leurs propriétés, à la différence que les silencers régulent négativement les gènes qui leur sont liés. Certains silencers peuvent, selon les observations, devenir des enhancers et inversement. Ce constat m'a conduit, tout au long de mon travail, à ne pas différencier enhancers et silencers et à utiliser, dans ce manuscrit le terme enhancer pour désigner à la fois enhancer et silencer.

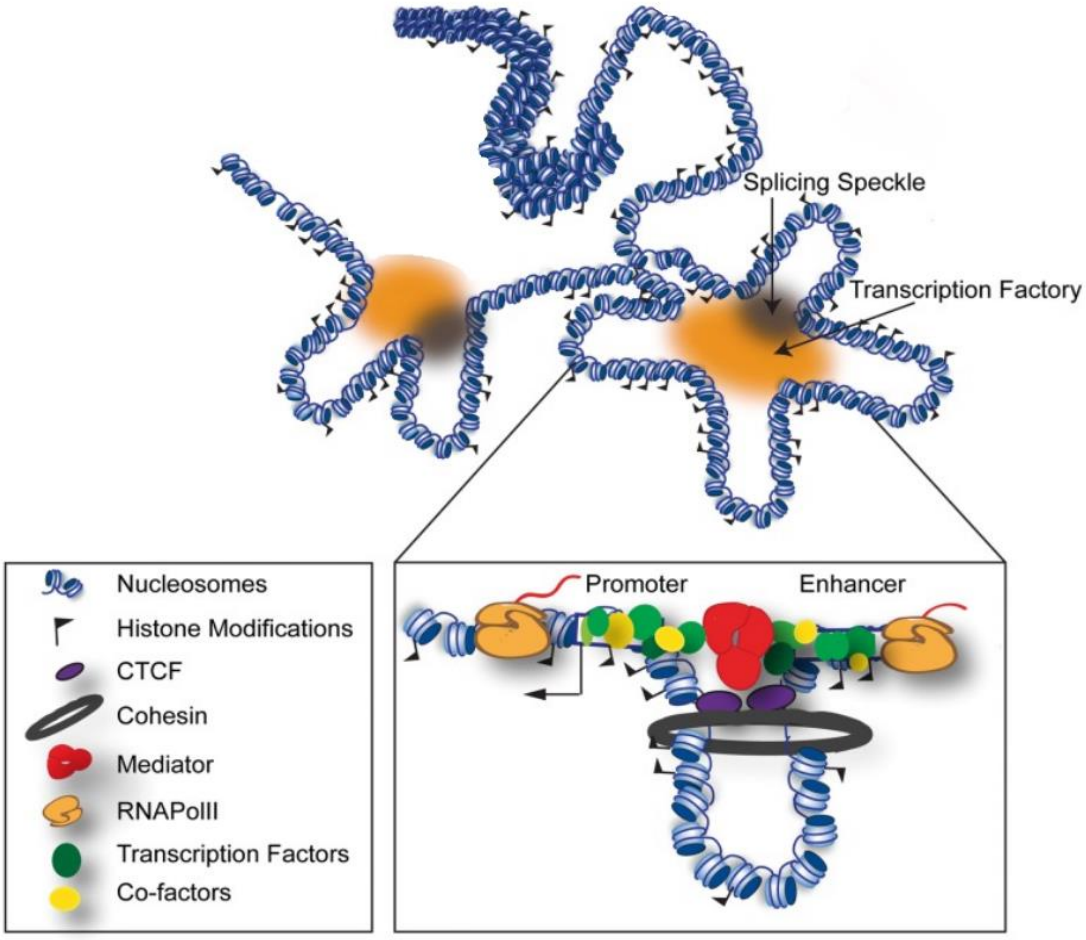


Figure I.10 ***: Structure d'une usine à transcription***
Les usines à transcription (Transcription Factories) regroupent plusieurs boucles d'ADN permettant le rapprochement d'un grand nombre de loci. Les boucles d'ADN permettent le contact entre les enhancers et les promoteurs assurant ainsi une colocalisation des protéines régulatrices de la transcription. Ces boucles sont maintenues par des anneaux de cohésine placés grâce à la présence du TF CTCF.

a. Les enhancers transcrits

En 2010, la transcription par la Pol II a été observée au niveau de certains enhancers[68]. La transcription au niveau de ces enhancers est corrélée avec la transcription d'ARNm des gènes voisins[69]. La transcription au niveau d'enhancers permet d'identifier des enhancers actifs. Ce groupe d'enhancers transcrits partage les mêmes caractéristiques épigénétiques que les enhancers actifs : ils se situent sur de la chromatine ouverte enrichie en histone H3K27ac et H3K4me1 (Figure I.11). Les enhancers transcrits produisent de courts ARN non épissés, coiffés et non polyadénylés (eRNA). Leur taille est comprise entre 0,5 et 2kb.

Figure I.10 : Mora, A., Sandve, G. K., Gabrielsen, O. S. & Eskeland, R. In the loop: promoter–enhancer interactions and bioinformatics. *Brief. Bioinform.* **17**, bbv097 (2015).

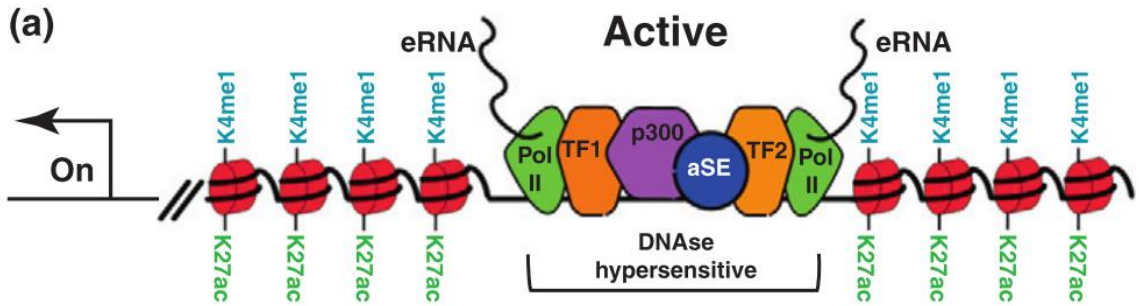


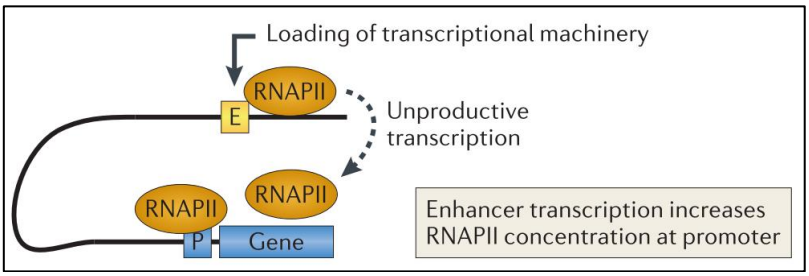
Figure I.11 ***: Profil d'un enhancer transcrit.***
La chromatine des enhancers transcrits est ouverte. Ils fixent des TFs et, en particulier, la protéine P300. Les enhancers fixent la Pol II qui transcrit bidirectionnellement l'ADN en eRNA. Ces eRNA sont coiffés et non polyadénylés. Les enhancers transcrits sont encadrés par des histones H3K27ac et H3K4me1.

L'absence de protection en 3' par la queue polyadénylée rend les eRNA sensibles aux exosomes, responsables de la dégradation des eRNA dans le noyau. Cette dégradation rapide, couplée à un taux de transcription des enhancers beaucoup plus faible que celui des gènes environnants, rend leur étude difficile. La cinétique de la transcription a été étudiée. Il semblerait que la transcription des enhancers précède celle des gènes qui leur sont liés[70]. Il est possible que la transcription débute au niveau des enhancers et non au niveau des promoteurs proximaux.

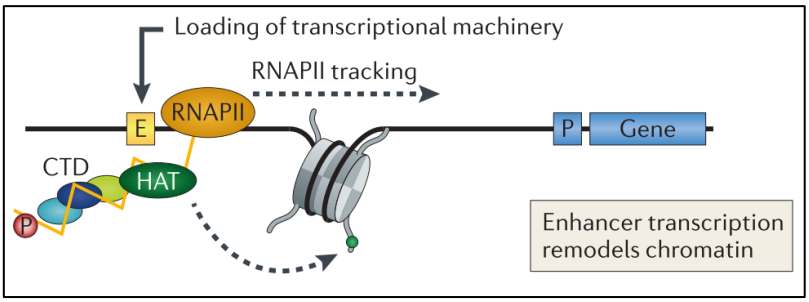
Plusieurs modèles ont été proposés afin d'expliquer la corrélation entre la transcription de certains enhancers et la transcription des gènes proches (figure I.12). Certaines études se sont intéressées à la fonction des eRNA. L'une d'entre elle montre que les eRNAs faciliteraient l'activation de la Pol II en capturant la protéine NELF (facteur de répression de l'élongation de l'ARN[71]). Cette capture de protéine semblerait aussi réguler positivement la fixation de TFs[72]. D'autres observations montrent que certains eRNA peuvent stimuler l'activation d'enzyme comme le coactivateur CBP[73]. Pour l'instant, aucune fonction globale n'est proposée quant à la transcription des enhancers. Certaines études mettent en doute l'existence de fonctions propre aux eRNA. Une étude réalisée par Rahman *et al* n'a pas trouvé de lien entre la présence d'eRNA au niveau des boucles enhancer-promoteur et la transcription des gènes présents.

Figure I.11 : Meng, H. & Bartholomew, B. Emerging roles of transcriptional enhancers in chromatin looping and promoter-proximal pausing of RNA polymerase II. *J. Biol. Chem.* **293**, 13786–13794 (2018).

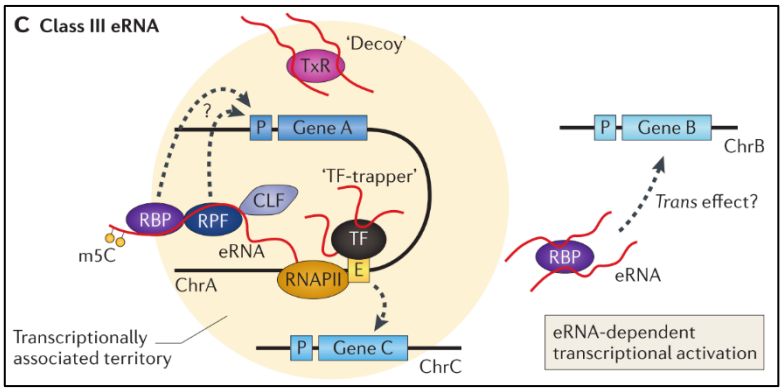
1.



2.



3.



4.

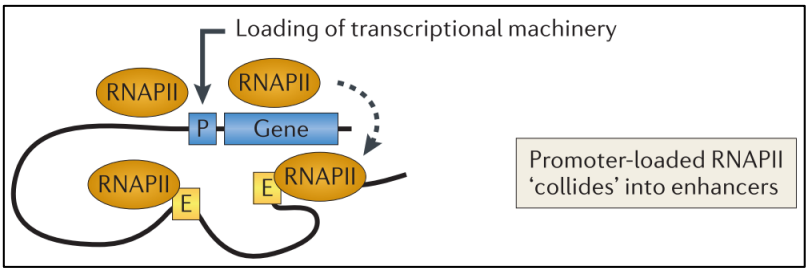


Figure I.12 ***: Fonction de la transcription au niveau des enhancers***
1. *La machine transcriptionnelle se fixe sur les enhancers. La Pol II active est ensuite transférée sur les gènes cibles.* ***2.*** *La Pol II fixe des protéines de remodelage de la chromatine au niveau de son CTD. Au fur et à mesure du déplacement de la Pol II le long de l'ADN, les protéines modifient les histones, permettant ainsi l'ouverture de la chromatine au niveau des gènes.* ***3.*** *Les eRNA capturent les protéines régulatrices situées au niveau de l'enhancer. Ce complexe eRNA/protéine permet le chargement des protéines sur les promoteurs par proximité ou par transport actif.* ***4.*** *La transcription au niveau des enhancers n'est due qu'à la proximité entre les promoteurs actifs et les enhancers. La transcription au niveau des enhancers ne serait qu'un "bruit" de la transcription.*

Figure I.12 : Li, W., Notani, D. & Rosenfeld, M. G. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat. Rev. Genet.* **17**, 207–223 (2016).

3. Les insulateurs

Les insulateurs sont des régions génomiques limitant les interactions entre différentes régions de la chromatine[74]. Pour cela, les insulateurs modifient la structure de la chromatine, empêchant le contact entre les promoteurs et les enhancers, et, par conséquent, l'initiation de la transcription. Différents modes de fonctionnement des insulateurs sont présentés dans la figure I.13.

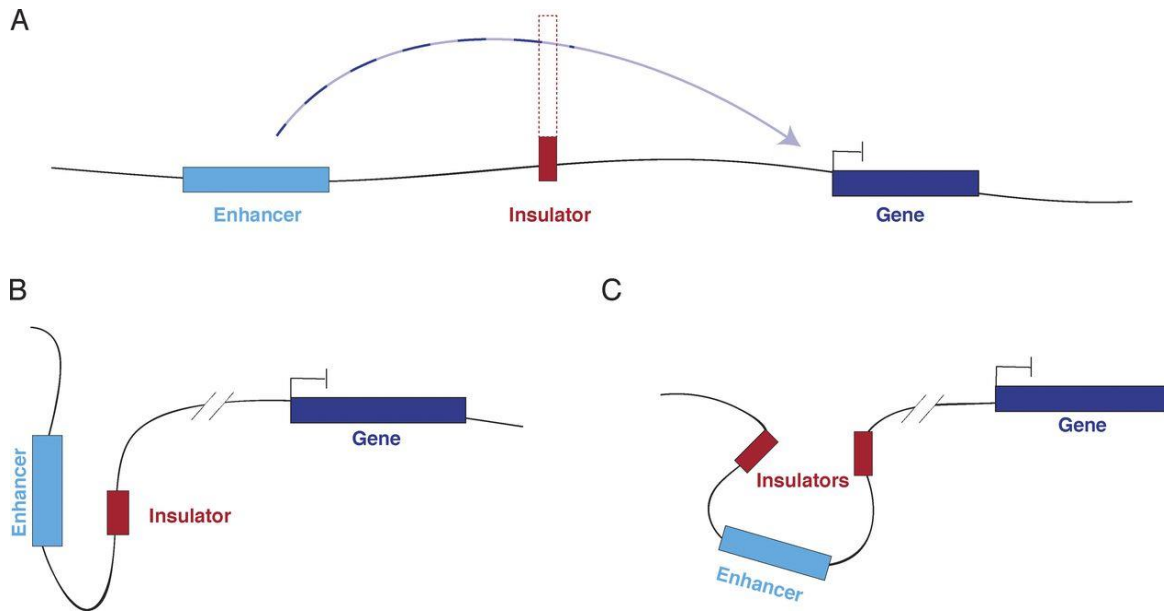


Figure I.13 ***: Action des insulateurs***
A. *Les insulateurs forment une barrière physique entre différents cis-régulateurs : Les insulateurs situés aux limites des TADs empêchent les interactions avec la chromatine de part et d'autre de l'insulateur.* ***B.*** *Compétition de fixation : Les insulateurs se lient aux cis-régulateurs empêchant la fixation de TFs sur les enhancers.* ***C.*** *Emprisonnement des CREs : les insulateurs forment des boucles courtes isolant les éléments cis-régulateurs des autres loci du génome.*

Figure I.13 : Vermunt, M. W., Zhang, D. & Blobel, G. A. The interdependence of gene-regulatory elements and the 3D genome. *J. Cell Biol.* **218**, 12–26 (2019).

II Régulation chez les plantes : différences et points communs

La transcription des gènes et leur régulation sont similaires chez la plupart des eucaryotes. Chez les plantes, la transcription est, elle aussi, assurée par la Pol II et régulée par les TFs. La fixation des TFs est soumise aux mêmes contraintes que chez l'Homme, à savoir, l'ouverture de la chromatine et la méthylation de l'ADN. Une différence se situe au niveau de la méthylation de la cytosine. Chez les plantes, la méthylation ne se produit pas seulement au niveau des nucléotides CpG. La cytosine peut être modifiée au niveau des sites CpG, CpHpG, et CpHpH (H représentant n'importe quel nucléotide). Chez Arabidopsis Thaliana, 14% des cytosines sont méthylées[75] contre 75% chez les mammifères[76].

Arabidopsis Thaliana est un organisme modèle chez les plantes. C'est une plante à fleurs de la famille des (*Brassicaceae* moutarde). *Arabidopsis Thaliana* a un génome nucléaire de 135 Mbp (parmi les plus petits chez les plantes) ce qui en fait un organisme modèle idéal pour l'étude des gènes et de leur régulation[77] (Arabidopsis Information Resource, TAIR). Une des conséquences de la petite taille du génome est la compaction de ses gènes et de ses éléments régulateurs.

A retenir pour mes travaux

Ce qu'il faut retenir : Chez l'Homme, la transcription des gènes est assurée par l'ARN polymérase II (Pol II). Le recrutement de la Pol II sur la chromatine, son activation et sa "pause" sont assurés par les protéines régulatrices de la transcription. Cette régulation de l'activation définit le niveau de transcription des gènes. Les protéines régulatrices peuvent être classées en trois catégories. Les facteurs généraux de la transcription (GTF) qui recrutent la Pol II au niveau des gènes, les Facteurs de Transcription (TF) qui se fixent directement sur l'ADN et les cofacteurs qui se fixent sur les TFs. Les TFs et les cofacteurs peuvent collaborer pour activer ou réprimer la transcription. Les activateurs et répresseurs peuvent agir : sur l'accessibilité de la chromatine en méthylant et acétylant les histones, sur l'activité de la Pol II en phosphorylant cette enzyme et en créant des boucles de chromatine permettant d'assurer la colocalisation des protéines régulatrices entres-elles.

Les TFs se fixent sur les sites de fixation des facteurs de transcription (TFBS). Les TFBS sont concentrés dans des régions spécifiques du génome, les éléments cis-régulateurs de la transcription (*cis-regulatory element,* CRE). Les promoteurs se situent en amont des gènes. Bien qu'ils présentent des éléments typiques (TATA-*box*, ilot CpG, etc.), leurs compositions exactes varient. Les enhancers et les silenceurs peuvent activer ou réprimer la transcription des gènes. Leurs positions varient grandement par rapport à leur gène cible (intron à plusieurs mégabases du gène). Leur activité dépend du type et du contexte cellulaire et peut varier au cours du temps. La formation de boucle ADN, maintenue par CTCF et RAD21, permet un rapprochement entre les gènes cibles et les enhancers/silencers. Les enhancers sont transcrits par la Pol II. Les insulateurs séparent le génome en TAD (Topologically Associated Domain). Les gènes interagissent préférentiellement avec les loci d'un même TAD.

III Annotation des éléments régulateurs

La complexité des processus biologiques chez le vivant est assurée, en grande partie, par la régulation de la transcription. Cette régulation s'effectue grâce à la coopération spatiale et temporelle entre les protéines régulatrices (régulation en trans) et les régions régulatrices du génome (régulation en cis). Il s'avère nécessaire d'identifier les CREs dans les différents organismes et lignées cellulaires afin de mieux connaître les mécanismes du vivant. A long terme, une meilleure compréhension du fonctionnement de la régulation de l'expression des gènes contribuera à l'identification des mécanismes d'action de nombreuses maladies et, donc, à développer des moyens d'actions permettant de les guérir.

Dans le but d'identifier les CREs, plusieurs groupes ont, dans les années 1990, lancé des recherches sur les séquences nucléotidiques composant le génome humain. Ces recherches ont abouti en 2001 à une carte complète du génome humain[78,79]. Cette carte du séquençage humain a permis d'identifier les régions codantes pour des protéines, mais aussi d'étudier les régions non codantes du génome. Avant le séquençage du génome humain, 98% du génome était considéré comme "junk DNA", c'est à dire non-codant et n'ayant pas de fonctions[80]. Cette théorie a été remise en question par le consortium ENCODE qui, grâce au séquençage du génome humain, a détecté que 80% du génome humain avait une activité biochimique[81]. La difficulté est maintenant d'identifier les régions fonctionnelles du génome non-codant. Les éléments régulateurs font partie de ces régions non-codantes.

La difficulté d'identifier et d'annoter les régions régulatrices à l'échelle du génome vient du fait qu'elles ne possèdent pas ou très peu de structures génomiques stables et connues. Même les promoteurs, qui sont les seules régions régulatrices possédant des éléments connus et caractérisés, ne possèdent pas de structure consensus (partie I.B.1. Promoteurs). L'identification des éléments régulateurs et leurs fonctions, nécessite de développer des méthodes expérimentales et informatiques ne se basant pas uniquement sur la recherche de séquences connues.

A. Méthodes expérimentales

1. Méthode à bas débit

Gène rapporteur

Une méthode expérimentale répandue permettant d'identifier les éléments régulateurs est la méthode des gènes rapporteurs[82]. Le but de cette méthode est de mesurer l'expression d'un gène en fonction de son contexte génomique direct. Pour cela, un gène, dont l'expression peut être suivie facilement, et une région du génome contenant un élément régulateur sont insérés dans un plasmide (ex : luciférase, green fluorescent protein : GFP, etc.). La comparaison entre l'expression du gène en présence ou en l'absence de la région régulatrice peut être observée de façon directe (Figure III.1). La méthode du gène rapporteur permet d'étudier la plupart des éléments régulateurs, le promoteur proximal et/ou distal, les enhancers, les silencers et les insulateurs, mais elle comporte certaines limites.

In vivo, il peut être difficile d'identifier les limites des éléments régulateurs à la paire de base près. En conséquence, il est compliqué de vérifier si l'élément régulateur a bien été inséré complètement dans le plasmide. Un plasmide contenant un élément régulateur incomplet ne rend pas compte de l'influence réelle sur l'expression du gène cible.

Les gènes rapporteurs ne prennent pas non plus en compte le contexte génomique complet, comme par exemple l'accessibilité de l'élément régulateur ou du gène, la méthylation du gène, etc.

De plus, rien ne permet de vérifier si l'élément régulateur est bien utilisé *in vivo*, ni dans quel contexte biologique il est activé. Certains éléments régulateurs, tels que les enhancers ne sont actifs qu'en présence de certaines protéines, et rien n'assure que ces protéines activent l'enhancer dans le contexte du gène rapporteur.

En conclusion, la méthode du gène rapporteur permet d'observer l'action de la région régulatrice sur le gène d'intérêt. Mais il est nécessaire d'utiliser d'autres méthodes en parallèle afin d'identifier et d'annoter les éléments régulateurs dans le génome humain.

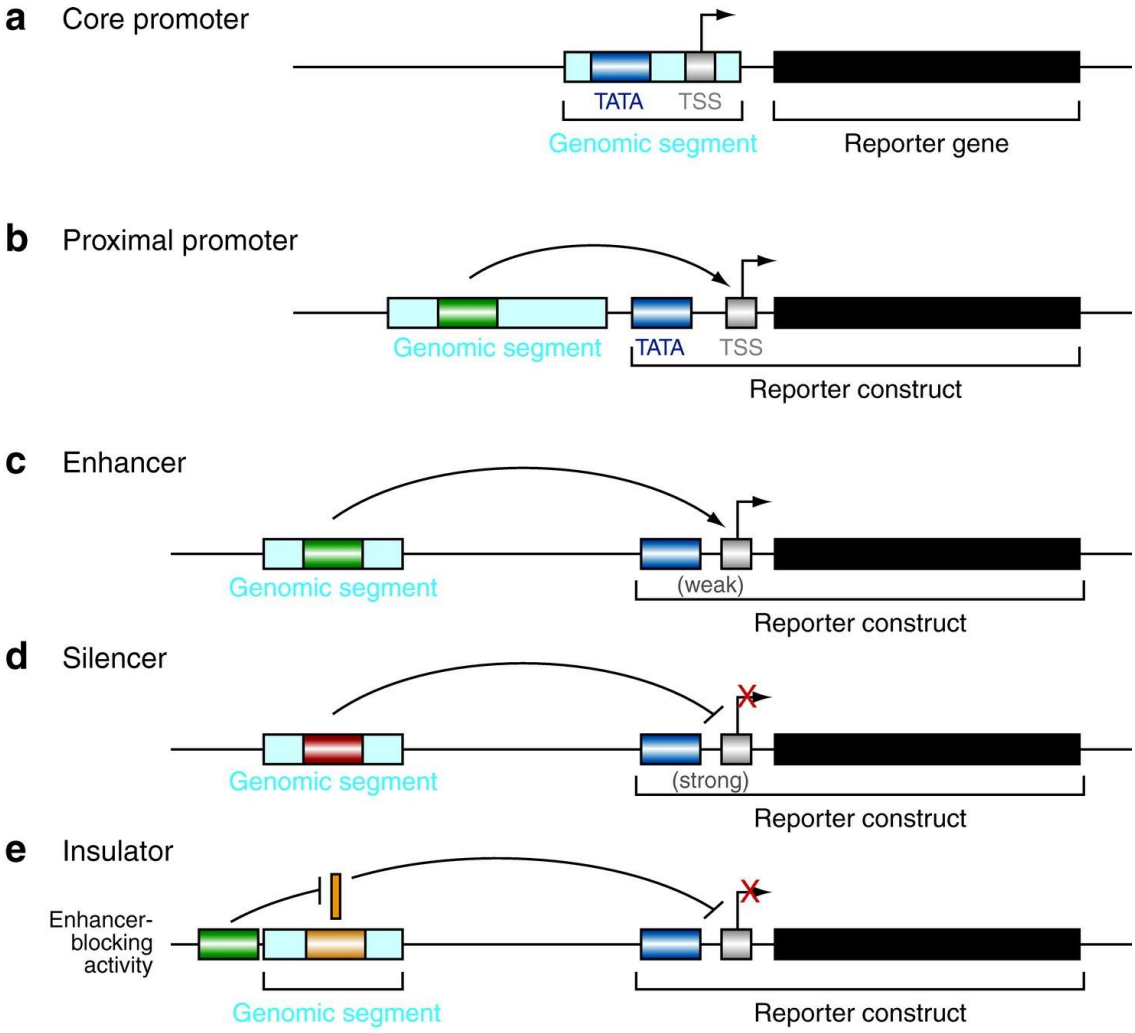


Figure III.1 ***: Exemples de construction de gènes rapporteurs.***
Dans la construction ***a*** *un promoteur de base hypothétique (core promoter) est inséré dans le plasmide au début du rapporteur (reporter gene). Dans cet exemple, la région à tester contient une TATA box et un Inr (TSS). L'expérience permet de contrôler si la région insérée permet, à elle seule, de fixer le PIC et de transcrire le gène cible (ce qui est une caractéristique des promoteurs de base. La construction* ***b*** *a pour but de tester un promoteur proximal. Dans ce cas et les suivants, la région rapportrice est composée du gène rapporteur et d'un promoteur minimal (TATA-box et Inr). Le gène est transcrit uniquement si la région insérée est un promoteur proximal. Dans le cas* ***c*** *le but est d'étudier la capacité qu'a une région d'augmenter le niveau de transcription d'un gène. Le gène rapporteur est transcrit faiblement. Si sa production d'ARNm augmente la région testée est un enhancer. La construction* ***d*** *teste les régions silencer. Le gène est fortement exprimé. Si son niveau de transcription baisse la région testée est un silencer.* ***e*** *Un des moyens d'action des insulateurs est de former une barrière physique entre les enhancers et le promoteur de son gène cible. Afin de contrôler ce type de régulation, la construction est composée d'un enhancer séparé de son gène cible par un insulateur potentiel. Si le gène est exprimé, on en déduit que la région testée n'est pas un insulateur.*

Figure III.1 : 1Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional Regulatory Elements in the Human Genome. *Annu. Rev. Genomics Hum. Genet.* **7**, 29–59 (2006).

2. Séquençage à haut débit

Le séquençage à haut débit a permis de réaliser des expériences à l'échelle du génome, et cela, à faible coût. Les méthodes à haut débit sont basées sur la parallélisation du processus de séquençage de millions de fragments d'ADN ou ARN par des machines de séquençage. Plusieurs types d'expériences peuvent être couplées au séquençage à haut débit ce qui permet d'étudier différentes facettes du génome comme la conformation de la chromatine, la fixation de protéine ou le niveau de transcription des gènes. L'émergence de ces technologies a rendu indispensable l'utilisation d'outils informatiques et statistiques afin d'analyser les millions de séquences produites par ces machines. La taille, le nombre de fragments séquencés, le coût et le temps par fragment dépendent de la technique de séquençage utilisée.

a. Techniques de séquençage à haut débit

Les premières techniques de séquençage à haut débit sont apparues sur le marché en 2005. Elles se basent sur le pyroséquençage[83] développé par la société 454 Life sciences et le séquençage SOLiD développé par Applied Biosystems (AB SOLiD). Les méthodes de séquençage à haut débit permettent d'identifier simultanément les séquences de millions de fragments d'ADN et, ce, à un coût de plus en plus faible. Ces techniques permettent de séquencer à des échelles allant de populations de cellules (bulk sequencing) à la cellule individuelle (single-cell sequencing). La constitution de la librairie de fragments à séquencer diffère en fonction de la méthode utilisée.

Le bulk sequencing est apparu en premier. Sa popularité a conduit au développement de plusieurs plateformes concurrentes utilisant des techniques différentes visant à séquencer toujours plus, tout en baissant le coût par nucléotide. De nos jours, c'est la plateforme Illumina© qui domine le marché pour le séquençage à haut débit de masse. La plateforme Illumina est basée sur un système optique de coloration de séquences. Plusieurs variantes de cette technique permettent différentes échelles d'analyse (Tableau III.1).

L'ensemble de ces méthodes présente un inconvénient. La qualité de lecture baisse au cours du séquençage de la librairie. C'est un facteur important qui limite la taille des fragments de lecture. Ces erreurs de lecture peuvent aussi survenir à d'autres moments. Pour cette raison, la plupart des séquenceurs attribuent un score de confiance pour chaque paire de bases séquencée.

Le bulk sequencing présente des biais techniques. Le plus fréquent est la sous-représentation des fragments ayant des compositions nucléotidiques extrêmes[84]. Ce biais est inhérent à l'amplification par PCR (Polymerase Chain Reaction) et conduit à la sur-représentation d'une population de fragments de lecture. Il est possible d'identifier les régions sensibles à la PCR pour chaque cellule en réalisant une expérience de contrôle. L'ADN ou l'ARN total brute est séquencé, ce qui permet d'identifier les régions sur-représentées en tant que "bruit de fond" (*Whole Cell Extract ou WCE*).

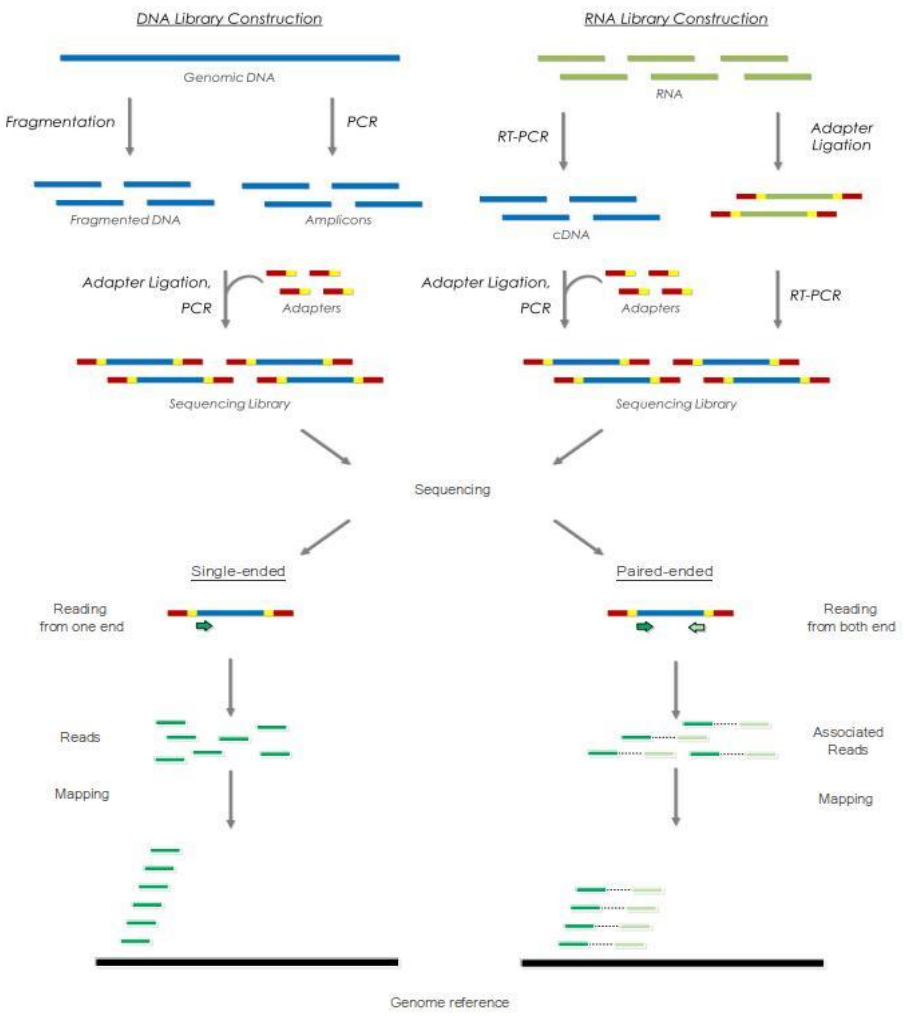


Figure III.2 ***: Schéma de deux expériences de séquençage à haut-débit.***
La première étape du séquençage à haut débit est la constitution de la librairie de séquençage. Pour l'ADN, il faut tout d'abord fragmenter le génome complet ou amplifier les régions d'intérêts par PCR. Pour le séquençage du transcriptome, il faut traduire les fragments d'ARN en ADN complémentaire (cDNA) par Reverse Transcription PCR (RT-PCR) ou directement lier les adaptateurs. Les adaptateurs sont essentiels pour amplifier les fragments par PCR et RT-PCR. Ils permettent aussi l'identification et la manipulation des fragments à transcrire. Les fragments sont ensuite amplifiés par PCR ou RT-PCR puis séquencés. Plusieurs types de traitement de la librairie permettent, soit le séquençage single-end, soit le paired-end. Le séquençage single-end permet de déterminer la séquence des fragments d'une seule extrémité contrairement au paired-end qui lit les deux. Les fragments de librairie ne sont pas séquencés sur la totalité de leur longueur, seules leurs extrémités le sont. Pour le séquençage paired-end, les paires de fragments sont associées afin de recomposer les extrémités d'un même fragment. A la fin du séquençage, des fragments de lecture sont obtenus (read). Ces fragments de lecture sont ensuite alignés sur le génome pour identifier leurs régions d'origine.

Figure III.2 : Global Next Generation Sequencing Market 2018 by Manufacturers, Countries, Type and Application, Forecast to 2023 - ReportsIntellect.com. https://www.reportsintellect.com/report/178746-global-next-generation-sequencing-market-2018-by-manufacturers-countries-type-and-application-forecast-to-2023.html.

Le séquençage single-cell a récemment pris une grande importance. Il permet d'analyser les séquences d'une seule cellule à la fois. En revanche, son rendement est plus faible[85]. A l'instar du bulk sequencing, plusieurs techniques permettant de séquencer individuellement une molécule d'ADN ou d'ARN, ont été développées. Une des premières méthodes à voir le jour dans les années 90 est celle du séquençage par nanopores[86]. Elle permet de déterminer l'ordre dans lequel les nucléotides sont disposés sur un fragment d'ADN donné, à l'aide de nanopores (trous ayant un diamètre de l'ordre du nanomètre). Cette méthode est lente et coûteuse, mais permet de séquencer des fragments de lecture très longs. Des méthodes alternatives, comme PacBio RSII, permettant de séquencer à plus faible coût, ont été développées par Pacific BioSciences.

Les séquençages de masse ou le single-cell peuvent être utilisés pour observer et identifier des éléments régulateurs à différentes échelles. La technique dite de single-cell permet d'analyser des phénomènes à l'échelle d'une seule cellule et donc, d'étudier l'hétérogénéité des populations de cellules précises dans un contexte biologique bien défini. Le séquençage de masse, quant à lui, permet d'analyser une large population de cellules et de déterminer l'élément régulateur "moyen" au niveau de la population entière. Si on veut annoter le plus d'éléments régulateurs possibles, le bulk sequencing est le plus approprié.

Analyses bioinformatiques du séquençage

Quelle que soit la technique de séquençage appliquée, il est nécessaire, par la suite, d'utiliser des outils informatiques permettant l'analyse des résultats. Les fragments de la librairie d'ADN, une fois séquencés en fragments de lecture, sont alignés sur un génome/séquence de référence afin d'identifier les régions d'intérêt. De nombreux outils ont été développés afin d'aligner ces fragments de lecture. Certains sont gratuits et donnent accès à leur code source comme Bowtie[87] ou BWA[88]. D'autres sont payants comme NovoAlign[a]. Les outils d'alignement diffèrent par leur algorithme. L'algorithme utilisé influe sur leur vitesse, leur justesse d'alignement, leur sensibilité et leur spécificité. De nombreuses études comparent ces aligneurs[89–91]. Ces études mettent en lumière les différentes fonctionnalités, ainsi que les forces et faiblesses des différents outils d'alignement. Le choix de l'outil à utiliser s'opère en fonction des caractéristiques des données à analyser (ex : pour l'ARN si l'ARNm est épissé ou non, la profondeur du séquençage, le choix single-end ou paired-end, etc) et de la machine sur laquelle l'analyse doit être lancée (machine personnelle, serveur, etc.).

Tableau III.1 ***: Description des différentes méthodes de séquençage à haut débit.***

Instrument	Plage de débit (Go)	Taille des fragments de lecture	Avantage	Inconvénient
Bulk sequencing				
Illumina				
MiniSeq	1.7–7.5	1×75 à ×150	Faible investissement initial	Longueur des fragments de lecture et temps de lecture
MiSeq	0.3–15	1×36 à 2×300	Taille des fragments de lecture, scalabilité	Temps de lecture
NextSeq	10–120	1×75 à 2×150	Rendement de production	Longueur des fragments de lecture et temps de lecture
NextSeq	10–1000	1×50 à 1×250	Précision de lecture, rendement de production	Haut investissement initial
NovaSeq 5000/6000	2000–6000	2×50 à 1×150	Précision de lecture, Rendement de production	Haut Investissement initial
Séquençage single-cell				
Oxford Nanopore				
MInION	0.1–1	Jusqu'à 100kb	Taille des fragments de lecture, portabilité de l'appareil	Taux d'erreur élevé, temps de lecture
Pacific BioSciences				
PacBio RSII	0.5–1	Jusqu'à 60kb	Taille des fragments de lecture, vitesse	Taux d'erreur élevé
Sequel	5-10	Jusqu'à 60kb	Taille des fragments de lecture, vitesse	Taux d'erreur élevé

Tableau III.1 : Besser, J., Carleton, H. A., Gerner-Smidt, P., Lindsey, R. L. & Trees, E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin. Microbiol. Infect.* **24**, 335–341 (2018).

Un des aligneurs les plus utilisés chez l'homme est Bowtie 2[92] qui utilise l'indexation de Burrows-Wheeler. Il est rapide, facilement paramétrable et prend en compte la qualité du séquençage. Ces caractéristiques en font un outil intéressant pour les analyses de grandes quantités de données hétérogènes. Il produit, au final, un fichier SAM[93] qui renseigne la ou les positions d'alignements de chaque fragment de lecture. En effet, un même fragment de lecture peut s'aligner plusieurs fois le long du génome. Les fragments longs et complexes ont moins de probabilité de s'aligner plusieurs fois. A contrario, certaines régions du génome peu complexes ou répétées ont tendance à aligner beaucoup de fragments de lecture (comme les centromères et les télomères).

Les outils informatiques peuvent aussi être utilisés pour corriger des biais du séquençage. Par exemple, l'outil bioinformatique mpileup de la suite d'outils samtools[94] a été développé pour identifier les fragments de lecture alignés sur-représentés du fait du biais de PCR. Des outils de "trimming" permettent de nettoyer les fragments de lecture avant l'alignement. Ils peuvent supprimer les adaptateurs Illumina qui n'ont pas été correctement retirés et les paires de bases dont la qualité n'est pas acceptable. Au cours de ma thèse, j'ai utilisé l'outil Trim Galore. Cet outil permet de trimmer les fragments paired et single-ended. Il utilise Cutadapt[95] et FastQC[b] .Fixation des facteurs de transcription.

Les méthodes ChIP

Les analyses dérivées de l'immuno-précipitation de la chromatine (ChIP) sont parmi les plus répandues pour identifier les interactions protéines/chromatine *in vivo*[96]. Le principe du ChIP est décrit dans la figure III.3.

Le ChIP-seq est une immunoprécipitation suivie d'un séquençage à haut débit. Cette méthode permet d'identifier toutes les régions de fixation d'une protéine spécifique sur la longueur de la chromatine. Elle peut donc être utilisée pour identifier tous les éléments de la régulation répondant à la présence d'une protéine particulière ou active grâce à la présence d'une marque d'histone spécifique. Elle peut également permettre d'identifier la présence d'éléments régulateurs activés par un GTF ou activateur spécifique.

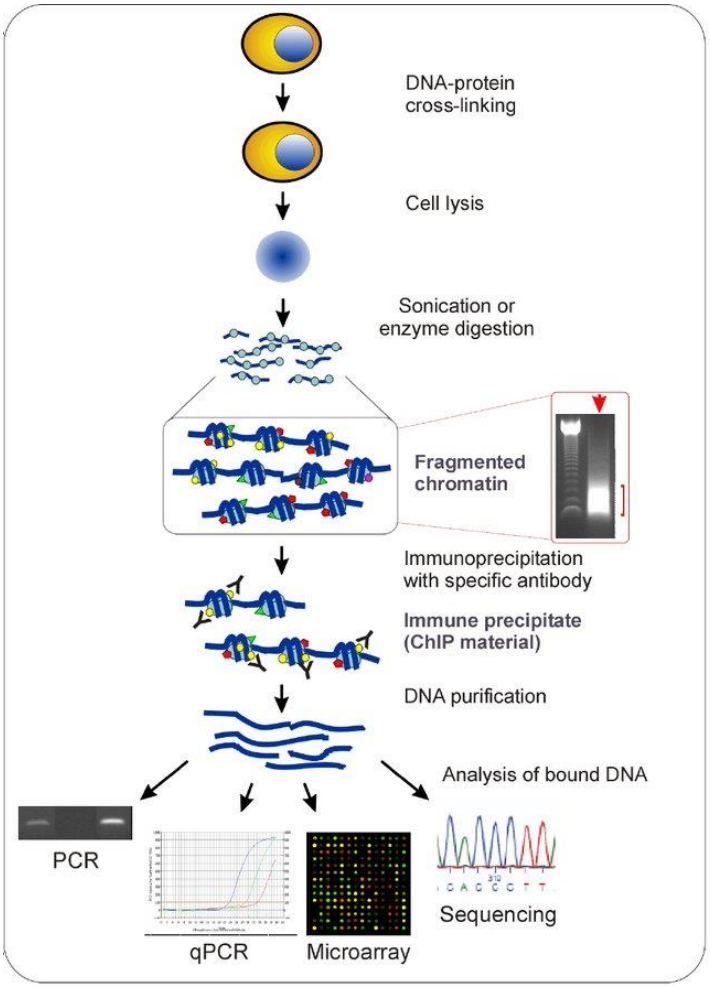


Figure III.3 ***: Étapes d'une expérience ChIP.***
Dans un premier temps, les interactions protéines/ADN sont figées par des liaisons covalentes (cross-linking). Généralement, le formaldéhyde est utilisé. Ensuite, le contenu du noyau est extrait par la lyse des cellules. Le génome, constitué des complexes protéines/ADN, est fragmenté en brins courts par sonication ou digestion enzymatique. Les fragments d'ADN complexés avec la protéine d'intérêt sont sélectionnés par l'utilisation d'un anticorps spécifique. L'ADN complexé avec la protéine d'intérêt et l'anticorps spécifique est précipité, l'ADN sans anticorps fixé est éliminé. L'ADN est purifié par séparation des complexes protéines/ADN. Ces fragments sont ensuite utilisés pour former la librairie à analyser. Plusieurs analyses sont possibles à partir de cette librairie. On peut réaliser une PCR pour vérifier la présence de séquences d'ADN précises. Une expérience de PCR quantitative (qPCR) permet d'analyser la composition en nucléotides des fragments de la librairie. Au final, la séquence nucléotidique des fragments peut être découverte par séquençage ou utilisation de puce à ADN (microarray).

Figure III.3: Collas, P. & Dahl, J. Chop it, ChIP it, check it: The current status of chromatin immunoprecipitation. *Front. Biosci.* **13**, 929–943 (2008).

Une variation du ChIP-seq est le ChIP-exo[97]. Pour une expérience de ChIP-exo, après sélection des fragments par immunoprécipitation, les complexes protéine/chromatine sont mis en présence d'une exonucléase qui dégrade l'ADN non protégée par la fixation de la protéine. Les régions obtenues sont donc plus précises (précision à la paire de base, contre +- 300 pb pour le ChIP-seq, Figure III.4).

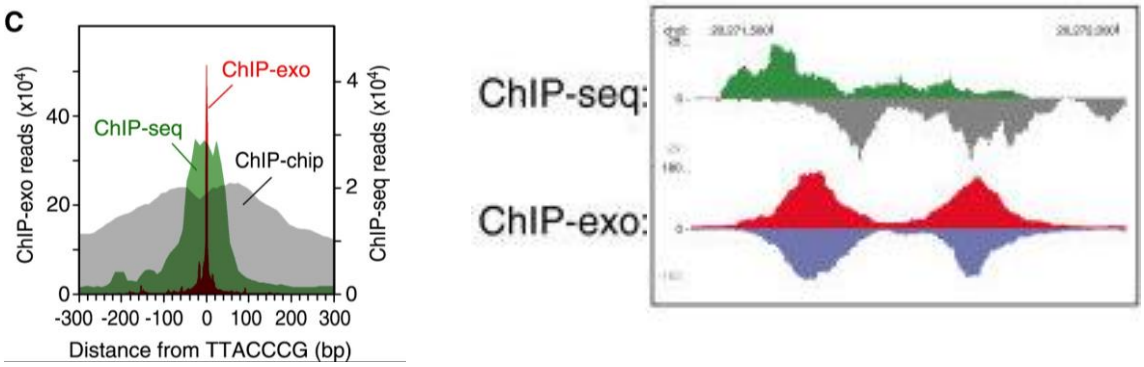


Figure III.4 ***: Comparaison de 3 signaux de ChIP-seq et ChIP-exo.***
Cette figure montre les signaux des fragments de lecture alignés sur le génome de référence. Les alignements sont issus d'expériences de ChIP séquencées en paired-end selon le même protocole. En vert et en rouge les fragments sont alignés sur le brin sens de l'ADN, en gris et en bleu les fragments sont alignés sur le brin antisens. La distance entre les pics sur le brin sens et antisens représente les extrémités séquencées des fragments de la librairie. Les fragments de librairie sont plus longs dans les expériences de ChIP-seq que dans les expériences de ChIP-exo. Les expériences de ChIP-exo permettent d'identifier plus précisément la région de fixation de la protéine d'intérêt (cadre 1). Dans le cas d'événements de fixation successifs de la protéine, une taille de fragments de librairie plus faible permet de différencier les régions de fixation des protéines.

Il est important de noter qu'une expérience de ChIP est limitée par la qualité de l'anticorps utilisé. Certains anticorps peuvent ne pas être assez spécifiques et capturer des protéines similaires ou d'autres régions du génome[98].

Il est donc souvent nécessaire de réaliser une étape de contrôle afin de limiter le biais dans la sélection des fragments. Le but est de repérer toutes les régions fixant des anticorps de façon non sélective. Pour cela, une expérience de ChIP est réalisée dans les mêmes conditions à la différence que l'anticorps est remplacé par un anticorps d'une autre espèce (souris, lapin) ou par un anticorps non spécifique comme celui de la globuline (IgG). Toutes les régions sélectionnées par cet anticorps sont repérées et marquées comme faux positif. Ce contrôle est complémentaire au contrôle du séquençage (WCE). Dans l'idéal, les deux opérations seront réalisées.

Appliquer la sélection par anticorps implique qu'il n'est pas possible d'effectuer une expérience de ChIP-seq sur des protéines ne possédant pas d'anticorps disponible, ce qui limite le nombre de TFs cibles possibles.

Figure III.4: 1.Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408–19 (2011).

Une des caractéristiques du ChIP-seq est qu'il peut aussi détecter les fixations indirectes de protéines. En effet, si une protéine, comme un cofacteur, est complexée avec une autre protéine qui, elle, est fixée sur de la chromatine, cet événement est détecté au même titre que des événements de fixation directe. Il est possible de limiter ce biais en réalisant plusieurs fois l'expérience. Une région présente sur tous les réplicas est plus probablement un vrai positif.

Une autre caractéristique est que la sensibilité de cette technologie dépend de la profondeur du cycle de séquençage (c'est-à-dire du nombre de marqueurs de séquence cartographiés), de la distribution du facteur cible[99] et de la taille ainsi que de la complexité du génome de référence[100].

DAP-seq

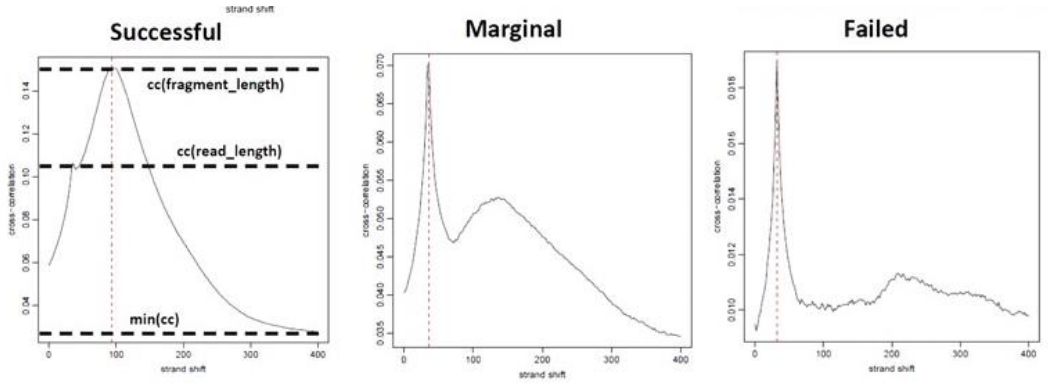
Une autre méthode d'identification des régions de fixation de TF est la purification par affinité d'ADN suivie d'un séquençage[101] (DAP-seq). La particularité du DAP-seq est que la librairie de fragments est issue d'ADN germinatif donc ne nécessite pas de PCR. Les TFs d'intérêt sont produits *in vitro* et fusionnés avec une étiquette (Tag) protéique. Les fragments d'ADN sont ensuite hybridés avec les TFs marqués. Les complexes protéine/ADN taggés sont ensuite prélevés. Cette méthode permet de dépister un plus grand nombre de TFs que les méthodes reposant sur l'immunoprécipitation. Aucun anticorps spécifique n'étant nécessaire, il est possible d'étudier des TFs sans anticorps connu. Cependant, cette méthode connaît un taux d'échec élevé, en raison de la complexité du protocole expérimental et d'une faible expression des protéines marquées. Le DAP-seq présente des inconvénients. Étant réalisé sans contexte chromatidique, la détection d'une fixation par cette méthode ne s'observe pas forcément *in vivo*.

b. Analyse bioinformatique des méthodes de fixation des protéines à l'ADN

Les analyses des données ChIP-seq et de DAP-seq permettent de générer des fichiers SAM contenant les fragments de lecture alignés sur un génome de référence. Ces fragments de lecture sont enrichis au niveau des régions de fixation de la protéine d'intérêt. La difficulté est d'identifier ces régions enrichies par rapport à l'environnement direct de la région.

La distribution des fragments de lecture sur le génome permet de donner une première évaluation de la qualité de l'expérience de ChIP. Afin d'évaluer la qualité de l'expérience de séquençage à haut débit, des scores de cross-corrélation peuvent être calculés[102]. Les fragments de lectures sont décalés dans la direction du brin auquel elles correspondent par un nombre croissant de paires de bases. Le score de corrélation de Pearson est calculé entre les vecteurs de comptage des lectures par position et, ce pour chaque brin. A partir de ces scores de cross-corrélation, deux métriques sont calculées.

Le NSC (Normalized Strand Coefficient) est le ratio normalisé entre la cross-corrélation de la taille des fragments de librairie et la cross-corrélation du bruit de fond. Le RSC (Relative Strand Correlation) est le ratio entre la taille des pics des fragments de librairie et la taille des fragments de lecture. Les données ChIP de haute qualité mettent souvent en évidence un pic de longueur des fragments de librairie plus grand que le pic de longueur des fragments de lecture (Figure III.5).



$$NSC = \frac{cc(fragment\ length)}{min(cc)} \qquad RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

Figure III.5 ***: Représentation graphique du NSC et RSC.***
Deux pics de cross-corrélation sont habituellement observés dans une expérience ChIP. L'un correspond à la longueur des fragments de lecture (pic "fantôme") et l'autre à la longueur moyenne des fragments de la bibliothèque. La hauteur absolue et la hauteur relative des deux pics sont des déterminants utiles pour le succès d'une expérience ChIP-seq. Un IP de haute qualité est caractérisé par un pic ChIP beaucoup plus élevé que le pic "fantôme", alors qu'un tel pic est souvent très faible ou nul dans les expériences ratées.

Une des caractéristiques des alignements d'expériences de ChIP est la distribution bi-modale des fragments de lecture autour des régions enrichies. Cette distribution vient du fait que seules les extrémités des fragments de la librairie de séquençage sont séquencées. La région de fixation de la protéine se retrouve donc entre ces deux sommets. L'identification de ces régions aux profils bien particuliers s'appelle la recherche de pics de fixation (peak-calling). Les différentes méthodes de peak-calling sont décrites dans la figure III.6.

Au cours de ma thèse j'ai utilisé le peak-caller MACS2[103] et testé JAMM[104]. Tous les deux fournissent, en sortie, un fichier texte BED qui contient les sites de fixation de la protéine le long du génome. Les fichiers BED peuvent facilement être visualisés grâce à des "genome browser" tels que IGV[105], Ensembl[106] ou UCSC[107].

Une métrique répandue pour évaluer une bonne recherche de pics est le FRiP (Fraction of Reads in Peaks). Cette méthode permet d'évaluer la précision de l'immunoprécipitation. Pour un anticorps sensible on s'attend à ce que la plupart des fragments de lecture se trouvent dans des régions enrichies identifiées par le peak-calling[102].

Figure III.5 : 1. Landt, S. G. et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 22, 1813–31 (2012).

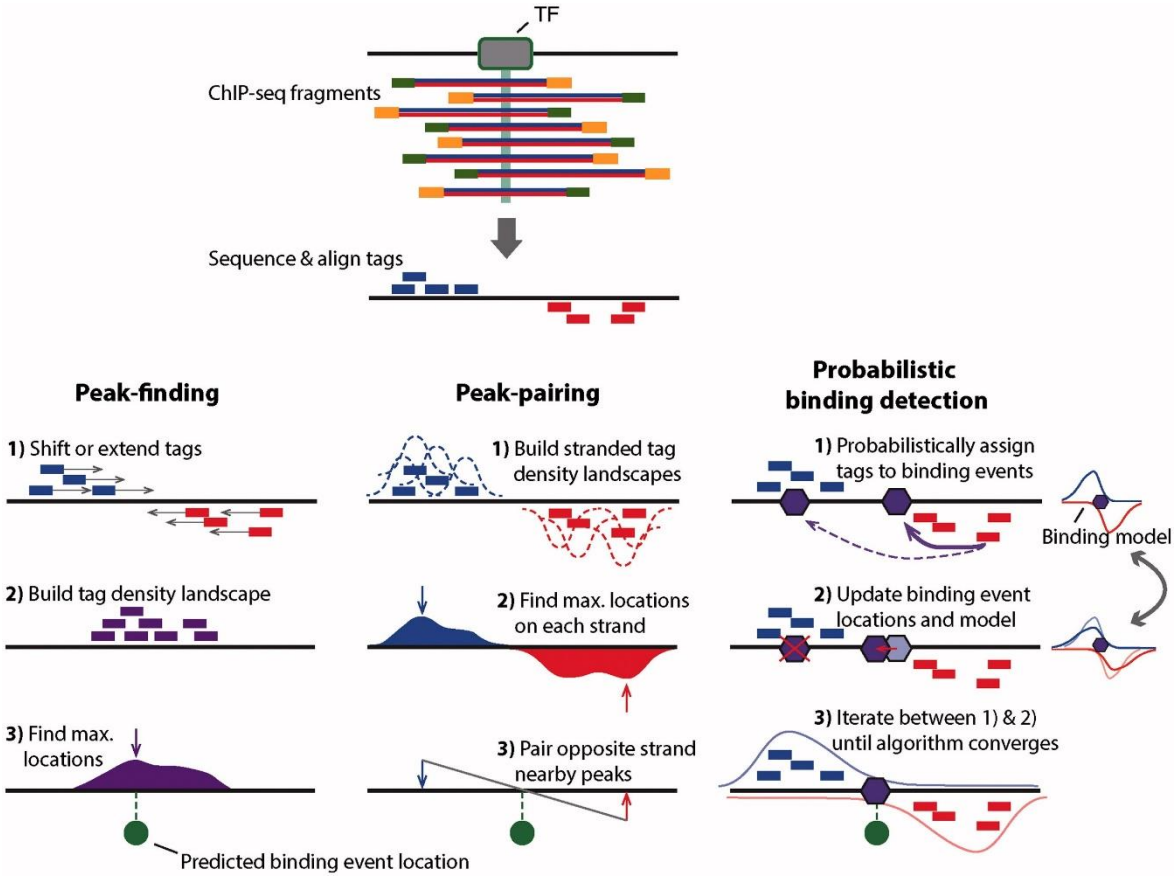


Figure III.6 ***: Méthodes de détection des événements de liaison de protéines à l'ADN à partir de données de séquençage à haut débit.***
Les méthodes de prédiction se basent sur l'observation suivante. L'événement de liaison se situe aux positions où la densité des fragments de librairie est la plus élevée. Leur objectif est donc d'identifier les centres des fragments de librairie à partir de leurs extrémités (fragments de lectures). Les méthodes sont appliquées pour chaque région contenant un enrichissement en fragments de lecture. La méthode de ***recherche de pics*** *(peak finding). 1) Les fragments de lecture correspondant aux extrémités des fragments de la librairie sont décalés vers le centre théorique du fragment de librairie. 2) Les fragments de lecture sur le brin sens et antisens sont fusionnés afin de construire un profil de densité. 3) La position où le profil est le plus élevé prédit où s'est produit l'événement de fixation. La méthode d'****appariement des pics*** *(peak-pairing). 1) Un profil de densité est réalisé à partir des fragments de lecture correspondant aux extrémités des fragments de la librairie. 2) La hauteur maximale du profil est définie pour chaque extrémité.3) L'événement est prédit au point médian entre les deux maxima. La* ***méthode probabiliste*** *(Probabilistic binding detection). 1) Un modèle représentant l'événement de fixation et la répartition des fragments de lecture l'entourant est réalisé. 2) Le modèle est ajusté au cours d'étapes successives. A chaque étape, les fragments de lectures sont associés de manière probabiliste à des événements de liaison proches. Cette association est réalisée en fonction de la distance entre les fragments et le lieu de l'événement de liaison. Le lieu de fixation et la distribution sont alors modifiés afin de se rapprocher au mieux de la réalité observée. 3) A partir du moment où le modèle reste inchangé entre deux itérations, l'événement de fixation final est prédit.*

Figure III.6 : Mahony, S. & Pugh, B. F. Protein–DNA binding in high-resolution. *Crit. Rev. Biochem. Mol. Biol.* **50**, 269–283 (2015).

c. Méthode et analyse de l'état de la chromatine

Un des points communs à tous les éléments régulateurs chez l'homme est la fixation directe de TFs. Il est donc possible d'identifier et de caractériser les régions régulatrices dans le génome grâce à la présence de TFs. Plusieurs méthodes expérimentales ont été développées afin d'étudier la fixation de TFs. Ces méthodes peuvent se classer selon l'information qu'elles tentent de capturer.

Certaines méthodes ont pour but d'identifier toutes les régions du génome où la chromatine est accessible. Comme il est décrit dans le chapitre sur les protéines régulatrices, les TFs ne peuvent se fixer que sur les brins ouverts de la chromatine. L'identification des régions accessibles du génome permet donc de repérer les régions régulatrices actives dans un contexte biologique précis. Les principales méthodes de séquençage à haut débit à l'échelle du génome pour identifier la chromatine ouverte sont le DNase-seq, MNase-seq, FAIRE-seq et ATAC-seq (Figure III.7). Elles présentent l'intérêt de ne pas nécessiter d'a priori (régions d'intérêt, fixation d'un TF particulier, etc.), mais elles ne permettent pas d'identifier le type d'éléments régulateurs présents. Chacune de ces méthodes présente des avantages et des inconvénients détaillés dans le tableau III.2.

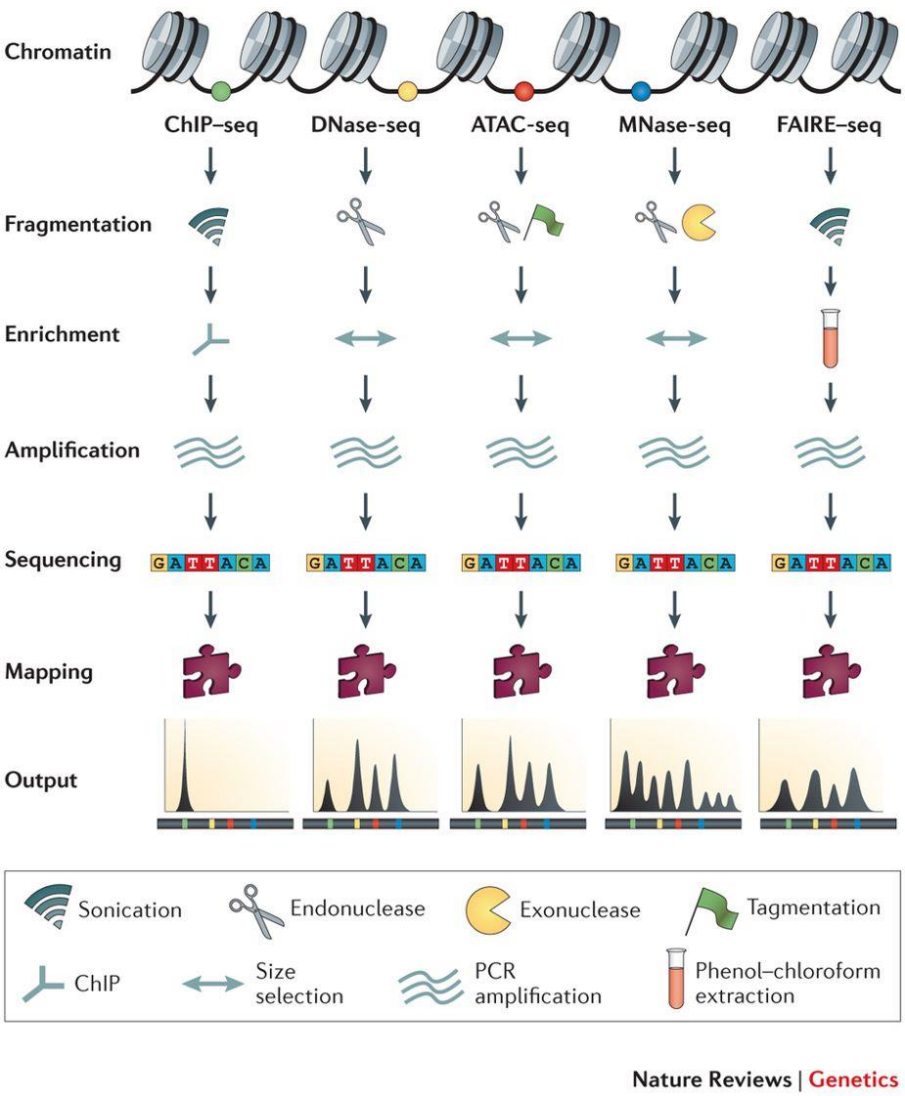


Figure III.7 ***: Comparaison des méthodes d'analyse de l'ouverture de la chromatine.*** *Le DNase-seq permet d'identifier les régions sensibles à la Désoxyribonucléase I (DNase I) qui est une endonucléase qui clive la chromatine au niveau des régions ouvertes et accessibles de celle-ci*[108–110]*. Le MNase-seq permet de différencier les régions de la chromatine occupées par des nucléosomes par digestion de la chromatine ou des nucléosomes par digestion de l'endo-exonucléase Nucléase micrococcale*[111,112]*. Le FAIRE-seq, basé sur la formation de liens covalents entre l'ADN et les protéines (cross-linking) par formaldéhyde, est plus efficace sur l'ADN nucléosomal que sur les régions nucléosome-déplétées du génome. L'ATAC-seq est une évolution du MNase-seq, qui utilise la transposase Tn5 pour cliver l'ADN*[113]*.*

Figure III.7: Meyer, C. A. & Liu, X. S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.* **15**, 709–721 (2014).

d. Limite des méthodes expérimentales

Les méthodes expérimentales permettent d'identifier la chromatine ouverte et les régions fixant les TFs, mais l'annotation s'avère plus compliquée.

Toutes ces méthodes sont expérimentales. Une des difficultés est de déterminer si les événements détectés représentent bien une réalité biologique. Par exemple, les événements de fixation d'un TF le long du génome ne signifient pas que toutes ces régions régulent l'activité d'un gène. C'est ce que montrent certaines études[114]. Cela peut être dû aux biais dans les méthodes expérimentales. Ces biais sont à prendre en compte dans l'analyse, que ce soit pour l'ouverture de la chromatine (tableau III.2) ou pour la fixation des facteurs de transcription. Une autre possibilité est que la présence d'un site de fixation du TF n'ait pas de fonction. Une des solutions pour pallier ces difficultés dans l'annotation des éléments régulateurs est de réaliser différentes expériences visant à confirmer l'activité des éléments régulateurs. C'est ce à quoi s'attèlent des consortia tels que le projet ENCODE (ENCyclopedia of DNA Elements) au NBCI ou FANTOM au Japon et à l'international.

La deuxième solution est de prédire l'activité d'une région en fonction de sa séquence nucléotidique. Par exemple, les promoteurs sont composés d'éléments ayant des motifs et des fonctions connus (TATA-box, boîte CAAT, etc) qui permettent de les identifier. De plus, tous les éléments régulateurs fixent des TFs. Ces TFs possèdent des motifs de fixation qui peuvent être connus. Il est donc possible d'identifier les régions régulatrices à partir de la composition en nucléotides et leurs séquences à l'aide d'outils statistiques et bio-informatiques tels que la suite d'outils MEME[115].

Tableau III.2 ***: Avantages et inconvénients techniques des différentes méthodes d'identification de l'ouverture de la chromatine.***

	Avantage(s)	**Inconvénient(s)**
DNase-seq	Le signal est plus important au niveau des promoteurs[116]	Nécessite un étalonnage minutieux des conditions de digestion et de la sélection des fragments[117]. Nécessite plus de matériel génétique
FAIRE-seq	Le signal est plus important au niveau des non-promoteurs[116] Peut analyser tous les types de cellules	Affiche un bruit de fond plus élevé, ce qui en fait une méthode moins sensible[118]
ATAC-seq	A besoin de moins de matériel génétique[119] Est rapide (~3h)	

B. Méthodes bioinformatiques et statistiques

La bioinformatique appliquée à la génomique connaît un essor sans précédent grâce à l'explosion du nombre de données génomiques favorisée par la facilité du séquençage, l'augmentation des capacités de calcul et par la baisse du prix du matériel informatique. De plus en plus de structures de recherche s'équipent et mutualisent leurs ressources (ex : Mesocentre Marseille[a]), et leur plateformes (Institut Français de Bioinformatique[b]).

Cette évolution rend possible des analyses d'intégration et de comparaison de données à grande échelle jusqu'alors impossibles. Cette progression du développement de la bioinformatique a favorisé la création de nouveaux outils permettant la découverte et l'annotation d'éléments régulateurs nouveaux (*Ab Initio*).

Une des difficultés dans l'identification d'éléments régulateurs est l'apparente hétérogénéité de leur structure fonctionnelle. Il faut donc se concentrer sur la similarité entre les séquences nucléotidiques, ainsi que sur la similarité avec des éléments régulateurs déjà annotés.

1. Identification d'éléments régulateurs de la transcription

a. Le cas des promoteurs

Seuls les promoteurs possèdent des éléments fonctionnels récurrents connus dans leur région proximale (TATA box, BRE, voir partie promoteur). Il est théoriquement possible de les identifier grâce à la présence de ces éléments[120]. En pratique, analyser les séquences du génome permet d'identifier leurs motifs. Mais la présence de ces éléments fonctionnels n'est pas systématique ce qui rend la découverte sans a priori difficile.

Il est plus efficace de comparer les séquences du génome à une liste de promoteurs déjà annotés (données d'entraînement) pour rechercher des motifs fonctionnels non définis précédemment. Cette méthode peut être utilisée, soit seule[121], soit en combinaison avec la modélisation des caractéristiques du promoteur, telles que la présence d'îlots CpG et d'un premier exon potentiel.

a : https://mesocentre.univ-amu.fr
b : https://www.france-bioinformatique.fr

Ces méthodes ont des limites.

- La spécificité et la sensibilité dépendent largement de la qualité des données utilisées pour l'entraînement.
- Ces méthodes se limitent à la découverte des promoteurs similaires à des promoteurs déjà annotés.
- Ces méthodes sont efficaces pour découvrir des promoteurs liés au gène de ménage car ils sont proche d'îlots CpG et de structures d'éléments fonctionnels plus "typiques". Or ces types de promoteur ne couvrent que la moitié des gènes annotés chez l'homme[122]

Ces méthodes doivent donc être utilisées en parallèle avec d'autres procédures, afin d'identifier tous les types d'éléments régulateurs.

b. Approche sans a priori

Tous les éléments régulateurs de la transcription ont pour point commun la fixation de facteurs de transcription. En conséquence, il est possible d'identifier de nouveaux éléments de la régulation en scannant le génome à la recherche des TFBS.

Les bases de données TRANSFAC[123] et JASPAR[124] annotent et identifient les sites de fixation détectés expérimentalement (ChIP-seq, SELEX) des TFs les mieux caractérisés. Des suites d'outils comme RSAT[5] proposent des outils tels que matrix-scan qui permettent de parcourir des séquences d'intérêt afin de détecter des TFBS à partir de plusieurs sources/bases de données[125].

Cette méthode renvoie souvent un grand nombre de TFBS putatifs car :

- La spécificité et la sensibilité dépendent grandement de la qualité des données utilisées pour construire les données de TFBS.
- Les motifs des TFBS peuvent apparaître de manière fortuite dans le génome.
- Tous les TFBS sont recherchés, qu'ils se trouvent dans des régions ouvertes de la chromatine ou non.

Ce type d'analyse est aussi limité par le nombre de TFBS identifiés dans les bases de données correspondantes. Ce type d'études peut être réalisé en collaboration avec d'autres chercheurs afin d'affiner la découverte de nouveaux éléments régulateurs.

c. Approche avec a priori

Les données de fixations des protéines ne sont pas les seules données utilisées pour identifier les éléments régulateurs de la transcription. Afin de limiter la recherche d'éléments régulateurs découverts à des régions d'intérêt, il peut être intéressant de se concentrer sur les gènes co-régulés (microarray, RNA-seq etc).

En étudiant les motifs dans les régions adjacentes à tous ces gènes co-régulés il est possible de découvrir de nouveaux TFBS ou éléments fonctionnels non annotés.

Différents programmes permettent de découvrir des motifs dans les régions co-régulées. MEME, Geneious[126] et RSAT sont parmi les plus utilisés. Ces méthodes permettent, d'identifier des clusters de TFBS, donc des TFs pouvant coopérer entre eux et, ainsi, d'établir un réseau d'interactions protéine/protéine.

d. Approche de conservation des séquences

La génomique comparative s'intéresse à la conservation des séquences d'ADN entre les différentes espèces. Elle a pour but d'identifier les régions génomiques sélectionnées au cours de l'évolution par pression de sélection[127]. La méthode d'empreinte phylogénétique consiste à comparer des séquences d'éléments de différentes espèces entre elles afin d'identifier les régions conservées[128]. Cette méthode est particulièrement efficace pour prédire la conservation de l'ADN non-codant.

Cette méthode présente des inconvénients :

- Tous les éléments régulateurs ne semblent pas soumis à cette pression de sélection. En effet, certaines régions se dégradent au fil de l'évolution et ne sont donc pas détectées[129].
- Elle ne permet pas, non plus, de détecter des éléments régulateurs "récents".
- De plus, la conservation de la séquence n'assure pas la fonction de l'élément détecté, et inversement.

A retenir pour mes travaux

Les éléments cis-régulateurs (CRE) peuvent être identifiés et catégorisés par différentes méthodes. Les méthodes expérimentales couplées au séquençage à haut débit permettent d'identifier des CREs le long du génome. Le séquençage à haut débit présente un biais, certaines séquences d'ADN sont sur ou sous représentées après alignement des fragments de lecture. C'est pour cela qu'il est nécessaire d'effectuer des contrôles qualités avec des métriques tel que le NSC (Normalized Strand Cross-correlation coefficient) et RSC (Relative Strand Cross-correlation coefficient).

Les CRE peuvent être identifier, entre autres, par l'accessibilité de la chromatine (ATAC-seq, DNAse-seq, etc.) ou la présence de protéines régulatrices de la transcription (ChIP-seq, ChIP-exo, DAP-seq). Les méthodes ChIP identifient les interactions protéine/ADN par immunoprécipitation. La qualité de cette capture dépend grandement de la spécificité de l'anticorps utilisé. Cette spécificité peut être mesurée par le score du FRiP (Fraction of Reads in Peak). Les expériences de captures de interactions protéine/ADN peuvent aussi identifier les interactions indirectes (cofacteur). Une des difficultés est de différentier les interactions directes des interactions indirectes à partir de données ChIP.

Les méthodes bioinformatiques permettent d'identifier les CREs à partir de la séquence d'ADN. Les sites de fixation de facteurs de transcription (TFBS) sont clustersés au niveau des CREs. La recherche de motifs ou de séquences consensus permet d'identifier les TFBS putatifs.

IV Consortia et données publiques

A. Consortia d'annotation d'éléments cis-régulateurs

Un des grands défis de la génomique est l'identification et la caractérisation des éléments fonctionnels. Chez l'Homme, le projet génome humain a permis la création du génome de référence permettant des études sur la séquence génomique.

1. Consortium ENCODE

Le consortium Encyclopedia of DNA Elements (ENCODE) a pour but de créer un catalogue public d'éléments fonctionnels chez l'Homme. Afin d'identifier et caractériser les éléments fonctionnels, le consortium produit de nombreuses données issues d'expériences biologiques ayant passé un contrôle qualité conforme à un protocole normalisé. ENCODE a été créé en Septembre 2003 par le National Human Genome Research Institute Home (NHGRI) aux Etats Unis. Le consortium est constitué de plusieurs groupes assurant différentes responsabilités, allant de la production des données à leur analyse (carte figure IV.1).

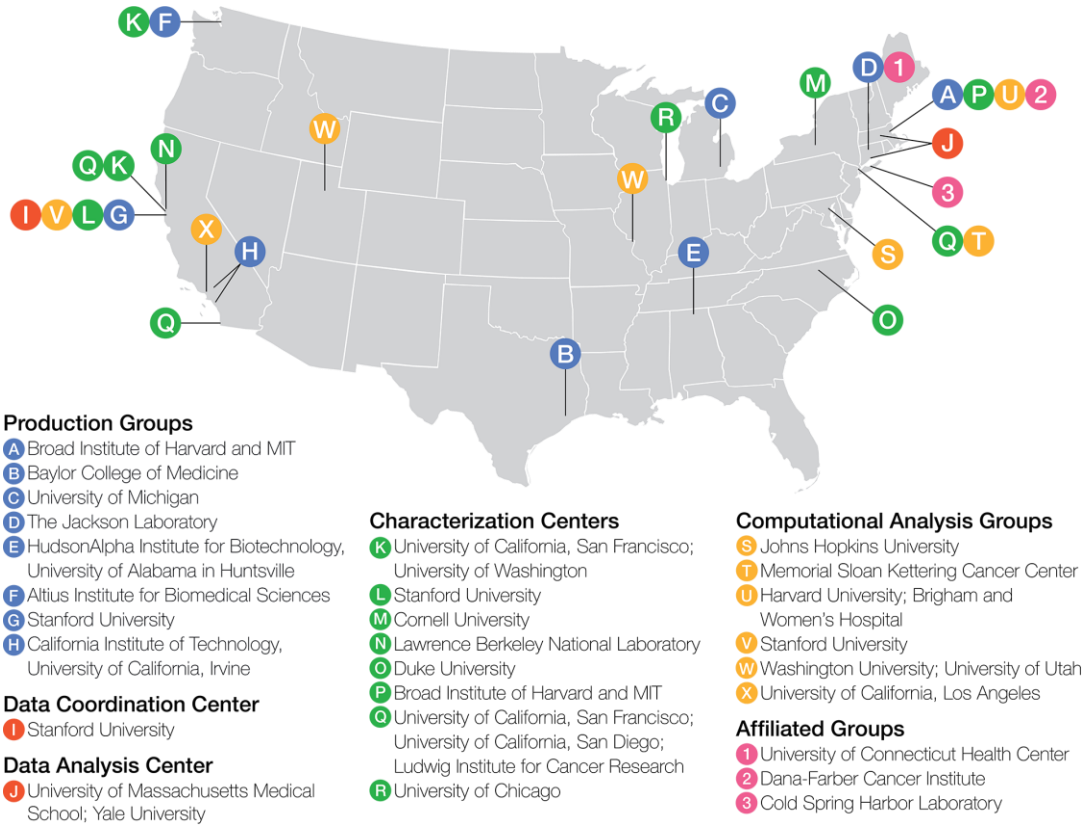


Figure IV.1 ***: Carte du consortium ENCODE pour la phase 4.***

Figure IV.1: https://www.encodeproject.org/help/project-overview/

Le projet ENCODE étudie des données issues de plusieurs lignées cellulaires cancéreuses et cellules primaires saines. Le projet génère des données afin de pouvoir : identifier les régions de fixation des TFs, la position des marques d'histone, la transcription au niveau des gènes et de l'ADN non-codant, l'ouverture de la chromatine, la fixation des protéines à l'ARN, la méthylation de l'ADN et la structure 3D de la chromatine (interaction chromatine et TAD).

Le projet ENCODE comporte déroulé en quatre phases.

Phase I : le projet pilote. L'intention de la phase pilote était de tester les différentes procédures et analyses afin d'identifier les méthodes les plus adaptées à des analyses à grande échelle. Leur prix, leur précision et leur robustesse face à l'utilisation à large échelle sont pris en compte afin de déterminer les analyses à employer pour les phases suivantes. Le but était aussi de développer les outils nécessaires pour la réalisation d'analyses à grande échelle. Dans cette optique, le consortium s'est concentré sur 1% du génome humain soit 30 Mb[130]. Sur les 44 régions retenues 50% ont été sélectionnées car elles comportaient des gènes et CREs connus et étudiés, 50% ont été sélectionnées au hasard. La phase pilote s'est achevée en 2007 et a permis le développement de nouvelles méthodes permettant l'analyse et l'identification des éléments régulateurs.

Phase II, III et IV : la phase de production du projet. Suite aux conclusions de la phase pilote, la phase de développement a été lancée. L'objectif de cette phase est double. Dans un premier temps, le but est d'analyser le génome humain dans sa totalité et de mener des étude pilotes supplémentaires.

Les données produites, brutes et analysées, sont mises à la disposition de la communauté scientifique sur leur site internet et visualisable sur les navigateurs de génome en ligne tel que UCSC genome browser. Au final, le consortium ENCODE a permis le développement de nombreuses méthodes et outils afin d'identifier et caractériser les CREs[131,132] et d'évaluer les méthodes de séquençage à haut débit[133–135]. Il a également fourni une liste de 1 310 152 CREs putatifs chez l'Homme pour environ 600 types cellulaires différents[136].

Le projet ENCODE met aussi à disposition des données et analyses sur les CREs de la souris. Mod ENCODE (MODel organism ENCyclopedia Of DNA Elements) est un projet issu des études pilotes menées lors de la phase III d'ENCODE. L'initiative a pour objectif d'identifier les CREs chez les organismes modèles Caenorhabditis elegans[137] et Drosophila melanogaster[138].

a. Roadmap

En 2008 le NIH (National Institute of Health) a créé ce consortium afin de produire des ressources publiques de données épigénétiques. L'objectif est de rapprocher les recherches fondamentales menées par ENCODE et l'étude de maladies. Ce consortium a permis de cartographier l'épigénome de 111 lignées cellulaires et tissus chez l'Homme[139].

2. Consortium FANTOM

Le projet FANTOM (Functional ANnoTation Of the Mammalian genome) est issu d'un consortium international créé au Japon. Comme ENCODE, il a pour but d'annoter les éléments fonctionnels dans le génome des mammifères. Le consortium se concentre principalement sur le modèle humain et murin. Le projet FANTOM est actuellement divisé en 6 phases successives.

FANTOM1 a permis de développer une méthode d'annotation des gènes fonctionnels, utilisée par le projet human genome[140]. La seconde phase de FANTOM fournit la séquence et l'annotation fonctionnelle de plus de 60 000 ADN complémentaires (cDNA) chez la souris[141]. Au cours de la troisième phase, le consortium FANTOM a développé la méthode CAGE (Cap Analysis Gene Expression) permettant l'étude de l'initiation de la transcription par l'analyse de l'extrémité 5' de l'ARN coiffé décrit dans la figure IV.2. L'utilisation de cette nouvelle méthode a permis d'identifier la transcription au niveau de 63% du génome humain, ainsi que la découverte de nombreux ARN non codants[142]. Le développement et l'amélioration des nouvelles techniques de séquençage ont rendu possible un séquençage plus profond des données issues d'expérience CAGE. Cette avancée a été employée par FANTOM4 afin d'étudier la cinétique entre l'expression au niveau du TSS et de la fixation de TFs[143]. La cinquième phase FANTOM est divisée en deux projets : l'identification des promoteurs et des promoteurs alternatifs, et l'identification des enhancers transcrits chez l'Homme et la souris[144] par la méthode CAGE. FANTOM5 s'est aussi intéressé à la cinétique de la transcription des enhancers par l'étude de 19 évolutions temporelles chez l'Homme et 25 chez la souris.

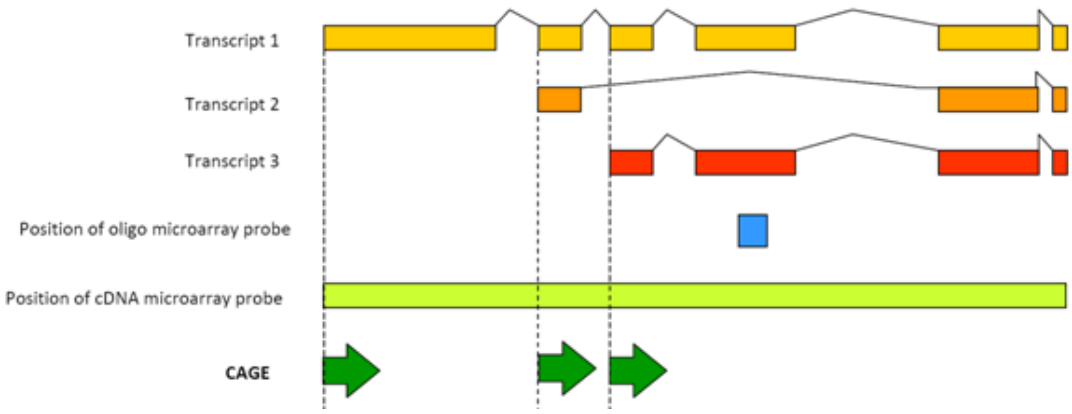


Figure IV.2 ***: Aperçu de la méthode CAGE pour l'identification de promoteurs alternatifs.***

Les analyses issues des différentes phases de FANTOM sont disponibles sur leur site internet. De plus, le consortium met à disposition les outils développés au cours de ces analyses[145]. La sixième phase de FANTOM est en cours. Elle s'intéresse à la fonction des longs fragments d'ARN non codant (lncRNA).

Figure IV.2: http://fantom.gsc.riken.jp/protocols/basic.html

B. Archivage des données de publication

Lors de la publication d'articles scientifiques, les données brutes et les résultats produits doivent être disponibles et accessibles au même titre que le manuscrit. Ces données doivent être stockées dans des bases de données publiques. Gene Expression Omnibus (GEO) est actuellement l'entrepôt de données le plus utilisé pour les données issues de microarray et pour le séquençage à haut débit[146]. Cette base de données est financée par le NCBI (National Center for Biotechnology Information) aux Etats Unis. Les données sont rassemblées par papiers et peuvent contenir de nombreuses expériences biologiques différentes. Les métadonnées sont accessibles et consultables via le portail de GEO.

Une Alternative à GEO est la base de données ArrayExpress[147]. AE express est géré par l'EMBL-EBI (European Molecular Biology Laboratory) en Europe. Contrairement à GEO, AE propose une supervision dans l'annotation des métadonnées des expériences soumises. AE importe les données issues de GEO. Les données brutes des expériences accessibles par AE sont stockées sur European Nucleotide Archive (ENA).

A retenir pour mes travaux

La réalisation de l'annotation d'éléments fonctionnels dans le génome nécessitant un important coût financier et humain, certains laboratoires se regroupent en consortium publique afin de mutualiser leurs moyens. Ces consortia réalisent des centaines d'expériences biologiques dont les données brutes et analysées sont disponibles via leurs bases de données en ligne. Ces données sont annotées uniformément pour faciliter le partage. Parmi ces consortia, ENCODE produit, entre autres, des données de ChIP-seq pour plusieurs lignées et tissus cellulaires humains, ainsi qu'une liste d'éléments cis-régulateur hypothétiques. Array Express est l'équivalent Européen d'ENCODE. Le consortium FANTOM met à disposition des données issues d'expériences CAGE permettant d'identifier les enhancers transcrits.

A cette importante quantité de données s'ajoute celles produites par la communauté scientifique stockées dans des entrepôts de données tel que GEO et AE/ENA. L'inconvénient majeur de la base de données GEO est que le dépôt des données et métadonnées biologiques n'est pas soumis à des consignes d'annotation uniformes (nom des protéines, cellules, etc.). De plus, les données de séquençage brutes ne sont pas disponibles au format le plus couramment utilisé (FASTQ). Pour ces raisons je télécharge mes données à partir d'AE et ENA.

V Reproductibilité, interopérabilité, accessibilité des données

La facilité à produire des données biologiques a conduit à une explosion du nombre de bases de données. Les nouvelles technologies ont considérablement facilité la mise en place de base de données biologiques.

Ces avancées peuvent être d'ordre expérimental comme le séquençage à haut débit, la méthode whole genome, etc.… ou d'ordre informatique, comme l'utilisation de logiciels tels que les frameworks pour l'aide à la mise en place de bases de données et de sites internet (Django, Laravel, etc) ou comme le traitement de données à haut débit (Nextflow, Snakemake, etc.). L'émergence du calcul haute performance ou HPC avec l'ouverture de grands centres de calculs, l'augmentation de la puissance de calcul, la parallélisation sur GPU, etc. a, elle aussi, contribué à l'enrichissement des connaissances.

Ces bases de données sont composées de deux éléments essentiels : les données elles mêmes et les métadonnées. Les données peuvent être un ensemble de faits, de mots, d'observations, de mesures ou la description d'un objet/événement/etc. Pour la biologie, ces informations peuvent être un nom de gène, sa position génomique, la protéine qu'il code, etc. Les données peuvent être exploitées pour en retirer des informations complètes. Elles peuvent être renvoyées ou analysées pour prendre certaines décisions. Les métadonnées sont décrites comme des données sur les données. Cela signifie que les métadonnées contiennent la description informative et pertinente des données originales. Elles aident l'utilisateur à connaître la nature des données et à déterminer leur utilité[a].

Dans l'exemple d'une expérience de ChIP-seq de la base de données ENCODE les données seraient les fichiers de données de séquençage brut, ainsi que tous les fichiers qui en découlent (fichier d'alignement, les pics de fixation de la protéine, etc.). Les métadonnées représentent toutes les informations afférentes, qu'elles soient liées directement à l'expérience (nom de l'expérimentateur, anticorps utilisé, etc) ou au contexte des données (le lien de téléchargement, date de mise en ligne ou mise à jour, les fichiers liés à l'expérience, etc.).

a : https://techdifferences.com/difference-between-data-and-metadata.html

Nucleic Acids Research recense plus de 1600 bases de données biologiques[a] recouvrant un grand nombre de sujets comme les séquences nucléotidiques, d'ARN ou de protéique, les voies métaboliques et cela pour un ou plusieurs organismes. Devant cette abondance de données il devient difficile de retrouver des informations pertinentes et correctement annotées. En effet, certaines de ces bases de données sont mal indexées ou manquent de métadatas ce qui rend leur utilisation complexe voire impossible. De plus, ces bases de données peuvent être redondantes, dans la mesure où elles exploitent les mêmes données brutes analysées de plusieurs façons différentes. Un autre point important réside dans l'accessibilité des données. Certaines sont bloquées par des barrières monétaires. Dans d'autres cas, le protocole de collecte et l'analyse des données ne sont pas expliqués. L'accessibilité a été cruciale dans mon travail de thèse car il repose sur l'analyse à haut débit de données issues de bases de données biologiques.

Face à l'explosion du Big Data, l'initiative FAIR (Findable, Accessible, Interoperable, Reusable) vise à normaliser l'ensemble des données et métadonnées afin de faciliter le partage de données et leur utilisation.

Tableau V.1 ***: Distinction entre données et métadonnées.***

DONNÉES	**MÉTADONNÉES**
Les données de base sont un ensemble de faits et de statistiques qui peuvent être exploités, référés ou analysés.	Les métadonnées décrivent les informations pertinentes sur les données.
Information Les données peuvent ou ne peuvent pas être informatives.	Les métadonnées sont toujours informatives.
Le traitement des données peut ou non avoir été effectué.	Les métadonnées sont toujours des données traitées.

a : www.oxfordjournals.org/nar/database/c

A. FAIR consortium

C'est en 2016 qu'est publié le premier article sur les principes FAIR (The FAIR Guiding Principles for scientific data management and stewardship). L'objectif est d'instituer des règles à suivre pour faciliter l'accès aux données, les rendre plus facilement utilisables, mais aussi plus faciles à partager. Ces principes reposent sur 4 critères : Facilité à trouver l'information (Findable), Accessibilité (Accessible), Interexploitabilité (Interoperable), Réutilisabilité (Reusable).

Pour partager les données et les rendre facilement utilisables. Encore faut-il les trouver. Les données et les métadonnées doivent impérativement être lisibles à la fois par l'être humain et la machine. Une donnée est consultable si elle répond aux critères suivants (www.go-fair.org/fair-principles) :

1. Les données et métadonnées se voient attribuer un identificateur unique et persistant universel.
2. Les données sont décrites par de multiples métadonnées.
3. Les métadonnées incluent clairement et explicitement l'identificateur des données qu'elles décrivent.
4. Les données et métadonnées sont enregistrées ou indexées dans une ressource consultable

Prenons l'exemple des gènes de la base de données ENSEMBL[148] qui donne des informations sur le génome de plusieurs espèces. 1. Chaque gène possède un identifiant unique (ex : ENSG00000102974 pour le gène codant CTCF). 2. Plusieurs métadonnées sont disponibles comme, l'organisme d'où provient le gène, le nombre de transcrits, l'identifiant de ce gène sur une autre base de données. 3. Les métadonnées sont disponibles et téléchargeables sous un format texte qui comprend l'identifiant ENSEMBL. 4. Les métadonnées sont indexées et formatées afin de pouvoir être lues par un outil d'indexation. Le World Wide Web Consortium (W3C) préconise l'utilisation du format (*Resource Description Framework*, RDF) afin d'indexer les métadonnées[a]. En pratique, le RDF peut être écrit dans différentes syntaxes de dictionnaire comme JSON et XML dans un fichier à part ou directement dans la page web d'un site internet ou d'une base de données. Il propose une formalisation des clés permettant de signaler des informations précises. Par exemple, le moteur de recherche Google utilise un outil capable de lire le format RDF qui permet d'afficher des informations (nom du domaine, description courte, web service, etc) directement issues de sites fournissant des fichiers RDF comme Uniprot[149]. Une des limites de ce format est que les mots clés sont définis par le W3C, les termes proposés sont donc liés à l'informatique générale. Pour pallier ce manque, la communauté scientifique du vivant s'est réunie, afin de proposer un schéma proposant des mots clés liés au domaine de la biologie[b].

a : www.w3.org/TR/rdf-schema
b : https://bioschemas.org

1. Accessible

Après avoir identifié (*Find*) les données, il faut pouvoir y accéder (*Access*) afin de pouvoir les manipuler. Il est donc nécessaire d'établir, de façon simple et universelle, un protocole d'accès à ces données (éventuellement l'authentification et l'autorisation).

1. Les données et métadonnées sont téléchargeables et accessibles par leur identifiant à l'aide d'un protocole de communication standardisé.
    a. Le protocole est "ouvert", libre et universellement applicable.
    b. Le protocole prévoit une procédure d'authentification et d'autorisation, le cas échéant.
2. Les métadonnées sont accessibles, même lorsque les données ne sont plus disponibles.

Reprenons l'exemple de la base de données ENSEMBL. 1. Plusieurs données sont disponibles au téléchargement, dont la séquence nucléotidique sous forme de fichier fasta par un simple lien de téléchargement FTP. Dans ce cas, l'accessibilité est "ouverte", c'est à dire libre d'accès. Les informations de la base de données sont aussi accessibles directement par des requêtes SQL ou REST. 2. La base de données ENSEMBL fournit aussi un historique des versions des gènes. Si, par exemple, un gène est annoté différemment, les métadonnées de son ancienne version restent disponibles.

2. Interoperable

Dans la plupart des recherches, les données doivent être intégrées et/ou comparées à d'autres données provenant de sources différentes. Cela implique que les données sont interopérables (utilisables) avec toutes sortes d'applications ou workflows pour l'analyse, le stockage et le traitement.

1. Les données et métadonnées utilisent un langage formel, accessible, partagé et largement applicable pour la représentation des connaissances.
2. Les données et métadonnées utilisent un vocabulaire qui suit les principes de FAIR.
3. Les données et métadonnées comprennent des références qualitatives à d'autres données.

En reprenant une nouvelle fois l'exemple d'ENSEMBL pour les gènes. 1. Les données et métadonnées concernant un gène sont téléchargeables grâce à une interface REST. 2. Malheureusement la base de données ENSEMBL utilise un vocabulaire qui ne suit pas toujours les principes FAIR. 3. Pour y remédier une liste d'identifiants externes est proposée provenant d'UniProtKB/Swiss-Prot ou CCDS (NCBI).

3. Reusable

L'objectif de l'initiative FAIR est d'optimiser la réutilisation des données par des moyens automatisés. Pour ce faire, les métadonnées et les données doivent être décrites avec précision afin d'être reproduites et/ou combinées dans différents contextes.

1. Les données et métadonnées sont richement décrites avec une pluralité d'attributs précis et pertinents.
    a. Les données et métadonnées sont publiées avec une licence d'utilisation des données claire et accessible.
    b. Les données et métadonnées sont associées à une provenance détaillée.
    c. Les données et métadonnées sont conformes aux normes communautaires relatives au domaine.

En reprenant, une dernière fois, l'exemple de la base de données ENSEMBL. 1. Selon le site il n'y a aucune restriction à l'accès ou à l'utilisation des données et métadonnées de la base concernée.

Tableau V.2 ***: Exemples de respect des directives FAIR par des bases de données biologiques.***

	Findable	**Accessible**	**Inter Usabilité**	**Réutilisabilité de données**
ENCODE	Oui	Oui	Non : Les noms de cellules et de protéines ne suivent pas les conventions externes (par exemple ENCODE n'utilise pas les noms définis par HGNC pour les protéines).	Oui : De nombreuses métadonnées sont disponibles et suivent un convention interne solide.
GEO	Oui	Non	Non : Il n'y a aucune convention interne ou externe pour les métadonnées.	Non : Il n'y a aucune convention interne.
Array Express	Oui	Oui	Non : Les noms de cellules et de protéines ne suivent pas les conventions externes.	Oui : De nombreuses métadonnées sont disponibles et suivent un convention interne solide.

B. Outils informatiques et reproductibilité

L'explosion des données biologiques a aussi conduit à un challenge dans l'analyse et le traitement des données. En prenant l'exemple d'expériences de tous types sur tous les organismes contenus dans la base de données ENCODE, les expériences sont passées d'une dizaine en 2010 à plus de 12 000 en 2019. Les autres bases de données telles que GEO ont connu une évolution similaire (Figure V.1).

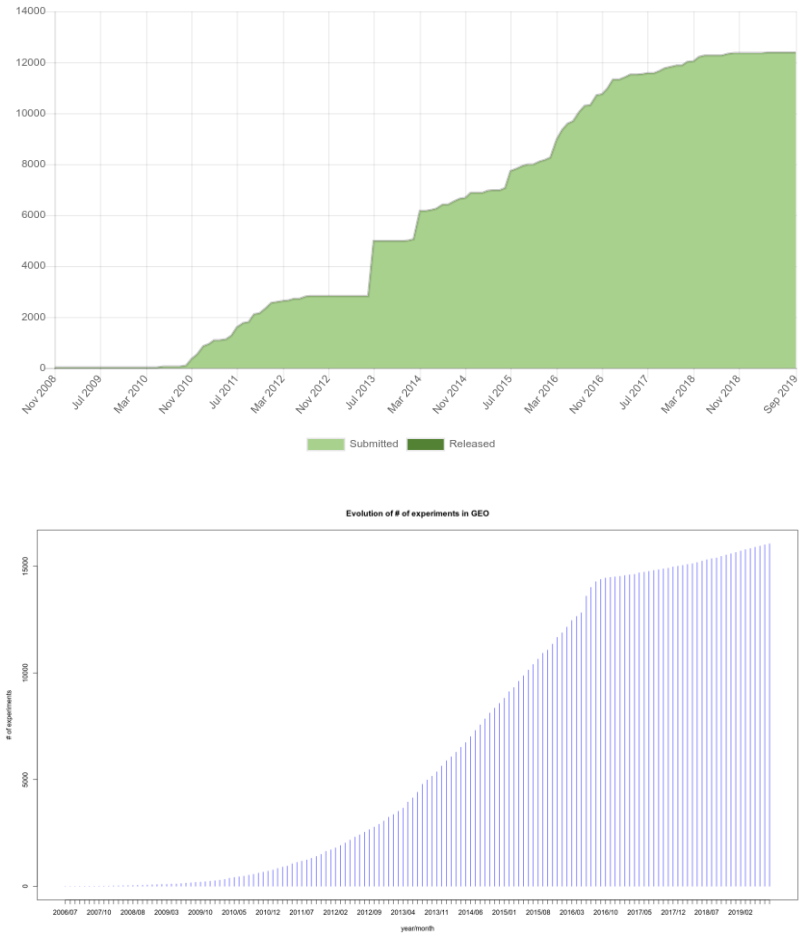


Figure V.1 : Nombre d'expériences dans les bases de données ENCODE et GEO.
Le graphique du haut représente le nombre d'expériences disponibles dans la base de données du consortium ENCODE de 2008 à 2019. L'évolution du nombre de données suit une augmentation constante. Les décrochements correspondent à la sortie des différentes phases d'ENCODE. Il y a actuellement plus de 12 000 expériences concernant l'annotation des CRE disponibles.
Le graphique du bas représente le nombre d'expériences disponibles dans la base de données GEO de 2007 à 2019 importées dans AE. Une augmentation exponentielle est observée jusqu'au début 2017. Cette occurrence est le résultat du changement de politique d'import de données d'AE. Depuis 2017, cette base de données n'importe que les résultats et données brutes issues de RNA-seq et de single-cell.

Ce nombre de données toujours croissant suscite des questions quant à leur traitement. L'analyse et l'intégration de ces données repose, de plus en plus, sur l'automatisation du traitement de l'information. On ne peut pas traiter l'analyse de plusieurs milliers de données de la même façon qu'une analyse isolée sur un ordinateur personnel. Il faut mettre en place des stratégies de traitement des données. Ces stratégies peuvent concerner la puissance de calcul ou les outils bioinformatiques mis en œuvre. Il faut mettre en place des solutions qui soient reproductibles et transposables afin de s'assurer de la pérennité des données produites. Un pipeline est généralement constitué de composants linéaires, où l'un des outils logiciel en alimente un autre.

1. Modularité des pipelines

Au vu de la taille des données biologiques produites, il devient nécessaire d'utiliser des supercalculateurs capables d'une plus grande puissance de calcul, d'un plus grand espace de stockage, ainsi que de la possibilité de parallélisation des processus.

Ces supercalculateurs ou HPC sont des clusters de "machines" ou nœuds reliés entre eux par un nœud maître capable de répartir les tâches et d'accéder aux données stockées. Ces nœuds sont eux-mêmes composés de cœurs (*node*), la plus petite unité physique de traitement de l'ordinateur (Figure V.1). Les cœurs peuvent, eux-mêmes, être divisés en fils d'exécution ou thread, mais cette division est purement logicielle. Les cœurs d'un même nœud peuvent communiquer entre eux. Ce n'est pas le cas pour les nœuds d'un même cluster. Il est donc nécessaire d'avoir un nœud maître qui peut envoyer et récupérer des tâches aux nœuds du cluster. Les logiciels capables de distribuer les tâches à travers les nœuds sont appelés Portable Batch System (PBS) tel que TORQUE[a] et SLURM[b] (Simple Linux Utility for Resource Management).

a : www.adaptivecomputing.com/products/torque
b : slurm.schedmd.com

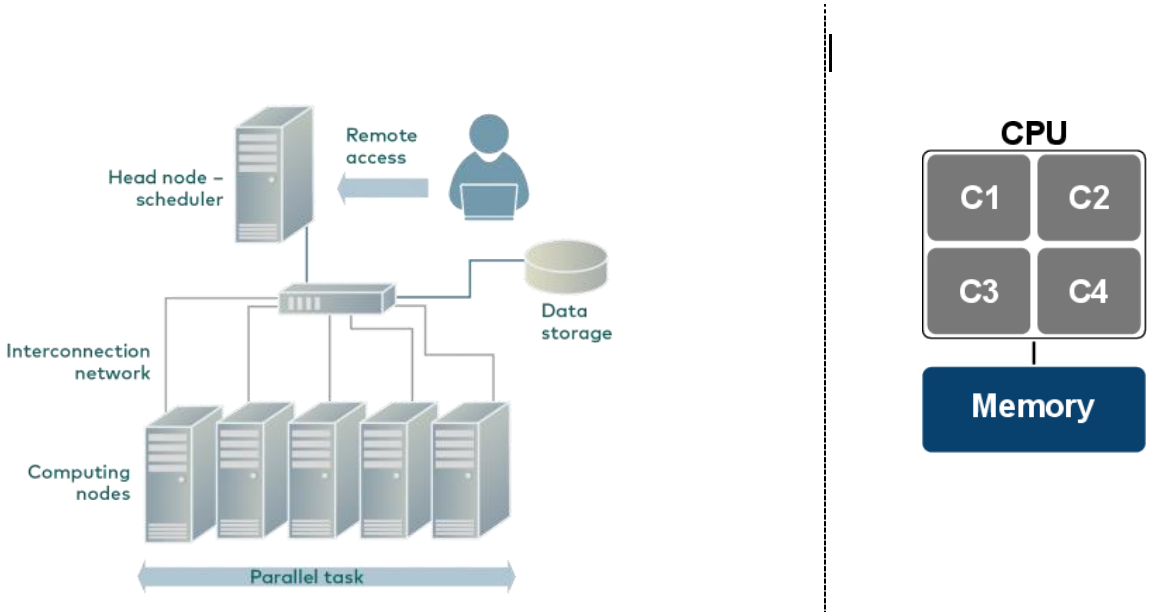


Figure V.2 : représentation de la structure d'un cluster de calculs (HPC) et de ses nœuds.
Le schéma du haut représente l'architecture simplifiée d'un HPC. Un HPC est composé d'un nœud maître (head node - scheduler). Le nœud maître est une machine capable : de contrôler et d'accepter les accès à distance, d'avoir accès aux données stockées, de consulter la disponibilité et de distribuer les tâches sur les nœuds de calculs (computing nodes) grâce au gestionnaire de tâche (scheduler). Les données stockées et les nœuds de calculs sont physiquement séparés sur des machines différentes. En conséquence, les interactions entre ces machines sont assurées par le nœud maître via le réseau interne.
Un nœud de calculs est composé d'un processeur unique (CPU) et de mémoires vives (memory) assignées à ce CPU. Un CPU est composé de plusieurs cœurs (core, C1 à C4) dont la coopération est assurée par le CPU.

En bioinformatique, la stratégie courante de parallélisation consiste à prendre une application non parallèle existante et à diviser les données en unités de travail discrètes (en jobs), à travers de multiples cœurs, et nœuds du cluster. La parallélisation est assurée par les pipelines. Elle permet de considérablement réduire le temps de calcul total. La parallélisation des tâches présente plusieurs inconvénients : le déploiement et la configuration des pipelines, la gestion et la complexité de la division des entrées, de la collecte et du rassemblement des données de sortie. De plus, les pipelines sont potentiellement fragiles, car il n'y a pas de communication directe entre les processus envoyés sur différents nœuds de calculs. Par exemple, il est difficile de prédire les conséquences d'une erreur de stockage ou de réseau survenue au cours d'une semaine ou d'un mois de calculs. Plusieurs gestionnaires de workflow ont vu le jour pour remédier à ce problème.

Figure V.2 : http://www.iue.tuwien.ac.at/phd/weinbub/dissertationsu16.html

Les gestionnaires de workflow se chargent de mettre en relation toutes les étapes du pipeline, de suivre leur exécution et de configurer la parallélisation sur différentes architectures. Ils utilisent leur propre langage et syntaxe. En bioinformatique, les gestionnaires de workflow les plus courants sont Snakemake[150] et Nextflow[151]. Au cours de ma thèse, j'ai utilisé Snakemake car il est basé sur le langage Python et est utilisé au sein du laboratoire TAGC.

Snakemake permet la composition de workflows basés sur un graphique de règles dont l'exécution est déclenchée par la présence, l'absence ou la modification de fichiers et répertoires attendus. Il génère lui-même le graphique de dépendance des règles (*Direct Acyclic Graph*, DAG) et peut relancer uniquement les parties pertinentes du workflow. La construction du DAG se fait à partir de la sortie finale et remonte jusqu'aux fichiers d'entrée. Il est possible d'écrire des règles utilisant du Bash, du Python et du R nativement. Snakemake permet de configurer, pour chaque règle, les ressources demandées, ainsi que de définir l'environnement à utiliser (voir partie suivante).

Nextflow est une alternative à Snakemake. A la différence de Snakemake le DAG est construit à partir des fichiers d'entrées et "descend" jusqu'à la sortie. En conséquence, les différentes étapes peuvent être générées dynamiquement au cours du workflow et permettre les embranchements dans le pipeline. Le gestionnaire de workflow Nextflow supporte l'utilisation du langage Groovy basé sur Java[a].

Grâce à l'utilisation des HPC et des gestionnaires de workflow tels que Snakemake il est donc possible de moduler relativement facilement les pipelines d'analyse et d'intégration de données.

2. Reproductibilité des pipelines

La reproductibilité des données est un enjeu majeur de la biologie. Passer du traitement des données sur un ordinateur local par l'utilisation d'un simple pipeline "maison" au Big Data et au traitement de données à haut débit n'est pas anodin. Une grande question se pose : Comment reproduire et vérifier les résultats obtenus par des pipeline qui ont opéré sur une machine particulière.

En science, les publications dans les revues scientifiques sont basées sur "l'évaluation par les pairs" (peer review). Une expérience impossible à reproduire a peu de valeur dans le milieu scientifique. Si ce principe est acquis dans le milieu expérimental de la biologie, il l'est beaucoup moins dans celui de la bioinformatique. Il n'est pas rare d'observer dans les matériels et méthodes des articles scientifiques un manque total d'informations permettant la reproductibilité des données (outils ou script bioinformatique utilisés, paramètres des outils, version, etc.). Plusieurs groupes et personnes se sont posés la question de la reproductibilité des données.

Une meilleure reproductibilité des workflows passe, tout d'abord, par une meilleure traçabilité des outils et scripts utilisés. En plus d'une documentation du code et du projet il faut également fournir les informations minimales permettant la reproductibilité des expériences :

- Système (OS) et version où les scripts ont tourné (Ubuntu, MAC, etc)
- Schéma du pipeline avec étapes + entrée et sortie.
- Scripts utilisés et commentés
- Outils bioinformatiques utilisés : numéro de version, paramètres utilisés.
- Description des fichiers d'entrée et de sortie (origine, format, information contenue, optionalité des fichiers, etc).

Pour toutes ces raisons, il est conseillé d'utiliser des contrôleurs de version décentralisés, tels que GIT afin de s'assurer de la traçabilité de son pipeline. GIT permet de créer des "points de sauvegarde" des dossiers et fichiers contenant le pipeline sur un serveur distant au cours du développement (Figure V.3). Cela permet de revenir facilement à une version précédente, mais aussi de créer des branches de développement sans toucher à la version principale du workflow. GIT permet donc de partager et de coopérer plus facilement sur la création et le développement de pipeline. L'utilisation de GIT couplée à des sites d'hébergement proposant des services comme la mise en place de documentation en ligne, et de communication entre développeur et utilisateur. GitLab, SourceSup et GitHub permettent de s'assurer d'une meilleure traçabilité et reproductibilité du workflow.

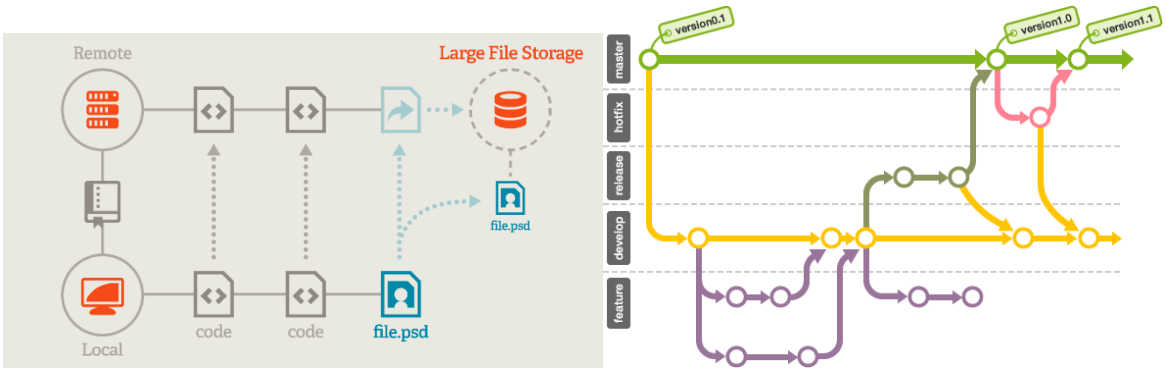


Figure V.3 : Schémas du fonctionnement d'un contrôleur de version tel que Git.
Le code développé localement peut être sauvegardé sur des répertoires distants. Chaque événement de sauvegarde crée une copie du code et des différences apportées par les nouvelles versions. Il est possible de revenir sur les versions antérieures du code. Les contrôleurs de version ne sont pas appropriés au stockage de fichiers non textuels car un de leur intérêt est de pouvoir comparer ligne à ligne les différences apportées d'une version à l'autre.
La plupart des contrôleurs supportent la création de branches de développement secondaires. Une branche permet la sauvegarde et la modification de fichiers sans influencer sur les fichiers sauvegardés sur la branche de développement principale (master). Cette caractéristique permet de poursuivre le développement sur un projet sans en altérer le fonctionnement tant que les branches ne sont pas fusionnées.

Une meilleure reproductibilité des pipelines passe aussi par la portabilité des pipelines. L'accès à un pipeline et à ses scripts n'assure pas forcément la reproductibilité des données. En effet, un pipeline développé pour une machine ou un serveur particulier n'est pas forcément adaptable pour d'autres architectures (PBS, OS différent, dépendance logicielle manquante, etc.). Une trop grande rigidité dans le développement du pipeline est souvent en cause. Utiliser des gestionnaires de workflow peut permettre de pallier cette rigidité. Comme il est dit précédemment, ces gestionnaires garantissent une flexibilité dans l'architecture matérielle et logicielle utilisée pour l'exécution de pipeline. Ils proposent souvent des options pour "emballer" toutes les dépendances pour l'exécution du workflow ce qui permet de l'exécuter sur d'autres machines contenant uniquement le gestionnaire de workflow. Cependant, les gestionnaires de workflow ne règlent pas le problème de l'environnement utilisé pour chaque outil ou script invoqué par le pipeline. Des outils informatiques ont donc été développés afin d'isoler l'exécution de chaque outil dans un environnement contenant toutes les dépendances nécessaires à son exécution.

Figure V.3 : https://git-lfs.github.com
https://sjkelleyjrblog.wordpress.com/2017/08/26/the-blockchain-from-a-git-perspective

Plusieurs solutions existent.

La première est Conda, écrit en python. C'est un gestionnaire de package permettant de créer facilement un environnement isolé. Pour cela, il propose une liste de logiciels pouvant être installés. Conda se charge de télécharger et d'installer toutes les dépendances et de gérer les conflits dans un environnement indépendant du système ce qui permet d'installer plusieurs versions du même outil. En environnement, Conda peut facilement être créé à partir d'un fichier de configuration listant les outils à installer. Cette fonctionnalité est utilisée par Snakemake. Il est possible de fournir, pour l'exécution de chaque règle, un environnement Conda à utiliser. La liste des outils par Conda est étoffée par Bioconda[152]. Bioconda est géré par une équipe à temps plein avec collaborateurs. Il y a actuellement 6975 packages (13 Août 2019). Cette base de données grandit rapidement. En Décembre 2016, elle contenait 161 packages. En Juillet 2018, elle en contient près de 3000. Un des avantages de Conda est de ne pas nécessiter de privilèges pour son installation. En effet, sur les HPC, il est courant de ne pas détenir les droits indispensables à l'installation des logiciels. Conda permet de contourner cette limite car il installe les logiciels dans un dossier local.

Docker et Singularity sont des alternatives à Conda. Ils se basent sur la création de conteneurs reproduisant n'importe quel OS. Il est facile d'importer une image Docker dans Singularity. Dans chaque conteneur, installer tous les logiciels compatibles, sans passer par une base de données est réalisable. Un des avantages de Docker/Singularity par rapport à Conda est donc de permettre une isolation totale des processus exécutés dans les conteneurs. De plus, il existe un Hub d'échange de conteneur ce qui évite la création d'un conteneur déjà disponible. Ce Hub n'étant pas géré, il faut, en contrepartie, gérer soi- même les dépendances et les conflits lié à l'OS utilisé. De plus, il faut détenir des droits administrateurs afin d'installer Docker/Singularity sur une machine. Ces droits administrateurs sont indispensables pour composer des conteneurs. L'utilisation de conteneurs Docker et Singularity est également gérée par Snakemake.

En conclusion, la reproductibilité des pipelines bioinformatiques passe par le suivi d'une ligne directrice simple. Il faut que le pipeline et les scripts soient documentés. Il est indispensable que le pipeline et ses dépendances soit facilement partageables, réutilisables et portables. Ces contraintes passent par une combinaison de l'utilisation de logiciel de contrôle de version à distance, de gestionnaire de workflow et d'environnement

A retenir pour mes travaux

Dans la recherche, la reproductibilité peut s'effectuer à plusieurs niveaux. Les données et métadonnées doivent être trouvables (findable), accessibles (accessible), interexploitables (interoperable) et réutilisables (reusable). La communauté scientifique s'est regroupée dans le consortium FAIR afin de formaliser ces principes. A cela s'ajoute la reproductibilité dans le processus scientifique. Pour la bioinformatique, de nombreux outils sont disponibles afin de faciliter le partage et la reproductibililté des données. Au cours de ma thèse, j'ai utilisé le gestionnaire de workflow Snakemake, les gestionnaires d'environnement Conda, Docker et Singularity et le contrôleur de version Git. L'augmentation des besoins en calculs m'ont conduit à utiliser différentes architectures de serveur possédant différents gestionnaires de tâches TORQUE et SLURM.

Résultats

I ReMap 2018

A. Introduction

La fixation des facteurs de transcription (TF) sur l'ADN est un moyen sans a priori d'identifier la présence d'éléments cis-régulateurs (CRE) dans le génome des eucaryotes. La fixation des protéines régulatrices est complexe. Elle dépend de nombreuses conditions épigénétiques et varie, au cours du temps, en fonction des cellules et du contexte cellulaire. Par conséquent, la création d'une cartographie complète de fixation directe et indirecte des protéines régulatrices nécessite un nombre important de données recensant les interactions protéines/ADN, et ce, pour toutes les lignées/tissus et conditions cellulaires possibles et pour toutes les protéines régulatrices. La méthode d'immunoprécipitation suivie de séquençage à haut débit (ChIP-seq) permet d'identifier toutes les interactions d'une protéine cible avec l'ADN le long du génome. Depuis 2008, les données ChIP-seq disponibles publiquement (GEO, Array Express, ENCODE) connaissent une évolution exponentielle. Ces deux critères font des données générées par ChIP-seq une méthode idéale pour l'analyse intégrative à haut débit (big data) dont l'objectif est d'identifier les régions de fixation des TFs et le CREs.

Les données de ChIP-seq produites par les laboratoires publiant dans des revues scientifiques et la plupart des consortia sont facilement accessibles à travers leurs sites internet propres (https://www.ebi.ac.uk/arrayexpress/, https://www.encodeproject.org) ou grâce à des entrepôts de données comme GEO et ENA (https://www-ncbi-nlm-nih-gov/geo, https://www.ebi.ac.uk/ena). Les analyses intégratives des données issues de données publiques peuvent s'avérer compliquées car les bases de données dont les expériences sont issues ne respectent pas toutes les principes FAIR, à savoir, la possibilité de parcourir, d'accéder, d'interopérer et réutiliser les données et métadonnées produites (Tableau I.1). En conséquence, il est nécessaire d'annoter uniformément les données. De plus, les expériences de ChIP-seq issues des différentes bases de données suivent des protocoles et des design différents (single-end/paired-end, réplica ou non, présence ou absence de contrôle, contrôle de type WCE ou IgG). Il devient indispensable de développer un pipeline robuste et adaptatif.

B. Le projet ReMap

L'objectif du projet ReMap est de fournir un atlas, le plus complet possible, des régions régulatrices, chez l'homme par l'analyse intégrative de données ChIP-seq issues des entrepôts de données GEO, Array Express (AE) et du consortium ENCODE. Le but est de fournir à la communauté scientifique des données de fixation des protéines régulatrices de la transcription, passant un filtre de qualité, de façon accessible et compréhensible par le plus grand nombre. Le pipeline général du projet ReMap est simple. Il repose sur le retraitement des données brutes issues d'expériences de ChIP-seq. Ces données sont présentées sous la forme de fichiers texte (FASTQ) contenant les séquences des fragments de lecture, ainsi que le score de qualité de leur séquençage. Au final, ReMap fournit des fichiers renseignant la position des régions de fixation de protéines régulatrices ou de CREs sous format texte BED (Browser Extensible Data). Le pipeline général repose sur 4 étapes simples :

1. Annotations manuelles des données publiques.
2. Traitement des données brutes de façon uniforme.
3. Contrôle qualité.
4. Intégration des données pour identifier les régions régulatrices.

1. Annotation manuelle

L'étape d'annotation manuelle est indispensable car les différentes bases de données ne suivent pas forcément les principes FAIR (Tableau I.1).

Tableau I.1 : Application de FAIR pour les bases de données utilisées par ReMap.

	ENCODE	**ARRAY EXPRESS**	**GEO**
Findable	✓	✓	✓
Accessible	✓	✓	✗
Interexploitable	✗	✓	✗
Reusable	✓	✓	✗

Une des difficultés rencontrées est l'absence d'interopérabilité entre les données et les métadonnées. Le nom des protéines et des lignées/tissu cellulaires issues des différentes bases de données ne suivent pas les mêmes conventions d'appellation et ne font pas référence, dans leurs métadonnées, aux noms alternatifs. Il est donc nécessaire d'uniformiser les données par l'annotation. Dans le cadre de ReMap 2018 nous utilisons HUGO Gene Nomenclature Committee pour le nom des protéines cibles[153] et le BRENDA Tissue Ontologies[154] (BTO) pour les lignées cellulaires et tissus.

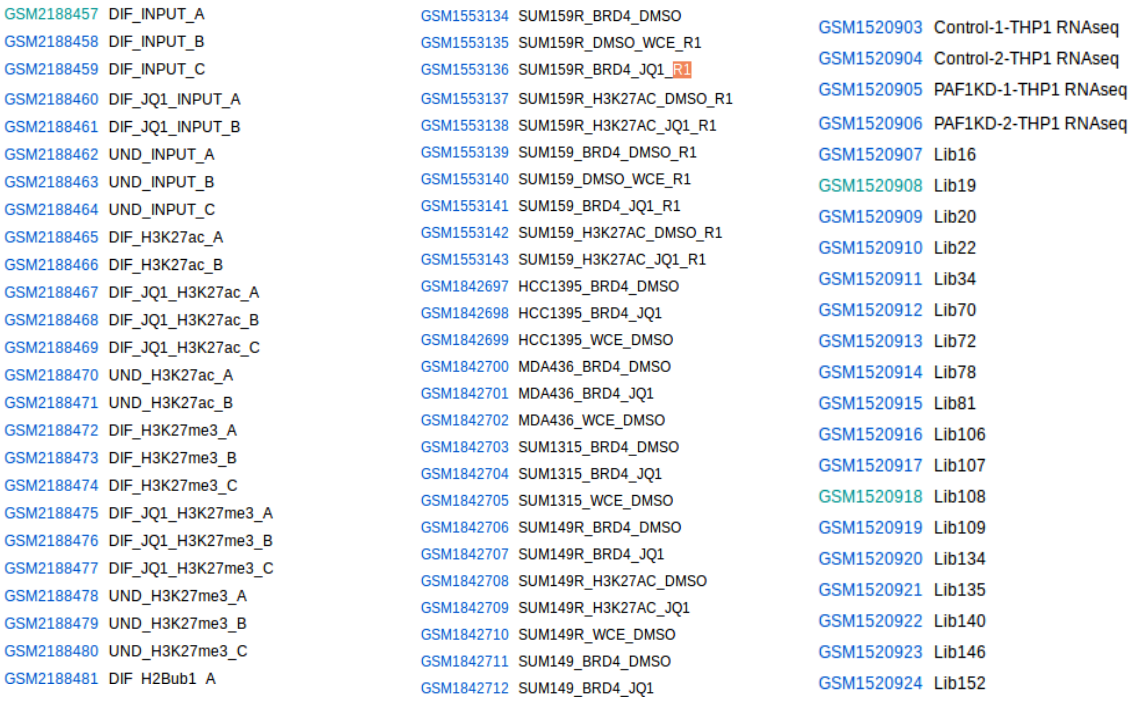


Figure I.1 : Exemples des métadonnées provenant de trois articles différents présentés sur le portail GEO.
A chaque article scientifique correspond une collection (serie) sur GEO. Chaque collection liste les fichiers soumis par le laboratoire. Cette liste comprend, dans le désordre, les différents réplicas et contrôles de toutes les expériences quels que soient leur types (ChIP-seq, RNA-seq, ATAC-seq, etc.). Il faut donc, dans un premier temps, regrouper les fichiers par expériences. Dans le cas des expériences ChIP-seq présentées ici, il est nécessaire d'annoter la protéine cible et la lignée cellulaire. Dans le premier exemple, le nom de la cellule n'apparaît pas, ce qui exige une recherche approfondie. De plus, les contrôles sont nommés A, B ou C, ce qui n'est pas le cas pour le deuxième exemple où ils sont nommés R1. Par ailleurs, le premier exemple nomme ces contrôles INPUT et le deuxième WCE. Dans le dernier exemple, aucune information sur la protéine cible, le nom de la lignée cellulaire ou de son type n'apparaît dans le nom du fichier. Les fichiers sont nommés Lib16, Lib19, etc… Ces différents exemples mettent en lumière la difficulté d'établir une annotation automatique, compte-tenu de la grande diversité des annotations proposées par GEO.

Les données issues de l'entrepôt de données GEO génèrent une autre difficulté. Les données et métadonnées proposées par GEO ne sont pas soumises à une convention d'appellation interne. Les métadonnées ainsi que leurs accessibilité, interopérabilité et réutilisation sont bridées par plusieurs limitations bien humaines. La mise en ligne des métadonnées et données sur GEO est requise pour la publication d'articles scientifiques, mais la soumission des métadonnées n'est régie par aucune consigne précise et la liste d'informations à fournir est peu ou mal documentée. Concrètement, cela signifie que les métadonnées soumises par différents expérimentateurs et/ou laboratoires ne suivent pas d'ontologies ou d'annotations identiques. De plus, les données et métadonnées ne sont pas nécessairement soumises par les expérimentateurs, ce qui peut mener à des erreurs ou à des oublis d'informations.

Pallier ces difficultés nécessite une annotation uniforme de toutes les expériences issues de GEO. D'autant plus qu'une annotation automatique n'est pas complètement réalisable au regard de l'extrême hétérogénéité des métadonnées, ce qui, dans certain cas, peut contraindre à lire le papier d'origine (Figure I.1).

L'annotation manuelle des métadonnées, tout comme l'uniformisation des noms de protéines cibles et des noms de lignées, sont donc des étapes longues et fastidieuses, mais indispensables au traitement et à l'intégration de toutes les données.

2. Traitement des données

Le traitement de données par ReMap est composé d'étapes successives représentées dans la figure I.2

Pour le retraitement des données, la difficulté repose sur le choix des outils bioinformatiques à utiliser. Actuellement il n'y aucun consensus sur un outil ou sur une suite d'outils les plus performants pour toutes les analyses bioinformatiques. Prenons pour exemple l'étape d'alignement des fragments de lecture sur le génome de référence. OmiX tools[a], entreprise recensant les outils bioinformatiques, dénombre actuellement 377 outils d'alignement. Chacun de ces outils se différencie par l'algorithme utilisé, la facilité d'utilisation, etc.

Devant ce grand nombre d'outils, le choix est compliqué. Des articles comparent ces outils[155,156]. Le but de ces articles est, le plus souvent, de comparer les algorithmes utilisés afin de déterminer les meilleurs outils dans des conditions bien précises. La qualité et la rapidité du résultat dépendent également des données à analyser. Le même problème se pose pour l'étape du peak-calling[157]. L'intégration de données à haut débit, si on la compare à des analyses isolées, présente une complexité supplémentaire. Les données étant d'origines diverses, les erreurs rencontrées lors de leur traitement peuvent être extrêmement variées (absence de lien de téléchargement, échec de l'alignement par manque de fragments de lecture, fichier corrompu, etc.).

a : omictools.com

Par conséquent, les outils utilisés doivent présenter une documentation complète, ainsi qu'un rapport des erreurs complet et facilement compréhensible. Le grand nombre de données traitées exige, pour chaque expérience, un compromis entre la rapidité d'exécution et l'adaptation des paramètres. Le choix s'est donc porté sur des outils rapides et robustes, fournissant une documentation complète et une communauté importante.

Le pipeline ReMap, dans ses différentes étapes, utilise les outils suivants. Pour le contrôle qualité Trim Galore est utilisé. Il regroupe Cutadapt qui se charge d'enlever les adaptateurs de séquençage et FastQC qui, pour sa part, se charge d'enlever les nucléotides dont la qualité de séquençage est en dessous du seuil fixé. Bowtie2 est utilisé pour sa rapidité d'exécution, sa facilité d'utilisation, sa documentation et son aide en ligne. Nous utilisons MACS2 (Model-based Analysis of ChIP-Seq), un outil de peak-calling basé sur la recherche de modèles (Introduction figure III.6). Le but, à long terme, est de rendre le pipeline d'analyse modulable afin d'incorporer de nouveaux peak-caller tels que SPP, développé par ENCODE[158] ou HOMER[159]. Cela permettrait de laisser le choix à l'utilisateur d'utiliser les différentes listes de pics en fonction de son usage. Pour la manipulation des fichiers textes SAM (Sequence Alignment Map) et fichiers binaires BAM (Binary Alignment Map) d'alignement, nous utilisons la suite de logiciels Samtools. Nous utilisons la suite d'outils Bedtools[160] pour manipuler les fichiers BED. En ce qui concerne le contrôle qualité du peak-calling, nous utilisons PhantomPeak, développé par le consortium ENCODE, pour sa capacité à calculer les scores de NSC et RSC.

3. Mise à disposition des données produites par ReMap

Le résultat de notre analyse intégrative est mis à la disposition de la communauté scientifique au travers du site http://tagc.univ-mrs.fr/remap (Figure I.4). Nous fournissons trois types de données :

- Des fichiers BED contenant tous les pics pour une expérience/protéine régulatrices de la transcription/lignée cellulaire et pour toutes les expériences.
- Des fichiers BED contenant les pics non redondants par TF et pour toutes les expériences (Figure I.3).
- Un fichier contenant les CRE putatifs (voir figure I.3 pour la méthode d'annotation). Nous considérons une région comme ayant un rôle régulateur dans la transcription si elle fixe au moins deux protéines régulatrices différentes.
- Les séquences nucléotidiques par pics de protéines régulatrices de la transcription sous format FASTA.

Le site fournit aussi un outil en ligne d'annotation des régions génomiques et de calcul d'enrichissement en facteurs de transcription basés sur le catalogue ReMap. Il permet aux utilisateurs de charger leurs régions d'intérêt afin d'identifier les protéines régulatrices de la transcription enrichies sur ces régions. Son fonctionnement est décrit dans le chapitre ReMap2020.

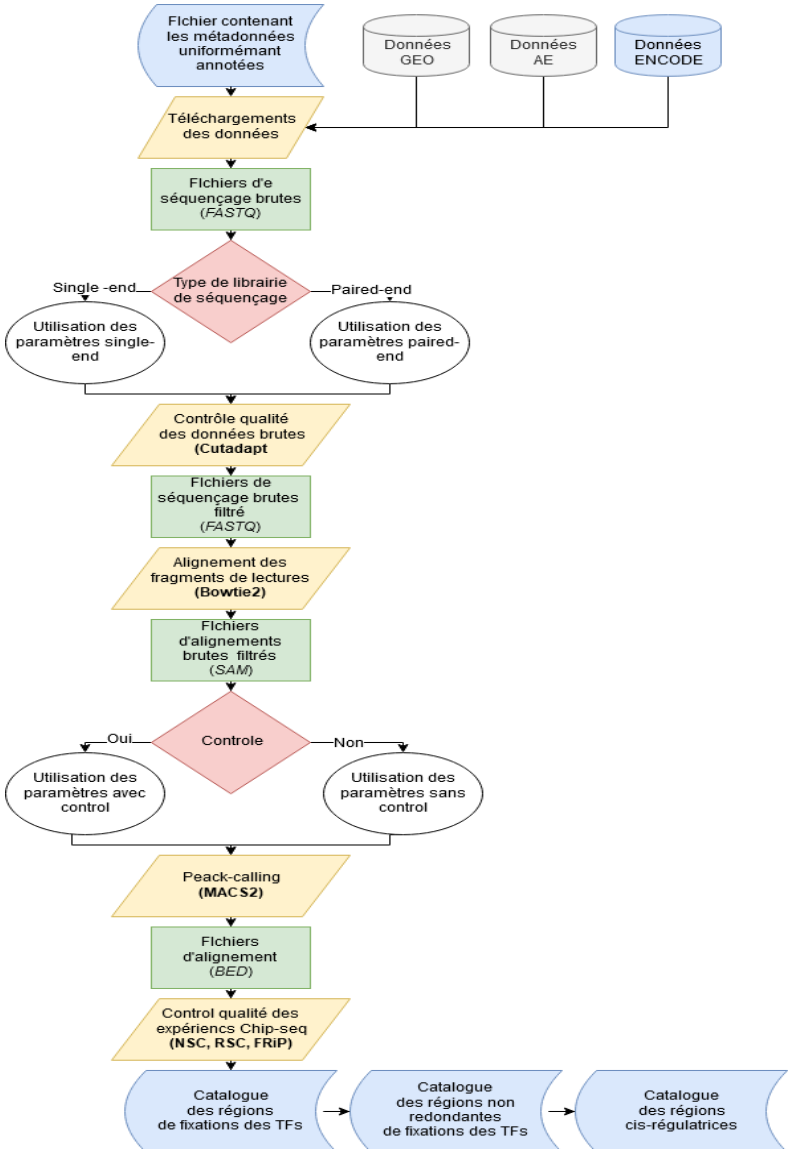


Figure I.2 : Diagramme du pipeline général de ReMap.
Dans un premier temps, on extrait les métadonnées nécessaires au téléchargement des données. Les données brutes, issues du séquençage sont téléchargées sous le format FASTQ. Les outils employés, de l'étape du contrôle qualité des FASTQ jusqu'à l'étape du peak-calling, traitent différemment les données issues de single ou de paired-end. Afin de leur appliquer les paramètres nécessaires, les expériences sont marquées en fonction de leur type de librairie. L'étape de contrôle qualité du séquençage des fragments de lecture est nécessaire car les erreurs de séquençage et la présence d'adaptateurs influent sur la qualité de l'alignement[161]*. L'outil utilisé pour cette étape est TrimGalore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore). Les fragments de lecture filtrés sont ensuite alignés sur le génome de référence par Bowtie2*[162]*. Bowtie2 produit un fichier texte indiquant pour chaque fragment de lecture, la ou les régions d'origine probables et un score évaluant la similarité entre la région d'origine et les fragments de lecture. Les pics correspondants aux régions de fixation sont détectés grâce à l'outil MACS2. Cette étape nécessite l'identification des réplicas et des contrôles pour chaque expérience de ChIP-seq. Les expériences sont filtrées par l'utilisation de métriques définies par le consortium ENCODE. Le NSC et le RSC permettent d'évaluer la qualité de la construction de la librairie, le FRiP permet d'évaluer la spécificité de l'anticorps utilisé (voir introduction). Les expériences sélectionnées permettent la création d'un catalogue de régions de fixation des protéines régulatrices. Ce catalogue permet d'extraire les fixations redondantes du même TFs pour plusieurs expériences et lignées cellulaires afin de fournir une liste de régions de fixation non redondantes. A partir de cette liste de fixations non redondantes, une liste de CREs putatifs est définie.*

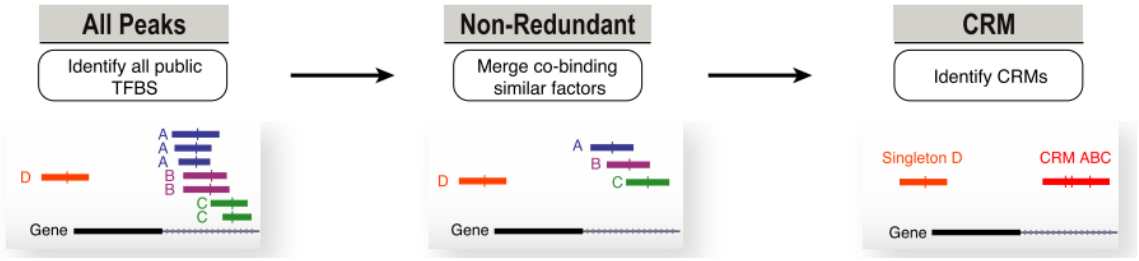


Figure I.3 : Méthode d'annotation des pics de fixation non redondants et des CREs. *Un fichier contenant les pics non redondants est produit pour chaque protéine. Un pic non redondant est le résultat de la fusion de plusieurs pics se chevauchant, provenant de plusieurs expériences de ChIP-seq. Dans cet exemple, les pics de la protéine A sont fusionnés entre eux. Il en est de même pour les protéines B et C. Les pics isolés (singleton) sont conservés. Le fichier de CRE putatif est constitué à partir du fichier contenant tous les pics redondants pour toutes les protéines. Les pics vont fusionner entre eux quels que soient la protéine et l'origine du pic. Nous considérons un pic fusionné comme étant un CRE s'il est constitué d'au moins deux pics non redondants de protéines différentes. Les singletons sont donc retirés. Les fichiers non-redondants et CREs putatifs sont réalisés à l'aide du logiciel Bedtools merge.*

Les régions de fixation des protéines, les pics non redondants et les CREs putatifs sont consultables via les navigateurs de génome en ligne d'UCSC et Ensembl. Nous mettons à disposition des sessions publiques, ainsi que des "Track Hubs". Les Track Hub sont des répertoires accessibles en ligne permettant d'accéder à des données stockées sur des serveurs HTTP ou FTP. L'avantage des Track Hub par rapport aux sessions publiques est qu'ils fournissent aussi les métadonnées liées aux données accessibles. Par exemple, nous fournissons dans notre Track Hub des informations sur le traitement des données, ainsi que sur le projet ReMap. La liste de ces Hubs est disponible sur les sites internet d'Ensembl et d'UCSC.

Figure I.3 : Griffon, A. *et al.* Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Res.* **43**, e27 (2015).

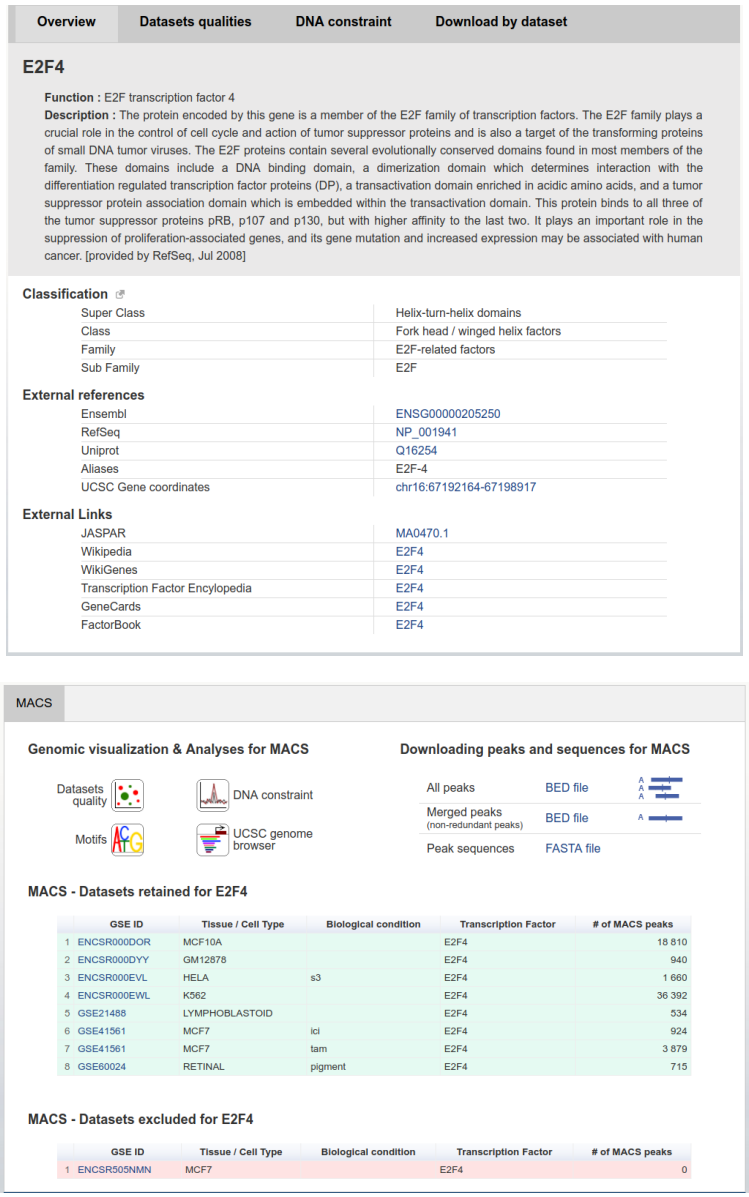


Figure I.4 : Accès aux métadonnées et données de ReMap par protéine.
Pour chaque protéine, ReMap fournit des métadonnées externes afin de faciliter l'interopérabilité. Nous mettons à disposition le nom officiel de la protéine ainsi que sa description et des informations concernant sa famille, provenant de la base de données Uniprot. Nous mettons, également, à disposition un lien externe vers la base de données JASPAR. Cette dernière utilise les données ReMap pour consolider sa base de données des sites de fixation des TFs. Les données des régions de fixation BED disponibles sont téléchargeables via un serveur HTTP. De plus nous renseignons, pour chaque protéine, la liste des expériences utilisées, leurs expériences d'origine, leurs lignées cellulaires ainsi que leurs modifications. Nous donnons également accès à un graphique représentant la qualité des expériences concernées (NSC, RSC). En complément, nous fournissons un lien vers le navigateur de génome UCSC afin de visualiser les pics, ainsi qu'un graphique renseignant l'état de la conservation de l'ADN à 1 Mb autour des pics de cette protéine. Ce graphique provient de l'outil Ensembl Compara[163].

4. ReMap2015 vs ReMap2018

La première itération de ReMap a été réalisée en 2015 avec des données allant de 2011 à 2014 par Benoît Ballester et son doctorant Aurélien Griffon[164]. ReMap 2015 contenait des informations des fixations de 132 protéines régulatrices de la transcription pour 83 lignées cellulaires différentes grâce au traitement de 395 expériences de ChIP-seq tirées de GEO, ainsi que l'intégration des données d'ENCODE. En 2014, les données d'ENCODE de ChIP-seq disponibles étaient réduites à un seul fichier contenant les pics de fixation des protéines régulatrices de la transcription pour toutes les lignées cellulaires étudiées sans moyen de les différencier.

En 2016, nous avons mis à jour ReMap avec de nouvelles données. En deux ans, le nombre d'expériences ChIP-seq est passé de ~600 expériences de ChIP-seq (avant filtrage par la qualité) à plus de 3000 en 2016. Cette considérable augmentation de données a amené à une réflexion sur la modularité du workflow de traitement ainsi que sur sa reproductibilité.

Afin de suivre les évolutions du traitement des données et de fournir des données de meilleure qualité, nous avons intégré de nouvelles étapes dans le pipeline d'analyse de ReMap. Nous avons ajouté une étape de trimming, permettant d'éliminer les bases nucléotidiques de mauvaise qualité, ainsi que les adaptateurs de séquençage. Nous avons aussi ajouté des étapes post alignement permettant de retirer les réplicas PCR et les alignements peu fiables. En 2015, le peak calling était effectué avec MACS 1.4. En 2018, nous avons mis à jour cette étape en utilisant MACS2 qui améliore l'algorithme de MACS1.4 ce qui a permis d'augmenter la précision de détection des pics. MACS2 calcule la p-valeur à chaque emplacement génomique alors que MACS1.4 recherche, en premier lieu, les pics potentiels, puis calcule la p-valeur. L'algorithme de MACS1.4 conduisait à la sélection de faux positifs par la présélection de régions enrichies[a]. Le peak-calling est une étape cruciale de la détection de régions de fixation. J'ai donc testé, sur une même expérience, les pics obtenus par MACS1.4 et MACS2. J'ai pu observer une différence de nombre de pics allant du simple à un facteur 10. C'est, suite à ce constat, que nous avons décidé de retraiter, pour ReMap 2018, la totalité des données de 2015.

De plus, la décision a été prise de tester un nouveau peak-caller en parallèle : JAMM. Son intérêt réside dans sa prise en compte des réplicas biologiques pour identifier les régions de fixation des protéines. L'ajout de JAMM a permis de mettre en perspective la modularité du pipeline d'analyse. Les données en entrées de JAMM étant différentes de celles de MACS2, Il fallait que le pipeline soit capable de traiter ces deux peak-caller de la façon la plus identique possible. Les différences sont résumées dans la figure I.5.

a : https://groups.google.com/forum/#!searchin/macs-announcement/macs2|sort:date/macs-announcement/Kzo6sdM4fTk/ye7BZeMb7E8J

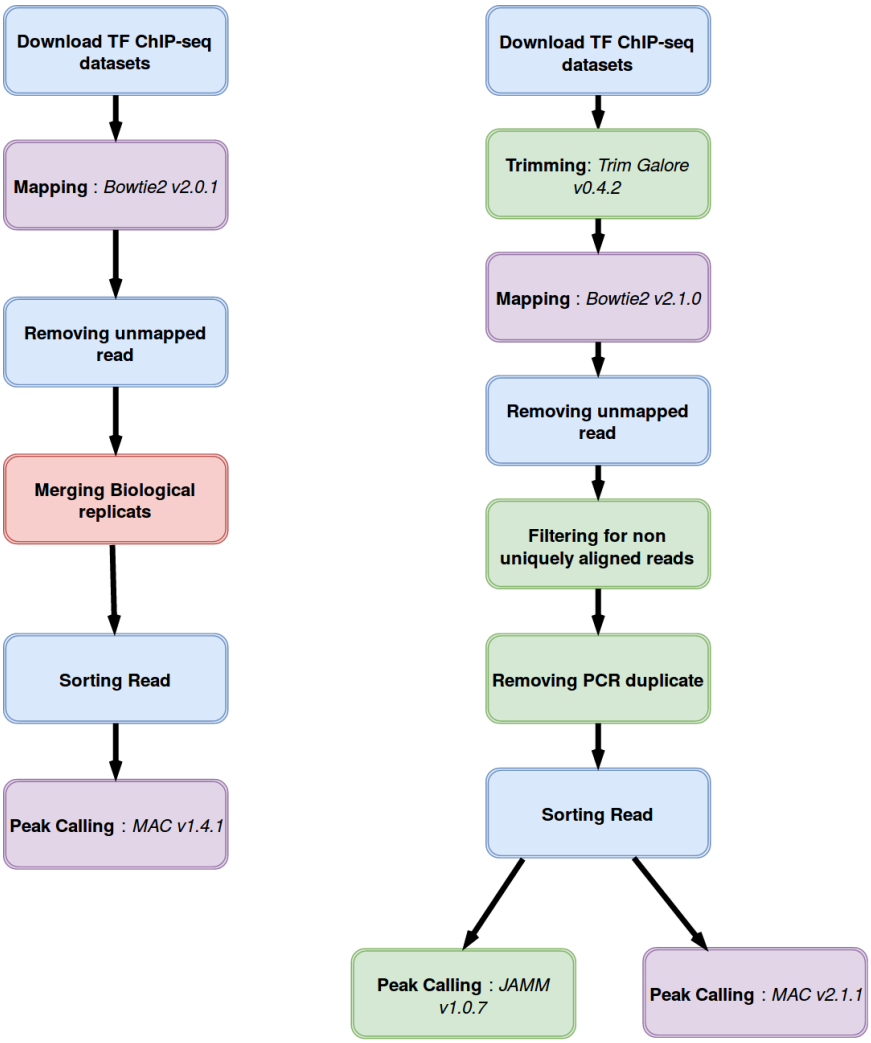


Figure I.5 : Différences dans le pipeline de ReMap2015 et ReMap2018.
En bleu, on retrouve les étapes identiques pour ReMap2015 et ReMap2018. En violet, les étapes identiques, mais avec un outil différent. En rouge, les étapes supprimées et en vert les nouvelles étapes. Une étape de trimming a été rajoutée afin d'améliorer la qualité de l'alignement des fragments de lecture. Nous utilisons la dernière version de Bowtie2 disponible lors du développement du pipeline. Dans les deux pipelines, nous enlevons les fragments de lecture qui n'ont pas été alignés sur le génome afin d'alléger les fichiers d'alignement. Dans la version 2018, nous ajoutons des étapes permettant de filtrer les fragments alignés. Enlever les fragments alignés plusieurs fois dans le génome permet de retirer les fragments peu complexes et les alignements faux positifs. Retirer les réplicas PCR est nécessaire pour empêcher que ces régions ne soient détectées en tant que pics lors du peak-calling. Pour toutes les étapes de manipulation des fichiers d'alignement, nous avons utilisé Samtools. Finalement, pour le peak-calling nous utilisons en parallèle les outils MACS2 et JAMM.

Le catalogue ReMap2015 était aligné sur la version hg19 du génome de référence qui date de 2009. Une nouvelle version du génome de référence est disponible depuis 2013 : hg38. La version hg19 contient environ150 000 *gap* dans le génome humain contre 250 pour hg38[165]. Nous avons décidé d'utiliser hg38. Néanmoins, nous avons converti les données hg38 en hg19, car de nombreuses équipes utilisent encore ce génome de référence.

Le pipeline de ReMap 2015 consistait en une suite de scripts bash exécutés sur un serveur unique. En raison de l'augmentation du nombre de données et pour réduire le temps de traitement nous sommes passés, en 2018, d'un serveur unique à un cluster de serveurs. Cette montée en charge nous a forcés à réfléchir au développement d'un workflow facilement transportable et modulable. Pour répondre à nos critères, nous avons décidé d'utiliser le gestionnaire de workflow Snakemake, développé par Johannes Köster. La difficulté a été d'adapter les suites de scripts de ReMap 2015 à un pipeline Snakemake en un temps limité. Le nouveau workflow est présenté dans la figure I.2 et les difficultés d'implémentation dans la discussion.

Pour la mise à jour de ReMap2018 les données issues d'ENCODE ont été traitées par Marius Georges de l'équipe d'Anthony Mathelier en collaboration sur le projet de mise à jour de JASPAR2018[124].

Published online 8 November 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D267–D275***
doi: 10.1093/nar/gkx1092

ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments

Jeanne Chèneby[1,2], Marius Gheorghe[3], Marie Artufel[1,2], Anthony Mathelier[3,4] and Benoit Ballester[1,2,*]

[1]INSERM, UMR1090 TAGC, Marseille F-13288, France, [2]Aix-Marseille Université, UMR1090 TAGC, Marseille F-13288, France, [3]Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway and [4]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, 0310 Oslo, Norway

Received September 15, 2017; Revised October 19, 2017; Editorial Decision October 20, 2017; Accepted October 20, 2017

ABSTRACT

With this latest release of ReMap (http://remap.cisreg.eu), we present a unique collection of regulatory regions in human, as a result of a large-scale integrative analysis of ChIP-seq experiments for hundreds of transcriptional regulators (TRs) such as transcription factors, transcriptional co-activators and chromatin regulators. In 2015, we introduced the ReMap database to capture the genome regulatory space by integrating public ChIP-seq datasets, covering 237 TRs across 13 million (M) peaks. In this release, we have extended this catalog to constitute a unique collection of regulatory regions. Specifically, we have collected, analyzed and retained after quality control a total of 2829 ChIP-seq datasets available from public sources, covering a total of 485 TRs with a catalog of 80M peaks. Additionally, the updated database includes new search features for TR names as well as aliases, including cell line names and the ability to navigate the data directly within genome browsers via public track hubs. Finally, full access to this catalog is available online together with a TR binding enrichment analysis tool. ReMap 2018 provides a significant update of the ReMap database, providing an in depth view of the complexity of the regulatory landscape in human.

INTRODUCTION

Transcription factors (TFs), transcriptional coactivators (TCAs) and chromatin-remodeling factors (CRFs) drive gene transcription and the organization of chromatin through DNA binding. TFs specifically bind to DNA sequences (TF binding sites) to activate (activators) or repress (repressors) transcription, TCAs enhance gene transcription by binding to activator TF. While CRFs modify the chromatin architecture to allow DNA access for transcription machinery proteins. In recent years, the development of high-throughput techniques like chromatin immunoprecipitation followed by sequencing (ChIP-seq) (1) has allowed to experimentally obtain genome-wide maps of binding sites across many cell types for a variety of DNA-binding proteins. The popularity of ChIP-seq has led to a deluge of data in current data warehouses (2,3) for TFs, TCAs and CRFs, collectively named transcriptional regulators (TRs). The rapid accumulation of ChIP-seq data in public databases provides a unique and valuable resource for hundreds of TR occupancy maps. There is a strong need to integrate these large-scale datasets to explore the transcriptional regulatory repertoire. Unfortunately, the heterogeneity of the pipelines used to process these data, as well as the variety of underlying formats used, challenge the analysis processes and the underlying detection of TF binding sites (TFBSs). Integrative studies would offer significant insights into the dynamic mechanisms by which a TF selects its binding regions in each cellular environment.

ReMap has been the first large scale integrative initiative to study these data, offering significant insights into the complexity of the human regulatory landscape (4). The ReMap 2015 resource created a large catalog of regulatory regions by compiling the genomic localization of 132 different TRs across 83 different human cell lines and tissue types based on 395 non-ENCODE datasets selected from Gene Expression Omnibus (2) and ArrayExpress (3). This catalog was merged with the ENCODE multi-cell peaks (5), generating a global map of 13M regulatory elements for 237 TRs across multiple cell types. However, since the 2015 publication of ReMap, an even greater number of ChIP-seq assays has been submitted to genomic data repositories.

*To whom correspondence should be addressed. Tel: +33 4 91 82 87 39; Fax: +33 4 91 82 87 01; Email: benoit.ballester@inserm.fr

Here, we introduce the ReMap 2018 update, which includes the integration of 2829 quality controlled ChIP-seq datasets for TFs, TCAs and CRFs. The new ChIP-seq datasets ($n = 1763$, defined as 'Public' for non-ENCODE) as well as the latest ENCODE ChIP-seq data ($n = 1066$) have been mapped to the GRCh38/hg38 human assembly, quality filtered and analyzed with a uniform pipeline. In this update, we propose a unified integration of all public ChIP-seq datasets producing a unique atlas of regulatory regions for 485 TRs across 346 cell types, for a total of 80M DNA binding regions. Each experiment introduced in this release has been assessed and manually curated to ensure correct meta-data annotation. Our ReMap database provides DNA-binding locations for each TR, either for each experiment, at cell line or primary cell level, or at the TR level in a non-redundant fashion across all collected experiments. This update represents a 2-fold increase in the number of DNA-binding proteins, 7-fold in the number of processed datasets, 4-fold in the number of cell lines/tissue types and 6-fold in the number of identified ChIP-seq peaks. While the first version of the ReMap catalog covered 26% (793 Mb) of the human genome, the regulatory search space for ReMap 2018 covers 46% (1.4Gb).

Finally, we give the community access to various options to visualize and browse our catalog, allowing users to navigate and dissect their genomic loci of interest co-occupied by multiple TRs in various cell types. Browsing the ReMap 2018 catalog using the Public Track hub, IGV data sever, Ensembl or UCSC sessions clearly exposes the abundance and intricacy of combinatorial regulation in cellular contexts.

This report presents the extensive data increase and regulatory catalog expansion of ReMap as a result of our large-scale data integration and genome-wide analysis efforts. The manual curation specific to the ReMap initiative offers a unique and unprecedented collection of TR binding regions. These improvements, together with several novel enhancements (search bars, data track displays, format and annotation), constitute a unique atlas of regulatory regions generated by the integration of public resources.

MATERIALS AND METHODS

Available datasets

ChIP-seq datasets were extracted from the Gene Expression Omnibus (GEO) (2), ArrayExpress (AE) (3) and ENCODE (5) databases. For GEO, the query '('chip seq' OR 'chipseq' OR 'chip sequencing') AND 'Genome binding/occupancy profiling by high-throughput sequencing' AND 'homo sapiens'[organism] AND NOT 'ENCODE'[project]' was used to return a list of all potential datasets, which were then manually assessed and curated for further analyses. For ArrayExpress, we used the query (Filtered by organism 'Homo sapiens', experiment type 'dna assay', experiment type 'sequencing assay', AE only 'on') to return datasets not present in GEO. Contrary to other similar databases (chip-atlas http://chip-atlas.org, (6,7)), ReMap meta-data for each experiment are manually curated, annotated with the official gene name from the HUGO Gene Nomenclature Committee (8) (www.genenames.org) and BRENDA Tissue Ontologies (9) for cell lines (www.ebi.ac.uk/ols/ontologies/bto). Datasets involving polymerases (Pol2 and Pol3), and some mutated or fused TFs (e.g. KAP1 N/C terminal mutation, GSE27929) were filtered out. A dataset is defined as a ChIP-seq experiment in a given GEO/AE/ENCODE series (e.g. GSE37345), for a given TF (e.g. FOXA1), and in a particular biological condition (e.g. LNCaP). Datasets were labeled with the concatenation of these three pieces of information (e.g. GSE37345.FOXA1.LNCAP).

A total of 3180 datasets were processed (Supplementary Table S1). Specifically, we analyzed 2020 datasets from GEO (1862) and ArrayExpress (158) repositories (July 2008 to May 2017). We define these non-ENCODE datasets as the 'Public' set, in opposition to ENCODE datasets (1160) (full list of experiments in Supplementary Tables S2 and 3).

ReMap 2015 contained the multi-cell peak calling processed from ENCODE release V3 (August 2013). For the ReMap 2018 update, we re-analyzed, starting from the raw data, all ENCODE ChIP-seq experiments for TFs, transcriptional and chromatin regulators, following the same processing pipeline as the Public set. We retrieved the list of ENCODE data as FASTQ files from the ENCODE portal (https://www.encodeproject.org/) using the following filters: Assay: 'ChIP-seq', Organism: 'Homo sapiens', Target of assay: 'TF', Available data: 'fastq' on 21 June 2016. Metadata information in JSON format and FASTQ files were retrieved using the Python *requests* module. We processed 1160 datasets associated to 161 TRs and 87 cell lines. We removed 2 TRs (POLR2A, POLR3G), and renamed TR aliases into official HGNC identifiers (e.g. p65 into RELA, see Supplementary Table) leading to a final list of 279 TRs from ENCODE.

ChIP-seq processing

Both ENCODE and Public datasets were uniformly processed and analyzed. Bowtie 2 (version 2.2.9) (10) with options –end-to-end –sensitive was used to align all reads on the human genome (GRCh38/hg38 assembly). For Public datasets, adapters were removed using TrimGalore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), trimming reads up to 30 bp. Polymerase chain reaction duplicates were removed from the alignments with samtools rmdup (11). For the ENCODE data, the adapter trimming step was not employed, as this data already passed certain quality assessment steps (https://www.encodeproject.org/data-standards/). TR binding regions were identified using the MACS2 peak-calling tool (version 2.1.1.2) (12) in order to follow ENCODE ChIP-seq guidelines (13), with stringent thresholds (MACS2 default thresholds, *P*-value: 1e-5). Input datasets were used when available. All peak-calling files are available to download. Among the 80M peaks identified, 99.5% of peaks (79 753 407) were below 1.5 kb in size (mean size: 286 bp, median size: 231 bp) and only 376 017 peaks were above 1.5 kb in size (mean size: 2209 bp, median size: 1859 bp).

Quality assessment

As raw data are obtained from various sources, under different experimental conditions and platforms, data quality differs across experiments. Since the ReMap 2015 release, our ChIP-seq pipeline assesses the quality of all

datasets, unlike similar databases (chip-atlas http://chip-atlas.org, (6,7)), (Supplementary Table S4). We compute a score based on the cross-correlation and the FRiP (fraction of reads in peaks) metrics developed by the ENCODE consortium (13) (Supplementary Figure S1). Descriptions of the ENCODE quality coefficients can be found on the UCSC Genome portal (http://genome.ucsc.edu/ENCODE/qualityMetrics.html). Our pipeline computes the normalized strand cross-correlation coefficient (NSC) as a ratio between the maximal fragment-length cross-correlation value and the background cross-correlation value, and the relative strand cross-correlation coefficient (RSC), as a ratio between the fragment-length cross-correlation and the read-length cross-correlation. The same methods and quality cutoffs were applied as in ReMap 2015 (4). Datasets not passing the QC were not included in the catalog of peaks available for download (http://remap.cisreg.eu).

DNA constraint scores

We provide the conservation profiles at the nucleotide level for each of the 485 TRs present in our catalog. We assessed the DNA constraint for each base pair by considering ±1 kb around the summit of each non-redundant peak (see below). Genomic Evolutionary Rate Profiling scores (GERP) were used to calculate the conservation of each nucleotide in a multi-species alignment (14). The computed GERP scores were obtained from the 24-way amniota vertebrates Pecan (15) multi-species alignment, and extracted from the Ensembl Compara database release v89 (16).

Genome coverage, non-redundant peak sets and CRMs

Genome coverages were computed using the BedTools suite (17) (version 2.17.0) using the 'genomecov' function with the option -max 2 that combines all positions with a depth ≥2 binding locations. Full details of the ReMap 2015 and 2018 genome coverage are available in Supplementary Table S5. ReMap also provides a catalog of discrete, non-redundant binding regions for each TR, a specificity not found in other databases (chip-atlas http://chip-atlas.org, (6,18)). We used BedTools to merge overlapping peaks (with at least 1 bp overlap) identified in different datasets for the same TR. The summit of the resulting peaks was defined as the average position of the summits of the merged peaks. Those peaks made of at least two or more peaks for a given factor are defined as non-redundant peaks. We observed a mean variation of 77 bp between the summits of the non-redundant peaks and the individual peak summits (Supplementary Figure S2). Similarly, to obtain the *cis*-regulatory modules (CRMs) in the genome, overlapping peaks of all TRs were merged using BedTools. Regions bound by several TRs are called CRMs, whereas regions bound by only one TR are labeled as singletons.

Roadmap human epigenome annotations

Two sets of chromatin accessibility data were used to better characterize the ReMap atlas. We employed BedTools for overlap analyses allowing a minimum of 10% overlap. The NIH Roadmap Epigenomics Mapping Consortium (19) data were downloaded from the roadmap data portal (http://egg2.wustl.edu/roadmap). Delineation of DNaseI-accessible regulatory regions were accessed from http://egg2.wustl.edu/roadmap/web_portal/DNase_reg.html#delieation. BED files with coordinates of each region type for each epigenome separately are available for 81 232 promoter regions (1.44% of genome), 2 328 936 putative enhancer regions (12.63% of genome) and 129 960 dyadic promoter/enhancer regions (0.99% of genome). The core 15-state model of chromatin combinatorial interactions between different chromatin marks was downloaded from http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state. Chromatin state definitions and abbreviations are: 1 Active TSS (TssA), 2 Flanking active TSS (TssAFlnk), 3 Transcr. at gene 5′ and 3′(TxFlnk), 4 Strong transcription (Tx), 5 Weak transcription (TxWk), 6 Genic enhancers (EnhG), 7 Enhancers (Enh), 8 ZNF genes + repeats (ZNF/Rpts), 9 Heterochromatin (Het), 10 Bivalent/poised TSS (TssBiv), 11 Flanking bivalent TSS/Enh (BivFlnk), 12 Bivalent enhancer (EnhBiv), 13 Repressed Polycomb (ReprPC), 14 Weak repressed Polycomb (ReprPCWk) and 15 Quiescent/low (Quies).

DATA COLLECTION AND CONTENT

Integration of data sources

The 2018 release of the ReMap database reflects significant advances in the number of binding regions, the number of TFs, transcriptional co-activators, chromatin regulators and overall the total number of datasets integrated in our catalog. We initially selected, processed and analyzed 3180 ChIP-seq datasets against TRs from GEO, AE and ENCODE. To ensure consistency and comparability, all datasets were processed from raw data, through our uniform ChIP-seq workflow that included read filtering, read mapping, peak calling and quality assessment based on ENCODE quality criterions. As the quality of ChIP-seq experiments vary significantly (20,21), we incorporated a critical data quality filtering step in our pipeline—not implemented in other databases (chip-atlas http://chip-atlas.org (6,7,18)). Specifically, we considered four quality metrics, two metrics independent of peak calling for assessing signal-to-noise ratios in a ChIP-seq experiment and two metrics based on peak properties. Following ENCODE ChIP-seq guidelines and practices (13), we used the NSC and the RSC (see 'Materials and Methods' section). Further, we used the FRiP and the number of peaks in the dataset (see 'Materials and Methods' section). After applying our quality filters based on these four ChIP-seq metrics we retained 2829 datasets (89%): 1763 datasets from GEO and ArrayExpress and 1066 from ENCODE (Figure 1A and Supplementary Figure S1). The significant increase of data is spread across almost all TFs when compared to ReMap 2015 (Figure 1B). Nevertheless, we observe TFs (e.g. AR, ESR1, FOXA1) and CRFs (e.g. BRD4, EZH2) displaying a larger data growth than other DNA-binding proteins. The majority of TRs show additional datasets integrated in ReMap 2018 (Figure 1B, dark blue bars).

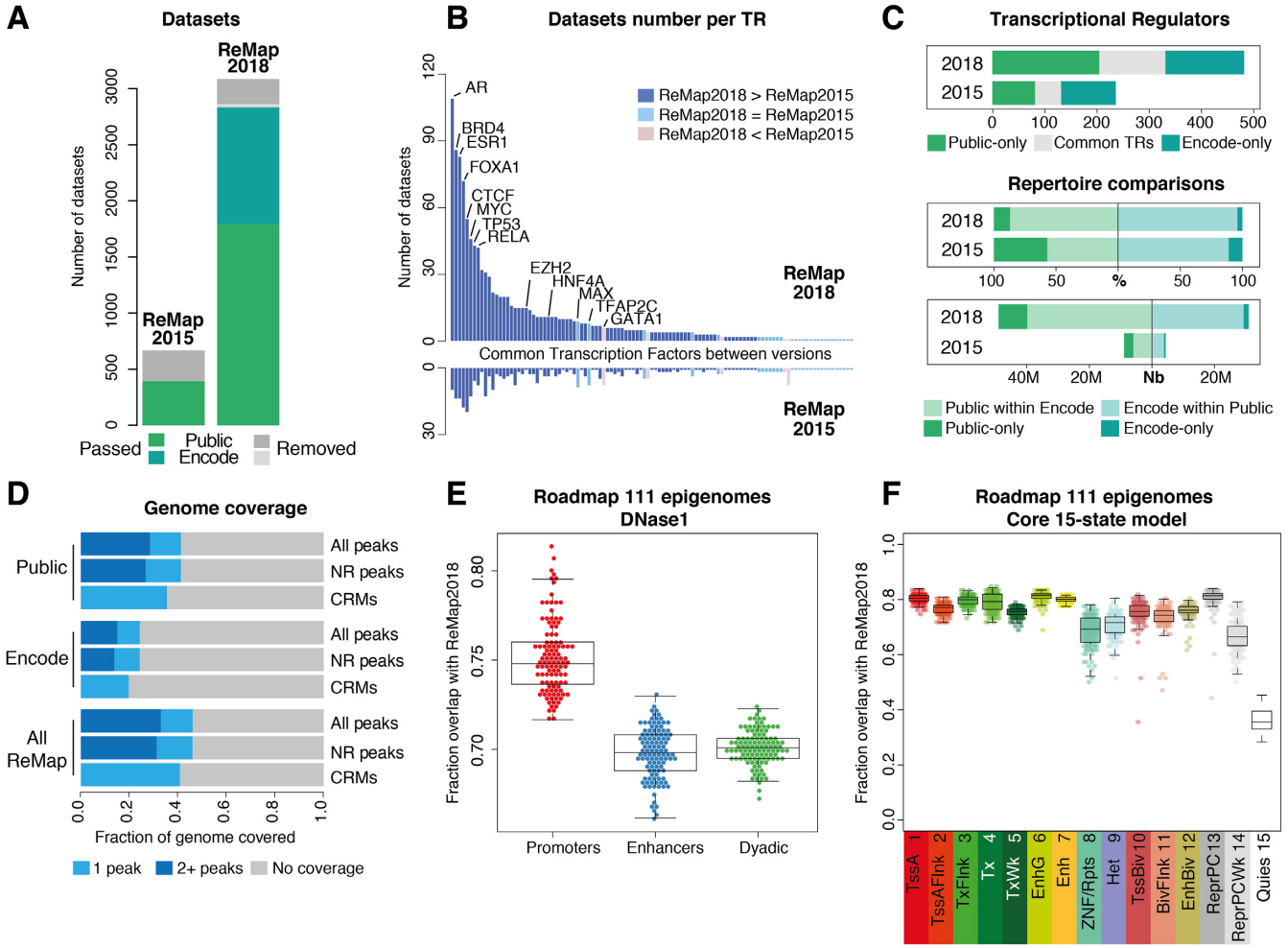


Figure 1. Overview of the ReMap database expansion. (**A**) Analyzed datasets growth in ReMap 2018 compared to ReMap 2015. (**B**) Evolution of the number of datasets per TRs, ranked across common between both ReMap versions. (**C**) Common TRs between Public and ENCODE sets of data (gray). Direct comparison of Public and ENCODE repertoire, defined as percentages (%), and as number (Nb) of peaks. (**D**) Genome coverage fraction of each ReMap dataset (NR non-redundant, CRM Cis Regulatory Modules). (**E**) Comparison of DNase I-accessible regulatory regions against the ReMap 2018, regions from the Roadmap Epigenomics Consortium defining promoter-only, enhancer-only or enhancer–promoter alternating states (Dyadic). Each dot represents the fraction overlap with ReMap 2018 for one of the 111 epigenomes. (**F**) Comparison of the Roadmap Epigenomics Consortium chromatin states annotations against the ReMap 2018 catalog, using the Core 15 chromatin states model, and a minimum overlap of 50% between regions. Each dot represents the overlap for one of the 111 epigenomes. Chromatin state definitions and abbreviations are as follows; 1 Active TSS (TssA), 2 Flanking active TSS (TssAFlnk), 3 Transcr. at gene 5′ and 3′(TxFlnk), 4 Strong transcription (Tx), 5 Weak transcription (TxWk), 6 Genic enhancers (EnhG), 7 Enhancers (Enh), 8 ZNF genes + repeats (ZNF/Rpts), 9 Heterochromatin (Het), 10 Bivalent/poised TSS (TssBiv), 11 Flanking bivalent TSS/Enh (BivFlnk), 12 Bivalent enhancer (EnhBiv), 13 Repressed Polycomb (ReprPC), 14 Weak repressed Polycomb (ReprPCWk), 15 Quiescent/low (Quies).

Regulatory catalog expansion

With all ChIP-seq data uniformly processed, the ReMap 2018 catalog displays ENCODE data down to the cell line and dataset level rather than the simpler multi-cell analysis provided by ENCODE DCC used in ReMap 2015. Our analyses produced 48 693 300 peaks for the Public-only (non-ENCODE) set across 331 TRs and 31 436 124 peaks for the ENCODE set across 279 TRs, leading to a final ReMap regulatory atlas of 80 129 705 peaks generated from 485 TRs (Figure 1C). We found 125 TRs common to the two sets, 154 proteins specific to ENCODE and 206 specific to the Public catalog (Figure 1C). We also found that 839 400 CRMs are shared between both catalogs. Taken separately, the ENCODE peaks overlaps by 96% the Public regions, and 87% of the Public peaks overlap ENCODE regions (Figure 1C). It suggests that merging both Public and ENCODE sets complements the annotation of DNA-bound regions, as it increases the number of regulatory regions in our atlas, hence improving the annotation of DNA-bound elements in the human genome (Figures 1C and 2).

Indeed, about 13% (405 Mb) of the human genome is covered by at least one feature only from the entire ReMap catalog and 33% (1.02 Gb) are covered by two or more features (Figure 1D and Supplementary Table S4). The Public-only and ENCODE-only sets cover the genome by two or more peaks by 28 and 15% respectively. The observed differences can be explained by the wide spectrum of cell lines and treatments included in the Public set (300 cell lines) compared to the ENCODE set (86 cell lines). As a comparison, the

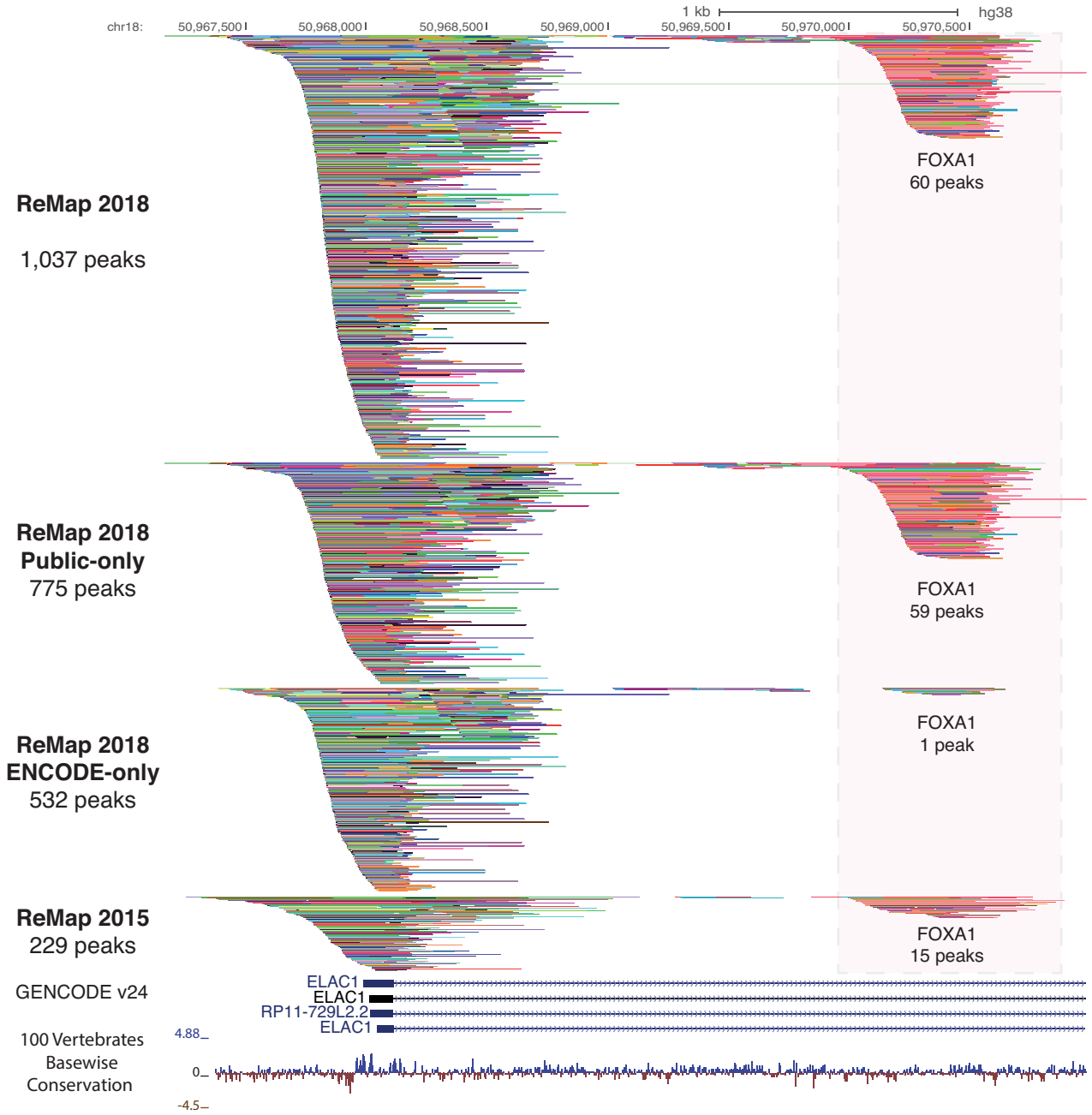


Figure 2. ReMap ChIP-seq binding pattern of 2829 datasets. A genome browser example of the ChIP-seq binding peak depth of the ReMap 2018 catalog compared to ReMap 2015 at the vicinity of the ELAC1 promoter (chr18:50,967,094-50,970,983). The tracks and peaks displayed are compacted to thin lines so the depth of ReMap 2018 bindings can be compared to ReMap 2015. A full and un-compacted screenshot is available as Supplementary Figures S2 and 3. On this location the ReMap 2018 catalog contains 1307 peaks, whereas the ReMap 2015 contains 229 peaks (ReMap 2015 lifted to GRCh38/hg38 assembly). The following genome tracks correspond to the GENCODE v24 Comprehensive Transcript Set and the 100 vertebrates base-wise conservation showing sites predicted to be conserved (positive scores in blue), and sites predicted to be fast-evolving (negative scores in red). A detailed view of the redundant peaks for a FOXA1 site is available in Figure 3.

ReMap 2015 catalog covered 10% (321 Mb) of the genome with one feature only, and 15% (471 Mb) with at least two or more features. Between the two ReMap versions, we observe that the fraction of the human genome covered by one feature remains extremely stable (+84 Mb from 2015 to 2018), whereas the fraction covered by two or more regulatory features increases by 545 Mb. With ReMap 2018, we increase the range of the regulatory space, and provide binding regions for similar TRs at a greater depth, revealing tight and dense co-localization sites (Figures 2 and 3).

Overlap with *cis*-regulatory genomic regions

Using the NIH Roadmap 111 epigenomes analyses, we asked whether the DNase I defined regions as well as the core 15 chromatin states model would better characterize

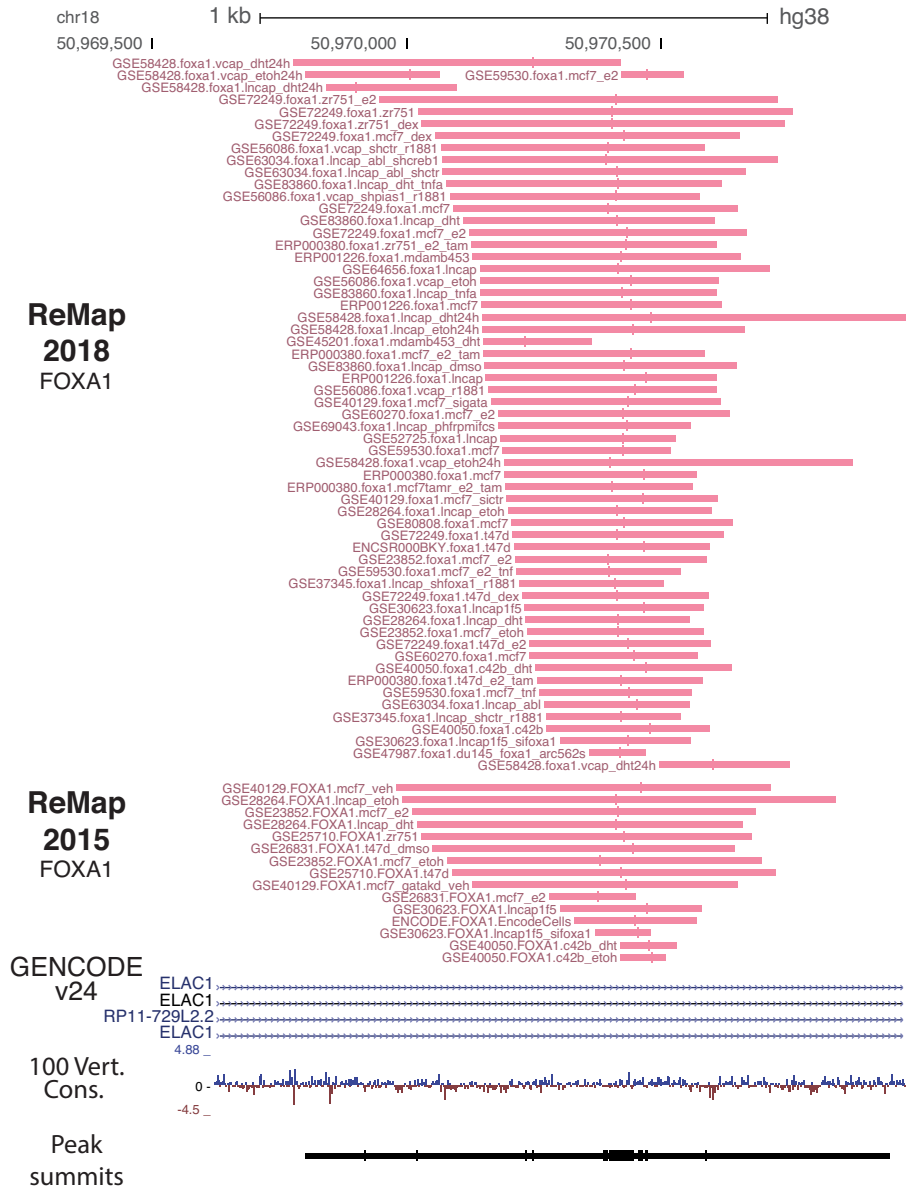


Figure 3. FOXA1 ChIP-seq peaks pattern evolution across ReMap versions. Detailed view of the FOXA1 peaks present in ReMap 2018 (60 peaks) compared to the FOXA1 peaks in ReMap 2015 (15 peaks) found at the genomic location chr18:50,969,638-50,970,931 in the first intron of the ELAC1 gene. Those 60 FOXA1 peaks are derived from GEO, ArrayExpress and ENCODE ChIP-seq across multiple cell lines. Interestingly, it can be noted that the peak summits (vertical bars) of each peak aggregate closely from each other, defining precisely the DNA binding location. Those aggregations of the FOXA1 summits are an illustration of what is globally observed for peaks of different TFs across the genome.

the ReMap atlas (Figure 1E and F). The Roadmap consortium defined a total of 3.5M DNase I-accessible regulatory regions by merging all DNase I hypersensitive regions across epigenomes, which were then annotated using the core 15-state model focusing on chromatin states for promoters, enhancers and dyadic (promoter + enhancer) ambiguous regions (see 'Materials and Methods' section). Among these three categories, the ReMap atlas could recapitulate on average 75.2% of the Roadmap promoter regions, 69.8% of enhancer regions and 70.1% of dyadic regions from the Roadmap annotation. Looking at the core 15-state model, we observe that the ReMap catalog recapitulates more than 70% of the regions covered by each state (Enhancer Genic (81%), Enhancer (80%) and TSS active (80%) states) with the exception of quiesent state (36%). Taken together, these results suggest that some promoter and enhancer activities from Roadmap may be cell type specific, as about 20–30% of those regions seem specific to Roadmap consortium cells. The ReMap initiative results from a large-scale integration of hundreds of diverse cell types, and leads to a regulatory landscape illustrating the large regulatory circuitry of those cells. The constant integration of novel data will allow for a greater definition of the regulatory space across the genome.

Large regulatory atlas

The ReMap database provides a large view of a unique regulatory landscape constituted by 80M binding regions forming 1.6M CRMs. The genomic organization of our occupancy map reveals dense co-localizations of sites forming tight clusters of heterogeneous binding sites with variable TRs complexity (Figure 2). For instance, the regulatory regions observed in the vicinity of the ELAC1 promoter illustrate the ReMap 2018 expansion ($n = 1037$ peaks). It highlights how the regulatory repertoire can be complemented by merging both Public and ENCODE sources. We observe a large cluster of peaks at the ELAC1 promoter followed by two clusters at +500 bp and 1 kb from the transcription start site. The third cluster exemplifies how integrating data from different sources improves genome annotations, as few peaks are available from ENCODE at this location. Additionally, this cluster was detailed in our previous ReMap publication (4) and consisted of 15 FOXA1 ChIP-seq peaks from different cells, antibodies, and laboratories (Figure 3). In this update, we consolidate this FOXA1 binding location with 60 peaks. The summit of each peak is represented by vertical bars aggregated closely from each other, providing an information about the putative location of the DNA binding site. The clustering of FOXA1 peaks and summits illustrates our genome-wide repertoire. However, this FOXA1 example shows overlapping sites derived from various experimental conditions, and therefore does not reflect the total number of discrete binding regions across the genome. To address redundancy between datasets, we merged binding regions for the same TR, resulting in a catalog of 35.5M peaks for all TRs combined. These merged peaks, defined as non-redundant peaks, are made of at least two or more peaks and singletons for a given factor across all experiments, and are available for download from the ReMap website. The TRs with the most merged binding regions across cell types are AR, FOXA1, CTCF and ESR1 (Supplementary Figure S6). These results indicate that most bindings are shared across different ChIP-seq experiments, either for similar or for different cell types. Overall, our ReMap update provides a unique opportunity to identify complex regulatory architectures containing multiple bound regions. We observe that by adding more cell lines, more experiments and more DNA-binding proteins, we increased the genome regulatory space and its depth (Figure 2), but also refined the current annotations of bound regions (Figure 3).

IMPLEMENTATION AND PUBLIC ACCESS

Web display

ReMap provides free public access to all data at http://remap.cisreg.eu. The results presented here provide an informative annotation for 80M ChIP-seq peaks coming from public data sources, which are derived from 485 TRs across 346 diverse cell lines. This catalog provides an unparalleled resource for dissecting site-specific TF bindings (e.g. FOXA1 in Figures 2 and 3) or genome-wide binding analyses. The ReMap web interface displays informations about the integrated TRs (description, classification, external references to Ensembl gene IDs, UniProt, RefSeq, WikiGene, JASPAR, FactorBook, TF Encyclopedia and other resources), peaks, and datasets (quality assessment, read mapping and peak calling statistics, conservation score under peaks). The interface provides a simple 'Dynamic Search' available from the TRs, Cell lines and Download pages and is the entry point for users to search for specific data. The search form allows users to narrow their searches based on gene aliases, dataset names or IDs, cell line names or ontology. For example, entering 'Oct' as search term in the 'Dynamic Search' returns three TFs POU2F2, POU2F1, POU5F1 having various 'OCT' aliases. Additionally, one could use the search box in the Cell or Download page to search for specific cell types containing the 'Colo' term for instance, or 'GSE66218' for a precise experiment from the Download page. Moreover, we provide a tool that allows the annotation of genomic regions provided by users. Those regions are compared against the ReMap catalog returning statistical enrichments of TR bindings present within user-provided input regions compared to random expectations. It allows for the study of over-represented TR binding regions.

Browsing and downloading data

Updates made in ReMap 2018 reflect significant improvement in the variety of genome navigation options. As the ReMap 2015 UCSC session was popular, we now provide more data navigation alternatives. The content of the ReMap database can be browsed through four options: (i) across two mirror sites of the UCSC Genome Browser (22) where a public session has been created (Figure 2 and Supplementary Figure S3), (ii) across three Ensembl Genome Browser mirrors (16) (Supplementary Figure S4), (iii) using the ReMap public track hub (23) or (iv) using the IGV data server (24) (Supplementary Figure S5). For each option, we provide four tracks, the full ReMap catalog containing all peaks, the Public-only peaks, the ENCODE-only peaks and a track containing only peaks above 1.5 kb. As the ReMap catalog expanded, it is crucial to allow visual exploration of regulatory regions across different platforms combined with public or user-specific genome-wide annotations. In addition, the entire ReMap 2018 catalog, as well as the Public-specific or ENCODE-specific peaks, have been compiled into BED files allowing further interpretations and computational analyses.

FUTURE DIRECTIONS

Next-generation sequencing technologies are playing a key role in improving our understanding of regulatory genomics. As ChIP-seq technology is applied to a broader set of cell lines, tissues and conditions, we will continuously maintain and update the database. In the near future, we propose on adding to the ReMap portfolio different peak-caller analyses to further consolidate the peak repertoire. Also, we aim to provide direct access to aligned reads through a FTP server, allowing users to upload and navigate aligned raw data of their choice. We plan on releasing a Bioconductor R-package for genomic region enrichment analyses for large genomic catalogs such as ReMap, which will be replacing our current web enrichment tool. In

the coming year, we would like to provide a Bioconductor R-package to search and download ReMap data for a specific study, to get genomic range objects, raw counts and/or metadata used for a specific study. Overall, determining the best approach to curate and annotate ChIP-seq data with a very broad level of submitted annotations and metadata into a simple-to-use, easy-to-analyze and up-to-date system will become a focus for the ReMap project.

CONCLUSION

The 2018 release of ReMap maintains the long-term focus of providing the research community with the largest catalog of high-quality regulatory regions by integrating all available ChIP-seq data from DNA-binding assays. The usefulness of ReMap is exemplified by the last release of the JASPAR database (25), for which ReMap ChIP-seq peaks were used to derive 45 new TF binding profiles that were incorporated in the 2018 release of the vertebrate CORE collection (Khan *et al*. 2018), providing a 9% increase from JASPAR 2016 (26) by solely relying on the ReMap 2018 catalog. Although new datasets are constantly added to repositories, we believe that our ReMap atlas will help in better understanding the regulation processes in human. In this update, we have (i) widely expanded the collection of datasets curated and analyzed from public sources with now 485 TFs, transcriptional co-activators and chromatin regulators; (ii) uniformly processed and integrated the ENCODE ChIP-seq data; (iii) enhanced the website usability by allowing dynamic search of TRs, aliases, cell lines and experiments, (iv) expanded the genome browsing experience by integrating ReMap in all UCSC and Ensembl Genome Browsers mirror sites and provided a Track Hub for data integration in other platforms; (v) improved the capacity to download all ReMap files in bulk or individually.

AVAILABILITY

ReMap 2018 can be accessed through a web interface at http://remap.cisreg.eu. Downloads are available in BED format for the entire ReMap catalog, the Public-only peaks, the ENCODE-only peaks, and in FASTA and BED formats for each TR. In addition, UCSC and Ensembl Genome Browsers users can navigate ReMap across their mirror sites, use ReMap in UCSC public sessions, or use the public track hub. Finally, Integrative Genome Browser (IGV) users have the option of loading an IGV optimized dataset directly in the application.

FEEDBACK

The ReMap team welcomes your feedback on the catalog, use of the website and use of the downloadable files. Please contact us at benoit.ballester@inserm.fr or remap@cisreg.eu for development requests. We thank our users for their feedback to make ReMap useful for the community.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

French Ministry of Higher Education and Research (MESR) PhD Fellowship (to J.C.); Norwegian Research Council (to A.M., M.G.); Helse Sør-Øst (to A.M., M.G.); University of Oslo (to A.M., M.G.). Funding for open access charge: Institut national de la santé et de la recherche médicale (INSERM).

Conflict of interest statement. None declared.

REFERENCES

1. Mardis,E.R. (2007) ChIP-seq: welcome to the new frontier. *Nat. Methods*, **4**, 613–614.
2. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
3. Kolesnikov,N., Hastings,E., Keays,M., Melnichuk,O., Tang,Y.A., Williams,E., Dylag,M., Kurbatova,N., Brandizi,M., Burdett,T. *et al.* (2015) ArrayExpress update–simplifying data submissions. *Nucleic Acids Res.*, **43**, D1113–D1116.
4. Griffon,A., Barbier,Q., Dalino,J., van Helden,J., Spicuglia,S. and Ballester,B. (2015) Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Res.*, **43**, e27.
5. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
6. Zhou,K.-R., Liu,S., Sun,W.-J., Zheng,L.-L., Zhou,H., Yang,J.-H. and Qu,L.-H. (2017) ChIPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data. *Nucleic Acids Res.*, **45**, D43–D50.
7. Yevshin,I., Sharipov,R., Valeev,T., Kel,A. and Kolpakov,F. (2017) GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res.*, **45**, D61–D67.
8. Yates,B., Braschi,B., Gray,K.A., Seal,R.L., Tweedie,S. and Bruford,E.A. (2017) Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.*, **45**, D619–D625.
9. Gremse,M., Chang,A., Schomburg,I., Grote,A., Scheer,M., Ebeling,C. and Schomburg,D. (2011) The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.*, **39**, D507–D513.
10. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
11. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
12. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
13. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
14. Davydov,E.V., Goode,D.L., Sirota,M., Cooper,G.M., Sidow,A. and Batzoglou,S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**, e1001025.
15. Paten,B., Herrero,J., Beal,K., Fitzgerald,S. and Birney,E. (2008) Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.*, **18**, 1814–1828.
16. Aken,B.L., Achuthan,P., Akanni,W., Amode,M.R., Bernsdorff,F., Bhai,J., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P. *et al.* (2017) Ensembl 2017. *Nucleic Acids Res.*, **45**, D635–D642.
17. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
18. Mei,S., Qin,Q., Wu,Q., Sun,H., Zheng,R., Zang,C., Zhu,M., Wu,J., Shi,X., Taing,L. *et al.* (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.*, **45**, D658–D662.

19. Roadmap Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
20. Mendoza-Parra,M.-A., Saleem,M.-A.M., Blum,M., Cholley,P.-E. and Gronemeyer,H. (2016) NGS-QC generator: a quality control system for ChIP-Seq and related deep sequencing-generated datasets. *Methods Mol. Biol.*, **1418**, 243–265.
21. Marinov,G.K., Kundaje,A., Park,P.J. and Wold,B.J. (2014) Large-scale quality analysis of published ChIP-seq data. *G3 (Bethesda)*, **4**, 209–223.
22. Tyner,C., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Eisenhart,C., Fischer,C.M., Gibson,D., Gonzalez,J.N., Guruvadoo,L. *et al.* (2017) The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.*, **45**, D626–D634.
23. Raney,B.J., Dreszer,T.R., Barber,G.P., Clawson,H., Fujita,P.A., Wang,T., Nguyen,N., Paten,B., Zweig,A.S., Karolchik,D. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.
24. Thorvaldsdóttir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.
25. Khan,A., Fornes,O., Stigliani,A., Gheorghe,F.N., Castro-Mondragon,J.A., van der Lee,R., Bessy,A., Chèneby,J., Kulkarni,S.R., Tan,G. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles andits web framework. *Nucleic Acids Res.*, doi:10.1093/nar/gkx1126.
26. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.Y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.

SUPPLEMENTAL INFORMATION:

ReMap 2018: An updated regulatory region atlas from an integrative analysis of DNA-binding ChIP-seq experiments

Supplementary Tables

Table S1 Overview of the content growth in ReMap 2018 compared to ReMap 2015.

	ReMap 2015	ReMap 2018
Total TRs	237	485
Public TRs	132	331
ENCODE TRs	155	279
Common TRs	50	125
Datasets (selected / processed)	396 / 668	2,829 / 3,180
Total peaks	13 million	80 million
Merged peaks	8.8 million	35.5 million

Table S2 Rejected datasets in ENCODE, MACS analyses.

Rejected ENCODE datasets-MACS	
ENCSR000DNO_GM12878_KAT2A	ENCSR038RXU_MCF-7_GABPA
ENCSR000EHO_K562_SMARCA4	ENCSR101FJM_HCT116_ZNF274
ENCSR000EUI_GM08714_ZNF274	ENCSR178NTX_K562_CUX1
ENCSR000EVR_HepG2_ZNF274	ENCSR219NRT_HepG2_GTF2F1
ENCSR000EWE_K562_ZNF274	ENCSR265WJC_MCF-7_KLF4
ENCSR000EWH_K562_NR2C2	ENCSR272JAT_K562_CBX5
ENCSR000EXH_SUZ12_NT2D1	ENCSR448UKK_K562_ZKSCAN8
ENCSR000EXR_erythroblast_GATA1	ENCSR475SOC_MCF-7_ELF1
ENCSR000FAC_K562_XRCC4	ENCSR505NMN_MCF-7_E2F4
ENCSR000FCI_HEK293_ZNF274	ENCSR511CUH_neural_cell_EZH2
	ENCSR532KTI_K562_GTF2E2

Table S3 Rejected datasets in Public, MACS analyses.

In Excel spreadsheet

Table S4 Evaluation of Databases of ChIP-seq analyses.

We evaluated ChIP-seq databases similar to ReMap from a user point of view, with the comparison table 1 from GTRD as a starting point (Yevshin et al. 2017).

Database	**ChIPBase** http://rna.sysu.edu.cn/chipbase/	**ReMap** http://remap.cisreg.eu	**Cistrome** http://dc2.cistrome.org/#/	**ChIP-Atlas** http://chip-atlas.org/	**GTRD** http://gtrd.biouml.org/
Release dates	2013, 2017	2015, 2018	2017	none	2017
Human assemblies	hg38?	hg38, hg19	hg38	hg19	hg38
Human datasets	3549	3157	5774	6869	5078
Sources	GEO, ENCODE	GEO, ARRAYEXPRESS, ENCODE	GEO, ENCODE, SRA, ENA	SRA	GEO, SRA, ENCODE
QC filtering	no	yes	yes	no	no
Manual annotation	no	yes	no	no	no
Human TFs	480 TFs	486 TFs and non-TFs	260 TFs and non-TFs	not stated	476 TFs
Peak callers	Multiple (import)	MACS2	MACS2	MACS2	MACS, SISSRs, GEM, PICS
Atlas provided	no	yes	no	no	yes
Non-Redundant analysis	no	yes	no	no	metacluster
Clustering analyses	no	CRMs	no	no	metacluster
Uniform data processing	no (import)	yes	yes	yes	yes
Displays, Genome browser	Self-developed	UCSC, ENSEMBL, Track Hub, IGV	UCSC	IGV	Self-developed
Annotation tool	no	yes	no	no	no
Standard format	no	BED	BED	BED	no
Bulk Download	no	yes	yes	yes	yes

Table S5 Genome Coverage of ReMap 2015 and ReMap 2018.

For each version of ReMap, the coverage is computed for regulatory regions overlapping the genome at a depth of only 1 feature, or 2 or more features.

	ReMap 2015 (lifted hg38)			**ReMap 2018**		
	Depth 1	Depth 2+	Total	Depth 1	Depth 2+	Total
Percent	10.4	15.2	25.6	13.1	32.9	46
Bases (Mb)	321.4	471.4	792.8	405.1	1016.6	1421.6

Table S6 List of ReMap 2018 datasets

In Excel spreadsheet

Table S7 Modified name of transcriptional regulator

In Excel spreadsheet

Supplementary Figures

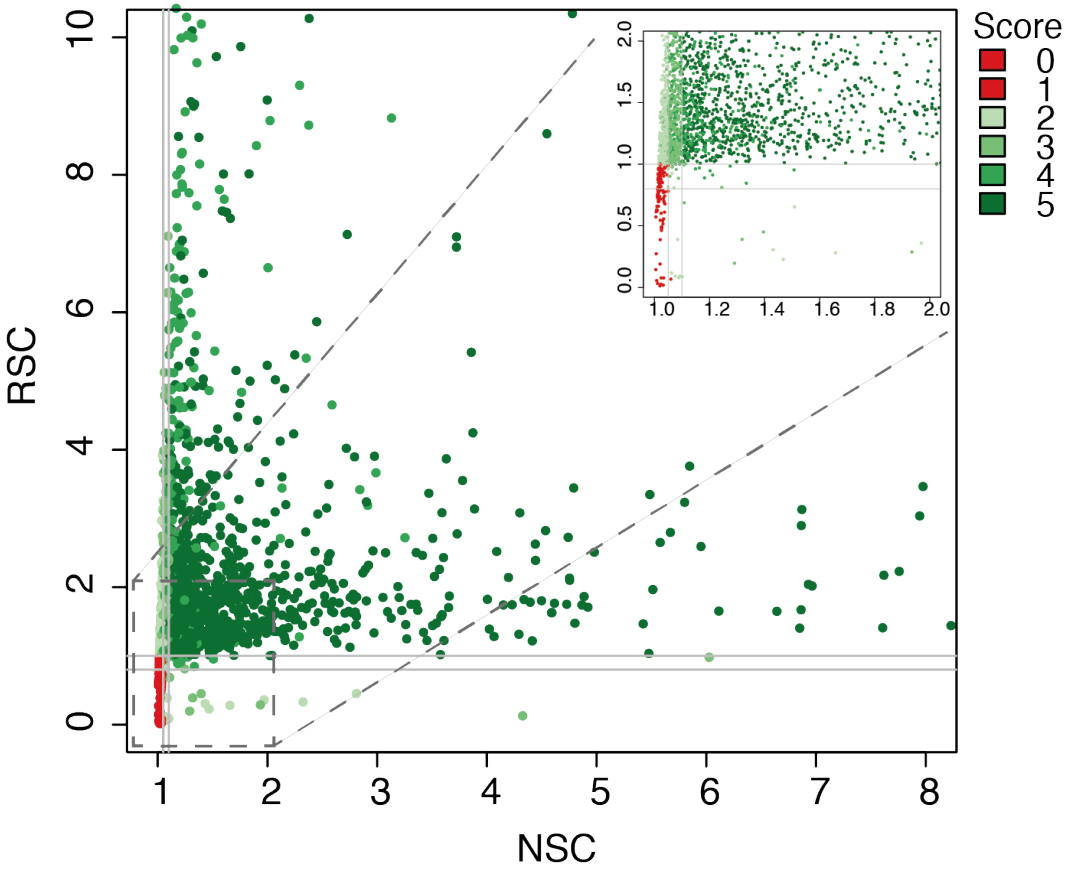


Figure S1 Quality assessment of 3,180 Public and ENCODE datasets.

Here each analysed dataset (n=3,180) is represented by a colored dot according to its assigned score. Indeed, to assess the quality of public datasets, we computed a score based on the cross-correlation and the FRiP (fraction of reads in peaks) metrics developed by the ENCODE Consortium and the phantompeak tools. This score is computed as follow. Two thresholds based on ENCODE studies were defined for each of the two cross-correlation ratios (Grey lines; Normalized Strand Coefficient: 1.05 and 1.10 on the x-axis; Relative Strand Coefficient: 0.8 and 1.0 on the y-axis, see M&M for details). A basal score ranging from 0 to 4 was assigned to each dataset corresponding to the number of thresholds it exceeds for NSC and RSC (2 thresholds for each score). Finally, this basal score was incremented by one if the FRiP is equal or higher than 1%. We observe that datasets having a minimum score of 2 exceed at least one threshold of RSC or NSC, which are both scores independent of peak calling procedures. Thus, for our analyses, datasets having a score less than or equal to 1 (red dots), as well as datasets with fewer than 100 identified peaks were discarded for further downstream analyses. Reds dots within accepted thresholds are datasets with less than 100 peaks. Datasets with scores greater to 1 were kept for further analyses (green dots, n=2,834).
Score = 0 : the dataset has NSC and RSC outside thresholds
Score = 1 : the dataset has one (RSC or NSC) above 1st threshold
Score = 2 : the dataset has one (RSC or NSC) above optimum threshold, OR two RSC or NSC above 1st threshold
Score = 3 : the dataset has one (RSC or NSC) above optimum threshold, and one RSC or NSC

above 1st threshold
Score = 4 : the dataset has two (RSC + NSC) above optimum threshold
Score = 5 the dataset has two (RSC + NSC) above optimum threshold + 1 extra point if the FRiP is equal or higher that 1% (see our M&M)
In short, we give a score from 0 to 4 according to NSC and RSC, and then add 1 point according to FRiP. Datasets with a score below 2 are discarded.

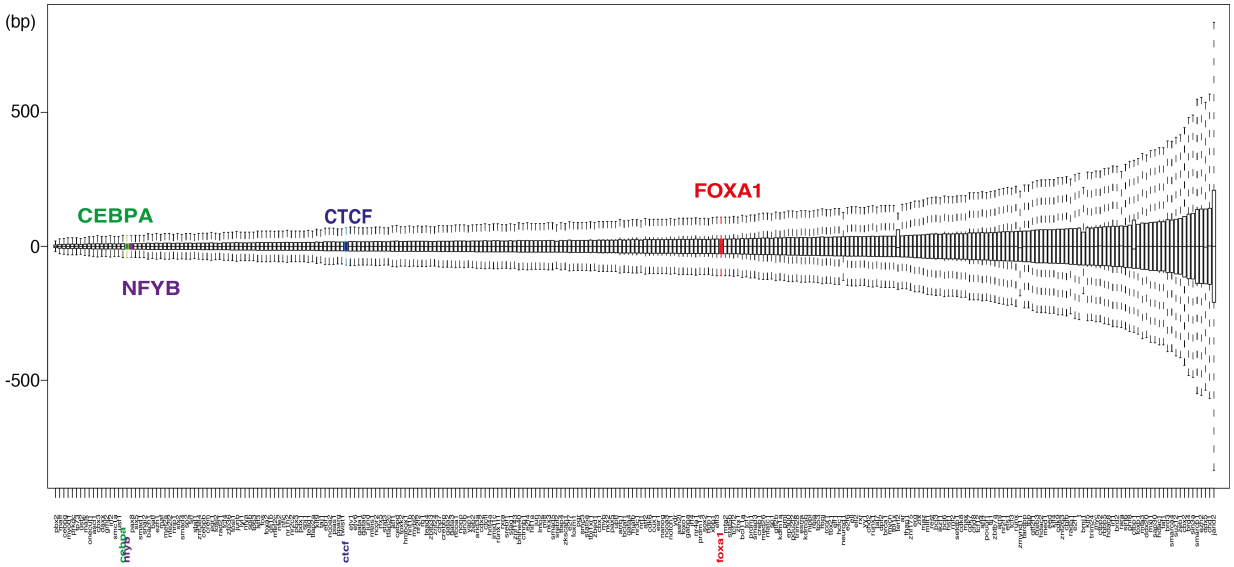


Figure S2 Variation of summits of merged peaks relative to the average summit of non-redundant peaks

Distribution of distances in bp between non-redundant peaks summit and the summits of all peaks they are made of, we selected and highlighted four factors (CEBPA, NFYB, CTCF and FOXA1). This boxplot is showing the summits variation for all factors having a minimum of 10,000 merged peaks.

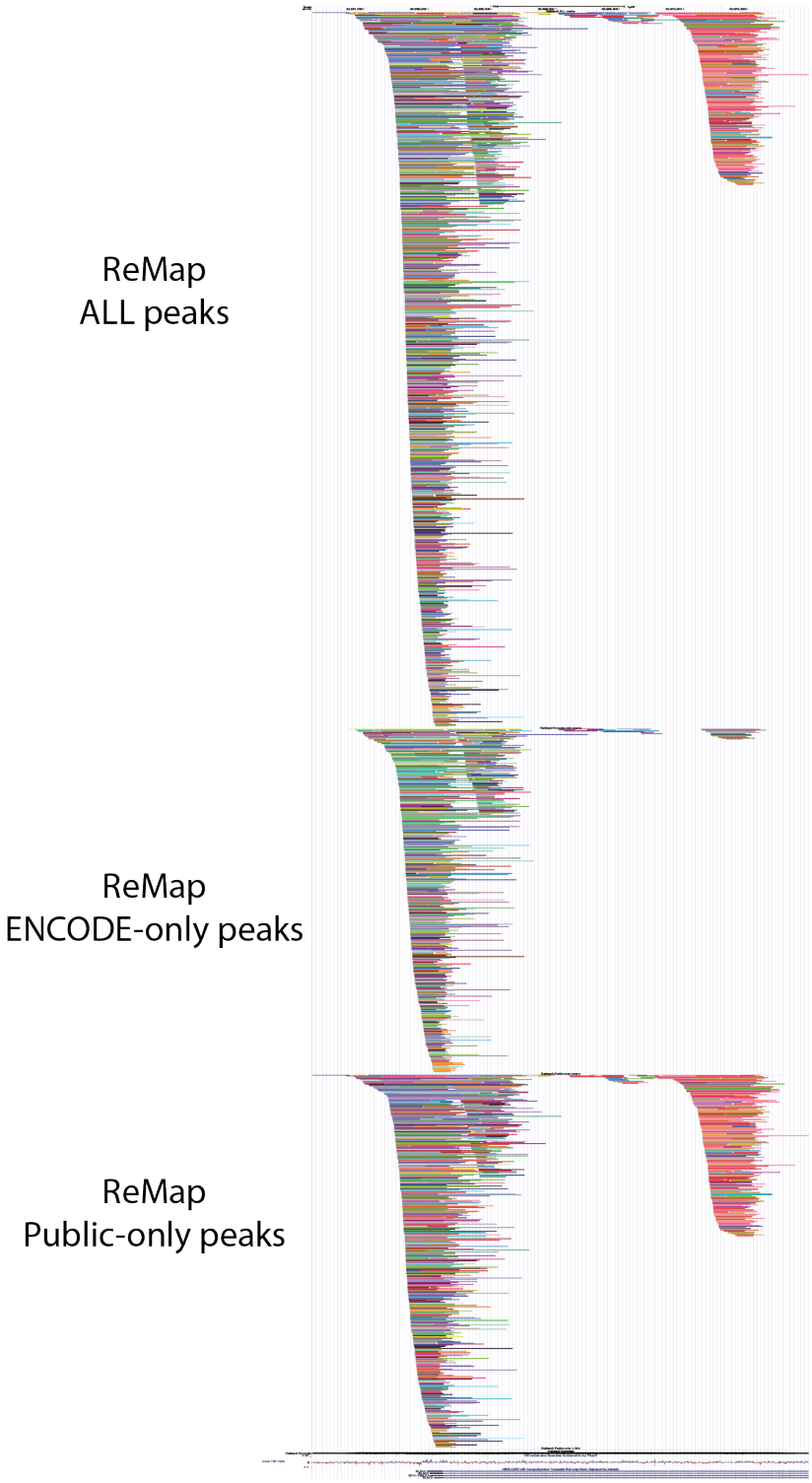


Figure S3 Visualization of the ReMap data on UCSC Genome Browser.

The ReMap catalogue can viewed through the Track hubs on both Ensembl and UCSC Genome Browser. Public sessions have been also created on mirrors of UCSC Genome Browser allowing the display of larger regions and fester rendering of the data.

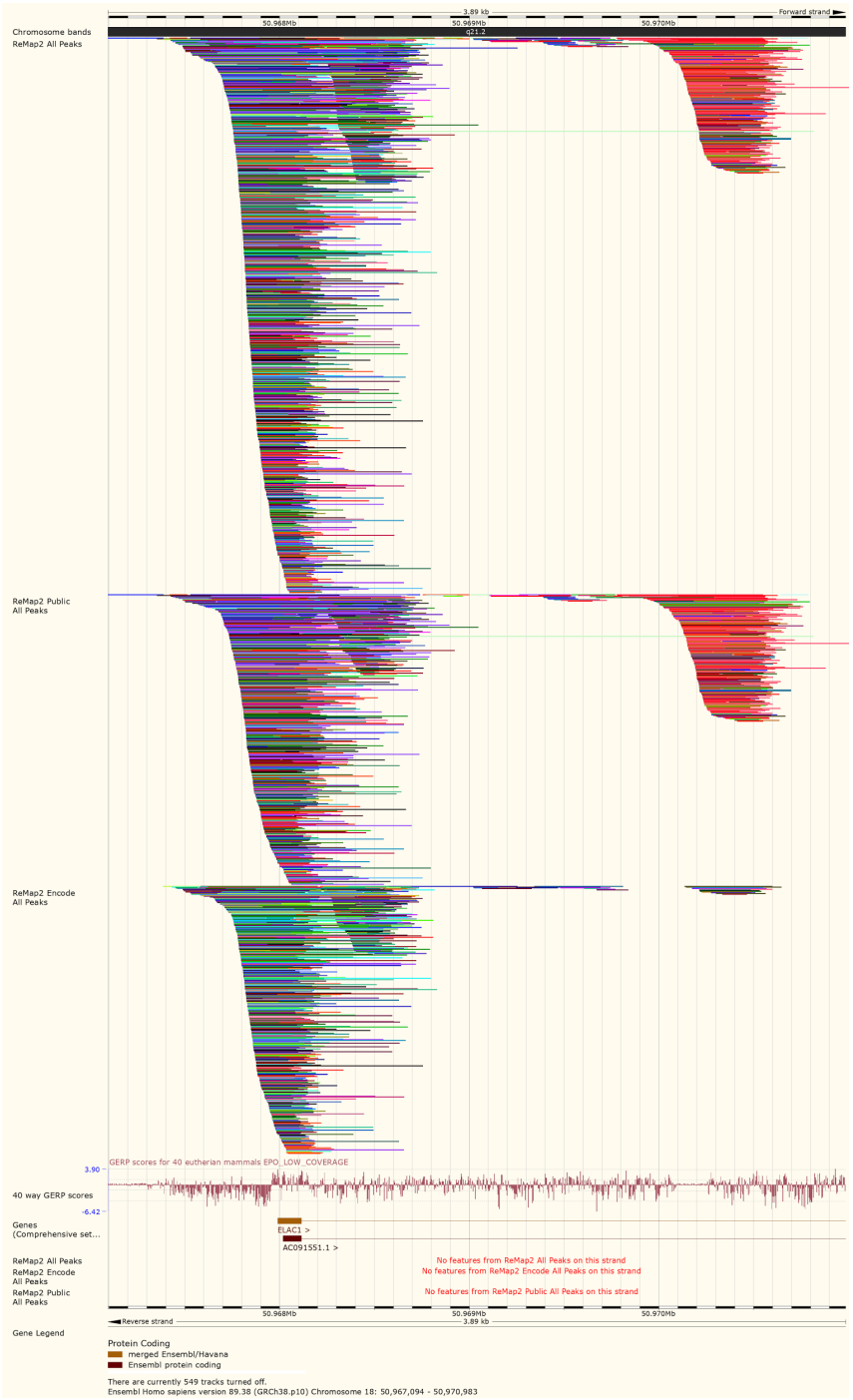


Figure S4 Visualization of the ReMap Track Hub on Ensembl Genome Browser.

The ReMap catalogue can viewed through the a public track hub. Track hubs are web-accessible directories of large collections of genomic data enabling the visualization of genome wide annotations on UCSC and Ensembl Genome Browsers. Here is a screenshot example of the ReMap track hub with the three tracks, the first track correspond the entire ReMap catalogue, the second the Public-only peaks and the third the ENCODE-only peaks. Users may chose to switch on or off any of the track as desired, and or change the rendering of the tracks (Compact, Full, Stacked, etc…). From this example, we can clearly see how merging Public and ENCODE peaks can greatly strengthen the regulatory annotation. The promoter and the first downstream regulatory regions can be both detected equally, but the intronic regulatory region (pink peaks) seems to be specific bindings found from cell lines used in the Public set.

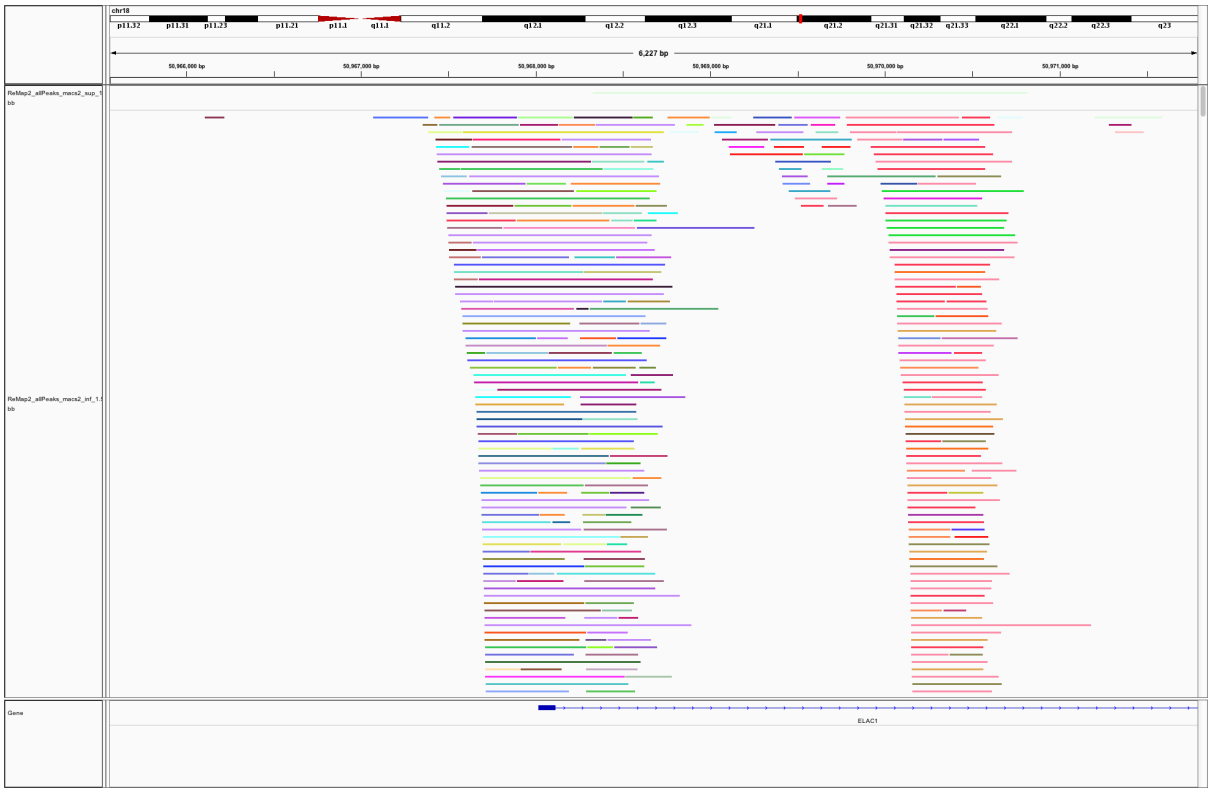


Figure S5 Visualization of the Tracks on IGV with the ReMap session.

The different tracks from the ReMap catalogue can also viewed using the Integrative Genomics Viewer (IGV) available from the Broad Institute, here the chr18:50,967,094-50,970,983 location around the ELAC1 promoter. The URL for the ReMap session is :

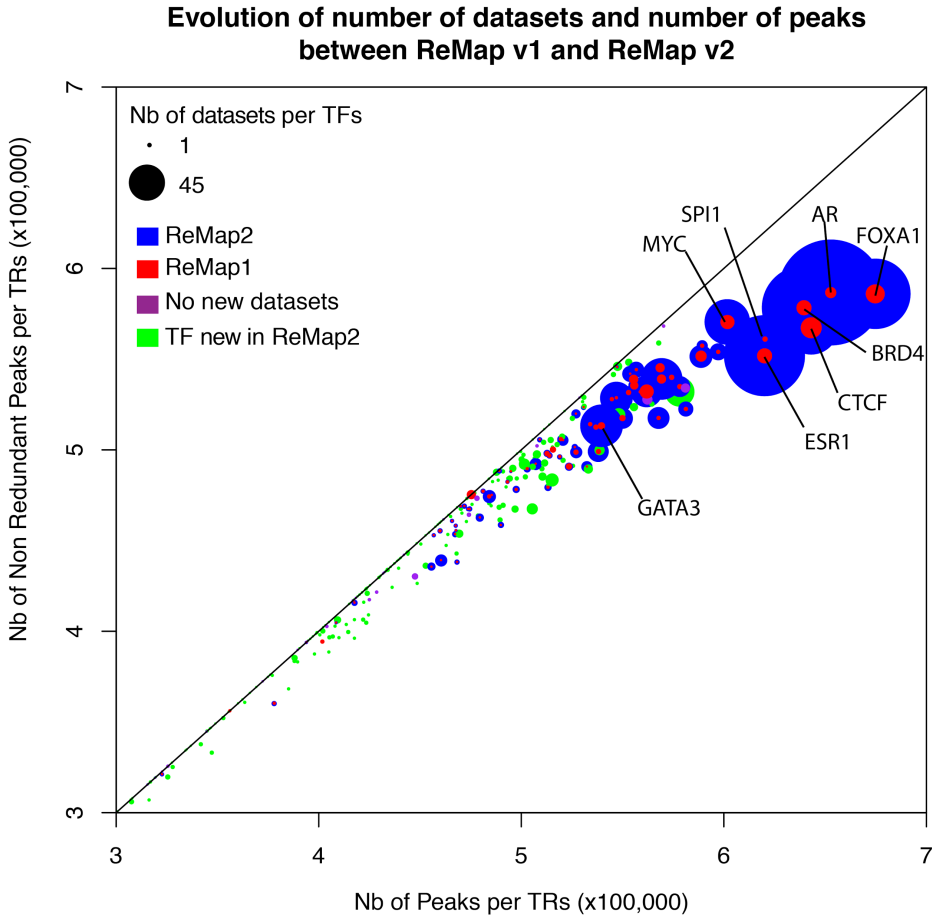


Figure S6 Evolution of peaks and datasets increase between ReMap 2015 and ReMap 2018.

Here are plotted the number (Nb) of peaks for each TRs on the x-axis, the number of merged (non-redundant) peaks on the y-axis. The number of datasets for each TRs are represented by circles of differents sizes.
New TRs in ReMap 2018 are colored in green. Datasets for similar TRs between ReMap 2015 colored in red, and in blue for ReMap 2018.

C. Conclusion

1. Résultats

Pour ReMap2018 nous avons traité 3 180 expériences de ChIP-seq différentes provenant d'ENCODE (1 160) de GEO (1 862) et d'Array Express (158). Après les étapes de contrôle qualité, 2 829 expériences ont été sélectionnées pour constituer le catalogue ReMap2018. Cela représente 89% des expériences traitées. En 2015, seulement 65% des expériences passaient les contrôles de qualité sur la base de 395 expériences traitées. Cette augmentation est peut-être due à la maturation du protocole ChIP-seq. Certaines données de ReMap2015 datent des premières expériences ChIP-seq (2011). Depuis, les méthodes de séquençage à haut débit ont évolué et permettent un séquençage de meilleure qualité (profondeur du séquençage, diminution des erreurs de séquençage, augmentation de la taille des fragments de lecture, etc.).

Le catalogue ReMap2018 contient environ 80 millions (M) de régions de fixation pour 485 protéines régulatrices de la transcription et pour 346 lignées cellulaires et tissus différents. En 2015, ce catalogue contenait 13M de régions concernant 132 protéines régulatrices de la transcription et 83 lignées cellulaires. Entre 2015 et 2018, le nombre d'expériences a été multiplié par 10 et le nombre de protéines et de lignées cellulaires par 4.

Cette nette augmentation de données nous a permis d'annoter 1,6 millions de CREs, ce qui représente 1 million de plus qu'en 2015. Nous sommes passés d'une couverture de 26% du génome à 46%.

L'augmentation des données de fixation des protéines permet, non seulement de découvrir de nouvelles régions régulatrices, mais aussi de consolider les CREs existants. Des pics issus des nouvelles expériences chevauchent des sites de fixation des protéines déjà présents en 2015 comme on peut l'observer sur la figure I.6.

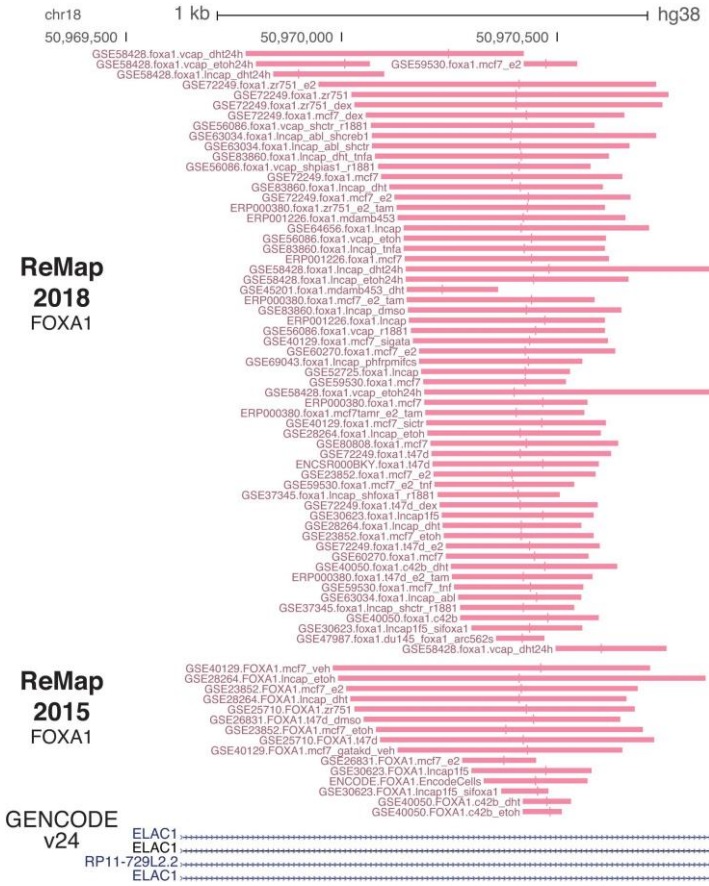


Figure I.6 : Régions génomiques présentant un regroupement de fixations de la protéine FOXA1.
Cette figure compare les pics de fixation issues du fichier BED de ReMap2015 avec ceux de ReMap2018 représentés par le navigateur de génome UCSC. Cette région correspond à un pic non redondant de FOXA1. En 2015, ce pic non redondant était constitué de pics provenant de 15 expériences de ChIP-seq différentes issues de différentes lignées cellulaires. En 2018, ce pic est constitué de données issues de 60 expériences différentes. On peut aussi constater que les sommets de pics représentés par une barre sont regroupés.

a. Autres bases de données de fixation des protéines

D'autres bases de données similaires, retraitant les données ChIP-seq publiques, existent. Les plus connues sont ChIP-Atlas développé au Japon par l'équipe de Chikara Meno[166]. Cistrome DB développé en Chine par l'équipe de X. Shirley Liu[167]. GTRD développé en Russie par l'équipe de Fedor Kolpakov[168]. Ces bases de données contiennent souvent plus de données (environ 10000 expériences pour environ 1200 GTR) car elles utilisent une annotation automatique (sauf ChIP-Atlas). Ce type d'approche permet une automatisation de l'annotation au détriment de la supervision de l'annotation. Il semble extrêmement difficile de se passer de la supervision humaine (cf article cistrome DB).

De plus, il est important que ces bases de données soient accessibles et utilisables par toute la communauté scientifique et pas seulement par les informaticiens. Des bases de données comme GTRD ne fournissent que très peu de métadonnées biologiques ce qui limite leur utilisation par des néophytes.

Tableau I.2 : comparaison des bases de données de fixation des protéines

	ReMap2018	**ChIP-Atlas**	**Cistrom DB**	**GTRD**
Source des données	GEO, ENCODE, Array Express	NCBI SRA contenant GEO, Array Express, DDBJ, ENCODE et Roadmap	GEO, ENCODE, Roadmap	GEO, ENCODE, une partie de SRA
Nombre d'expérience ChIP-seq	2818	6869	13976 (dont histone)	10675
Nombre de TF	485	Information manquante	960 (dont histones)	1158
Contrôle qualité	Oui	Oui	Oui	Non
Annotation CRE	Oui	Non	Non	Oui

2. L'annotation des clusters de fixations des pics

La figure I.7 illustre une des difficultés du Big Data, à savoir, l'hétérogénéité des données et des analyses. Les pics FOXA1 proviennent tous d'expériences ChIP-seq, or, nous observons que la longueur des pics et la position des sommets n'est pas constante. Plusieurs explications à cela, la première est l'utilisation de différents protocoles ChIP-seq. Nous avons ainsi observé que près de 60% des expériences de GEO, soit n'utilisaient pas de contrôle, soit n'utilisaient qu'un seul réplica. Cela limite l'estimation du bruit de fond et les aléas du séquençage. La profondeur de séquençage et la taille des fragments de lecture influent, eux aussi, sur la résolution de l'expérience et la qualité de l'alignement. Le facteur humain est également à prendre en considération. Cette hétérogénéité dans les données ChIP-seq influe sur la définition des pics non redondants et donc dans la précision de l'annotation des éléments régulateurs. Par exemple, certaines protéines, comme BRD4 (bromodomain containing 4), ont des pics d'en moyenne 200 pb. Ce chiffre est principalement dû à l'encombrement stérique (152 219kDa) de la protéine et définit donc des régions de fixation anormalement grandes.

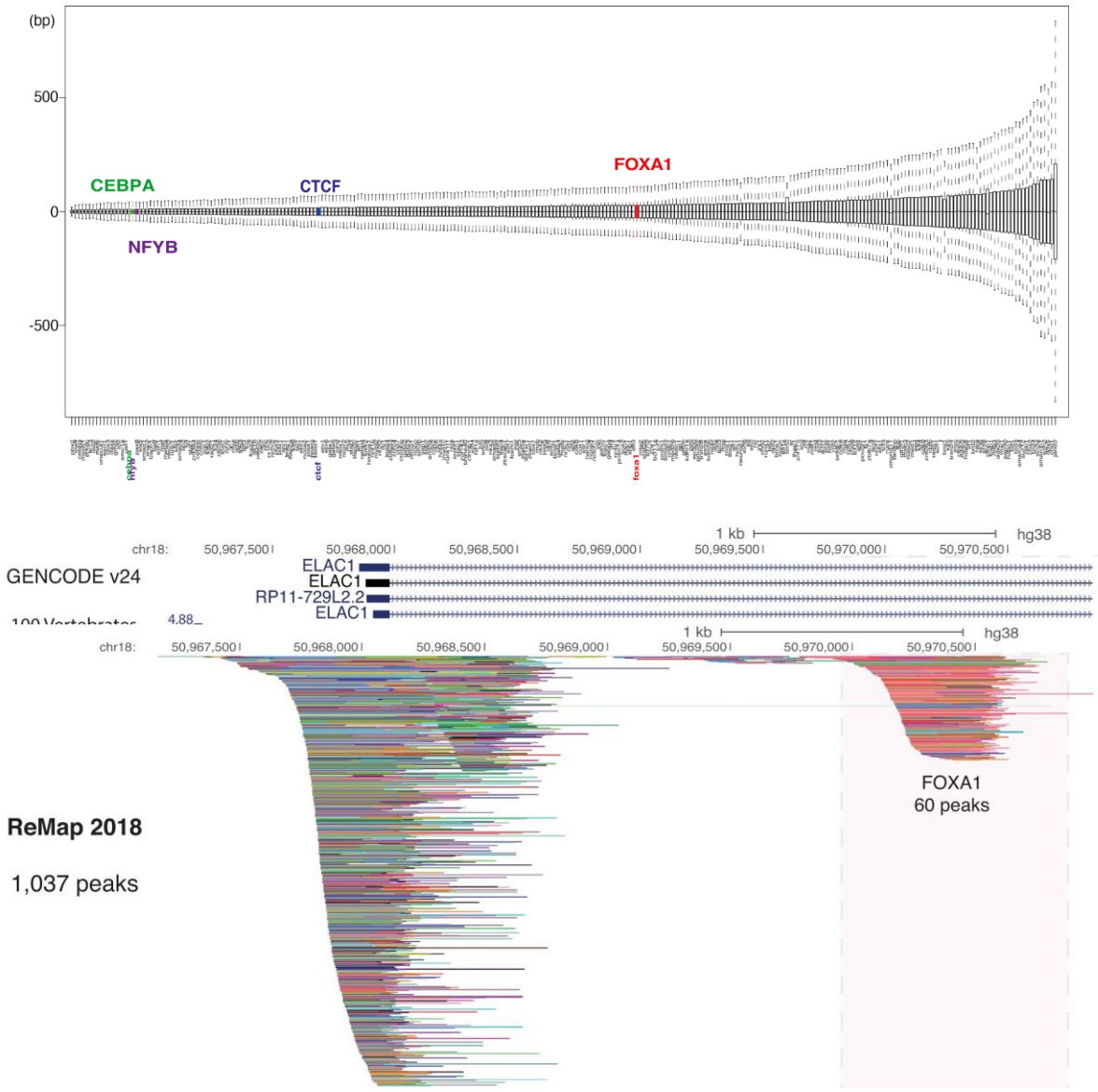


Figure I.7 : Variation de tailles de pics en fonction des protéines et exemples de CREs se chevauchant.

Le premier graphique représente la distribution des tailles des pics issus d'expériences de ChIP-seq en fonction de la protéine concernée. La taille moyenne varie de quelques dizaines de paires de bases (pb) à environ 500 pb. Le deuxième graphique est une capture d'écran du navigateur de génome Ensembl autour du gène ELAC1.

Selon la méthode actuelle de ReMap, la région représentée dans la figure I.7 ne décrit qu'un seul CRE. Or, nous pouvons distinguer 3 clusters de fixation différents. Un au niveau du promoteur, un autre tout de suite après décrivant sans doute un promoteur alternatif et le troisième dans l'intron. On en déduit que des éléments régulateurs de ReMap décrivent plutôt les régions de régulation que les éléments régulateurs uniques. J'ai, par conséquent, développé une nouvelle méthode permettant de distinguer deux éléments régulateurs très proches. Elle résulte de l'observation suivante : les motifs de fixation des TF se trouvent généralement sous le sommet des pics de fixation[169–171]. Elle se base sur la théorie de l'information de Mann-Kendall qui permet de trouver les pics locaux par rapport à l'environnement[172]. Le test de Mann-Kendall est utilisé pour déterminer si une série chronologique a une tendance monotone à la hausse ou à la baisse. Si ce test n'exige pas que les données soient normalement distribuées ou linéaires, en revanche, il exige qu'il n'y ait pas d'autocorrélation. Son intérêt est donc qu'il ne nécessite pas de modèle. J'ai, par conséquent, défini un CRM comme une "série chronologique" (profil) et les données de fixation ReMap comme le cumul des pics en pondérant le sommet (Figure I.9). J'applique ensuite le test de Mann-Kendall avec un seuil à 0.05 et je filtre uniquement pour les pics et non les creux. J'obtiens alors une liste de positions théoriques où se trouvent les centres des éléments régulateurs et des clusters des motifs de fixation.

J'ai testé cette méthode sur plusieurs régions régulatrices du catalogue ReMap et j'ai eu des résultats encourageants (Figure I.8). Malheureusement je n'ai pas eu le temps d'adapter ma méthode à grande échelle. Des contrôles de validité sont nécessaires. Pour un contrôle négatif, il faudrait tester cette méthode avec un remaniement aléatoire des pics. Ce contrôle permettrait de vérifier si les paires de bases définies comme centres d'éléments régulateurs sont bien dues à une concentration de sommets. En revanche le contrôle positif est plus fastidieux à mettre en œuvre. Il n'existe, en effet, aucun catalogue complet d'éléments régulateurs à ce niveau de définition. Une des méthodes envisageables serait de vérifier la concentration des motifs de fixation sous le centre du pic et de vérifier s'il y a enrichissement. La limite étant que les bases de données des motifs de fixation ne sont pas complètes et pourraient manquer de motifs pour certaines régions contenant des TFs peu connus. Une autre limitation est d'ordre technique. J'ai adapté cette méthode en un workflow Snakemake simple en 4 étapes. Or, la création du DAG, schéma des étapes à accomplir sur toutes les entrées, est anormalement longue et a saturé la mémoire du cluster. Si on applique ce workflow pour les 1,6M d'éléments régulateurs du génome, la construction du DAG serait extrêmement coûteuse en ressources. C'est une limite logicielle et matérielle que je n'ai pas su résoudre malgré la puissance de Snakemake.

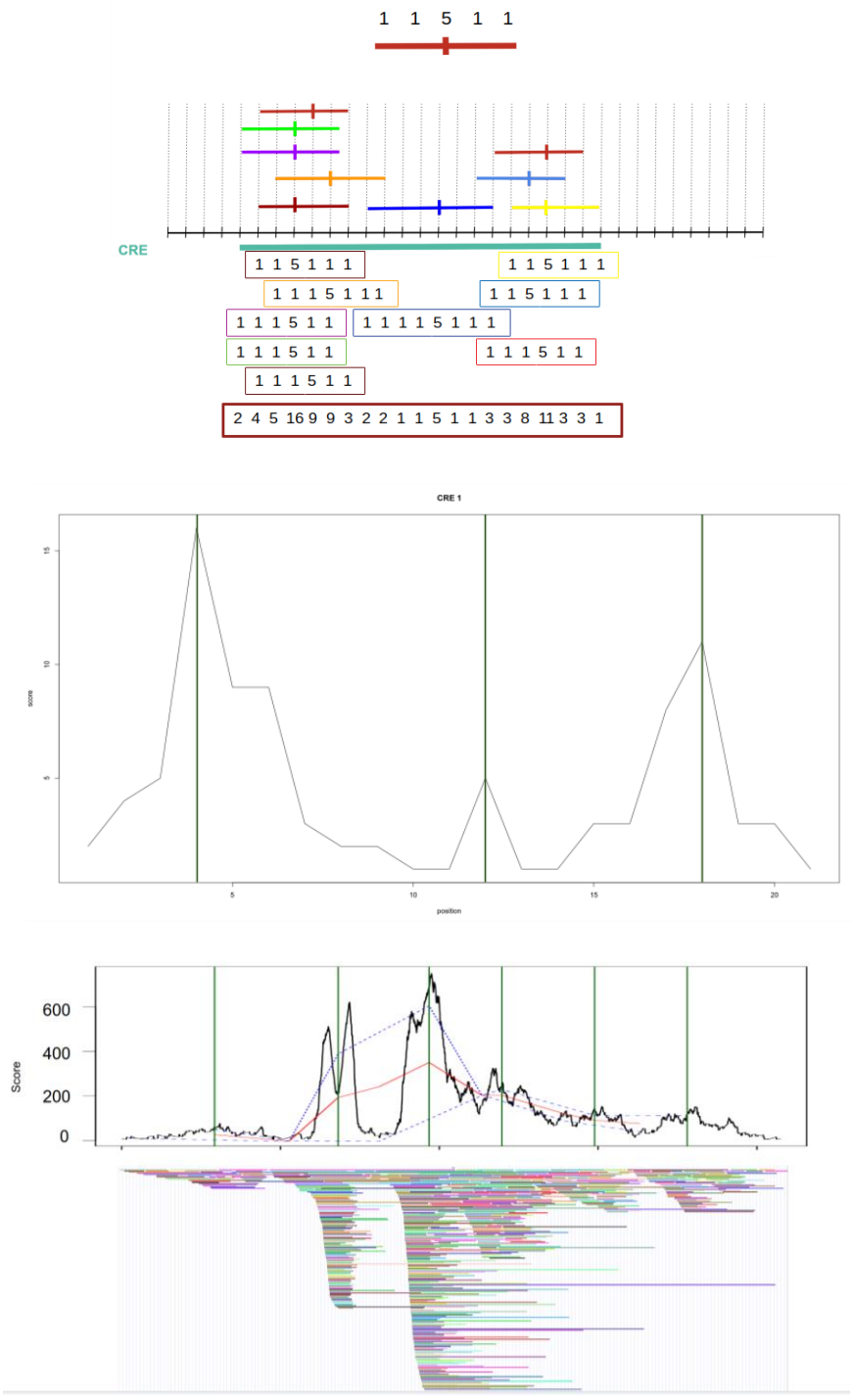


Figure I.8 : Représentation théorique et pratique de la méthode d'annotation des CREs.
Les deux premières images représentent un exemple théorique de la méthode d'annotation des CREs à partir des pics de ChIP-seq. Les pics de fixation des protéines sont convertis en profils centrés sur les sommets des pics. Dans la deuxième image, les lignes vertes représentent les centres des clusters de fixation des pics trouvés par la méthode Kendall.
La troisième image représente un exemple de la méthode appliquée à un CRE de ReMap.

3. Prémices d'un pipeline robuste

Le nouveau workflow développé pour ReMap2018 a permis le traitement de près de 3000 expériences et de 3000 réplicas et contrôles sur une nouvelle architecture. Il est toujours possible de l'améliorer. En effet le workflow de ReMap2018 est composé de 4 sous workflow traitant : les téléchargements, l'alignement des données single-end, l'alignement des données paired-end et du peak-calling. Le lancement de ces sous-workflow est manuel ce qui entraîne des contraintes de synchronisation, il faut s'assurer manuellement que tous les fichiers produits par un des sous workflows sont correctement générés avant de continuer le traitement des données.

Une autre difficulté réside dans le renommage des expériences et des fichiers. Les fichiers sont nommés ainsi : <ID>.<NOM_PROTÉINE>.<NOM_CELLULE> (ex : ENCSR000AHD.CTCF.MCF7, GSE49091.ERG.VCAP) ce qui implique que chaque correction du nom des protéines régulatrices de la transcription ou du nom des cellules induit le changement du nom du fichier. Ces fichiers sont alors considérés comme nouveaux et relancés par Snakemake. Il faut donc renommer tous les fichiers de toutes les étapes du workflow. Pour ReMap2018, ce renommage s'effectue à la main, ce qui entraîne des erreurs. Le renommage de fichiers est une contrainte incontournable de l'annotation manuelle. Il faut donc automatiser le plus possible le processus de renommage, mais aussi de nommage. La décision a donc été prise, pour les prochaines mises à jour de ReMap, de mettre en place un outil facilitant la correction des erreurs et de créer un pipeline de renommage automatique, mais aussi traçable.

II ReMap 2020

A. Introduction

ReMap a été mis à jour en 2019 afin de suivre la constante évolution des données issues d'expériences de fixation de protéines à l'ADN. Cette mise à jour s'appelle ReMap2020. Elle couvre, non seulement l'extension du catalogue d'éléments de régulateurs chez l'homme, mais aussi des données d'une nouvelle espèce : la plante modèle Arabidopsis thaliana (A. thaliana). Pour compléter notre catalogue, nous avons intégré des données ChIP-exo et DAP-seq. Cette mise à jour a permis l'actualisation des technologies et outils de notre pipeline contribuant ainsi à simplifier l'annotation et le traitement des données brutes. Un travail important a également été effectué afin de faciliter l'accès aux données produites par le projet ReMap.

B. Mise à jour du pipeline ReMap

Le workflow utilisé pour ReMap 2018 a mis en évidence la nécessité de planifier la structure d'un projet Big Data. Nous avons mis au point un cahier des charges précis en vue de pérenniser au mieux le travail produit. Il a été décidé :

1. D'utiliser le contrôleur de version GIT pour faciliter le partage des scripts produits ainsi que le suivi de leurs développements.
2. De fusionner, autant que faire se peut, les workflows de traitement des données dans le but de limiter les erreurs humaines.
3. De rendre le workflow portable afin d'utiliser de nouvelles architectures de cluster plus puissantes.

Le projet ReMap possède désormais un GitHub public permettant le partage ainsi que la collaboration à l'amélioration du workflow. Disponible à https://github.com/remap-cisreg/remap-pipeline.

Le pipeline de traitement de ReMap est maintenant constitué de trois workflow distincts correspondant aux dernières étapes du pipeline général de ReMap ; I) le retraitement des données II) le filtre qualité III) l'annotation des régions régulatrices. Les sous workflows du pipeline de traitement de données de ReMap 2018 ont été fusionnés en un seul script snakemake. Le pipeline a, de plus, été révisé pour le rendre complètement modulaire. Chaque étape peut être adaptée afin de permettre le traitement des données issues d'expériences biologiques variées (ChIP-exo, ChIP-seq, DAP-seq) ou d'organismes différents (*Homo sapiens*, *Arabidopsis thaliana*). Les paramètres pour chaques workflows sont réunis dans un unique fichier externe pouvant être facilement modifié en fonction des besoins.

a : https://mesocentre.univ-amu.fr
b : https://hub.docker.com/u/benoitballester

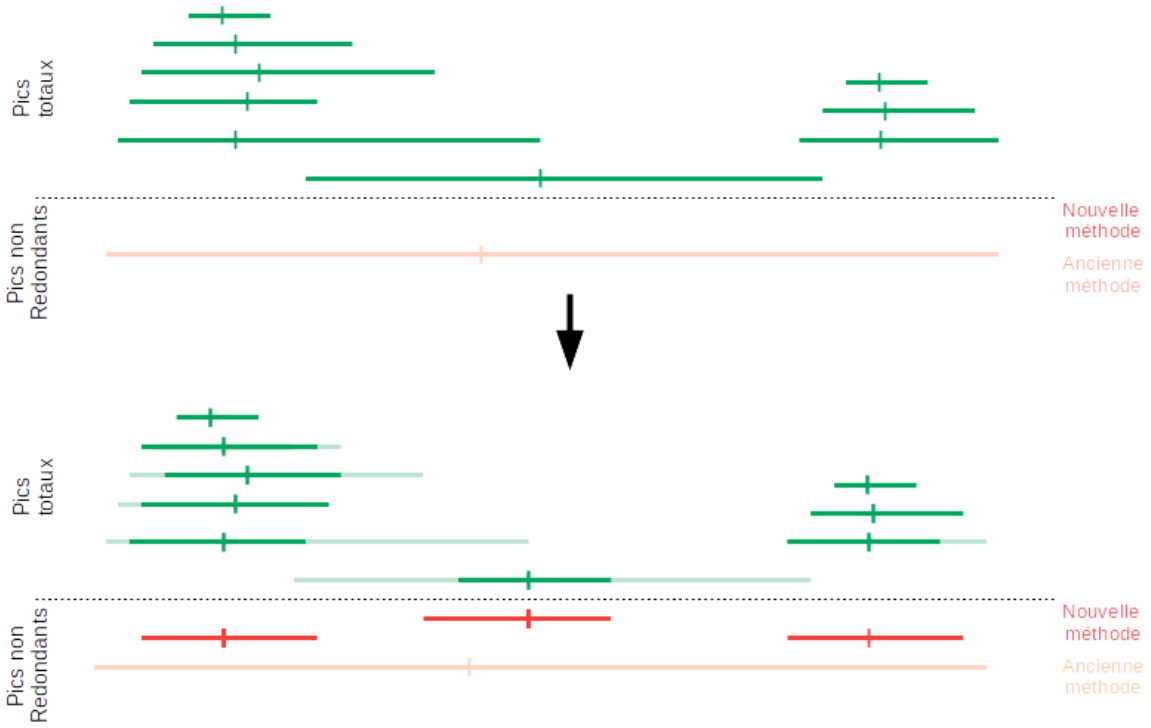


Figure II.1 : Méthodes de calcul de pics non redondants pour ReMap2018 et ReMap2020.
Cette figure représente une région contenant des pics de fixation d'une protéine. Les traits verticaux représentent le sommet des pics. Dans la première partie de la figure, les pics en vert foncé sont directement issus des expériences de ChIP-seq. En rouge clair, on trouve le pic non redondant calculé selon la méthode utilisée dans les précédentes versions de ReMap. Pour ReMap2018 et ReMap 2015, les pics se chevauchent sur plus de 25% de leur longueur sont fusionnés. Dans cet exemple, alors que nous pouvons distinguer deux clusters de fixation, la longueur des pics conduit à ne comptabiliser qu'un seul pic non redondant. La deuxième partie de la figure illustre la nouvelle méthode de calcul. Après calcul de la taille médiane de tous les pics de cette protéine, les pics dépassant la taille médiane sont réduits (les pics vert clair sont des pics bruts, les pics vert foncé, des pics réduits). C'est à partir de ces pics modifiés qu'est calculé le chevauchement des pics entre eux. Les pics se chevauchant sur plus de 25% de leur longueur sont fusionnés. On calcule le début et la fin des pics non redondants en faisant la moyenne de la position de début et de fin des pics qui les composent. Les pics non redondants obtenus par cette nouvelle méthode sont représentés en rouge foncé. La nouvelle méthode permet, dans les cas où plusieurs pics issus de clusters différents se chevauchent, de différencier les pics non redondants.

Le contrôle qualité des ChIP-exo est totalement différent de celui du ChIP-seq. Il a été décidé de créer un workflow spécifique pour cette méthode. Le filtrage par NSC et RSC n'est pas adapté aux expériences de ChIP-exo. En effet, la taille des fragments de librairie de ChIP-exo est proche de celle des fragments de lecture. Par conséquent, comparer la taille de ces deux types de fragments ne fournit d'indications ni sur la qualité du séquençage, ni sur celle de la librairie. Afin de garantir la qualité des données ChIP-exo, nous filtrons chacun des pics de chaque expérience suivant deux métriques telles que définis dans l'article scientifique d'où sont extraites les données de ChIP-exo utilisées par ReMap. Nous définissons le ratio du nombre de brins sens sur le nombre de brins antisens. Si ce ratio est supérieur à 4 le pic est retiré. Nous calculons, ensuite, le nombre de fragments de lecture dans la région du pic. Si on dénombre moins de 20 fragments pour une région de 500 pbs rapportés à 15 millions de fragments de lectures alignés, le pic est également retiré. La méthodologie de calcul des pics non redondants a, elle aussi, été revue afin d'augmenter la résolution des pics non redondants détectés (Figure II.1).

L'augmentation constante des données sur les bases de données publiques nécessite une puissance de calcul et un besoin de parallélisation de plus en plus avancés. Pour répondre à ce besoin, nous avons déplacé nos calculs au Mesocentre d'Aix-Marseille[a]. Ce centre propose des services de traitement à haut débit (HPC). Pour pallier les difficultés d'installation des logiciels sur un cluster, et augmenter la portabilité du pipeline nous avons été conduits à utiliser le gestionnaire d'environnement Singularity. Le gestionnaire de queues déployé sur le Mésocentre (SLURM) étant différent de celui utilisé dans nos précédents travaux (Torque), nous avons profité de la mise à jour de ReMap 2020 pour rendre tous nos processus portables. Grâce au système de gestion de workflows Snakemake, le pipeline ReMap 2020 est à la fois compatible avec Conda et Docker/Singularity. De plus, nous fournissons les environnements Conda nécessaires au workflow dans le GitHub et avons mis en place un Docker Hub permettant de stocker les images docker. Désormais, ReMap 2020 supporte différents gestionnaires de queue ce qui a permis de tester les workflows sur un serveur interne et de les déployer relativement facilement au Mesocentre.

C. Mise à jour de l'outil d'annotation web

Depuis le développement de ReMap 2015, nous fournissons un outil d'annotation en ligne permettant à l'utilisateur de comparer ses régions d'intérêt à notre catalogue. Pour la mise à jour ReMap 2020, nous avons actualisé cet outil d'annotation que nous avons renommé ReMapEnrich. D'un point de vue méthodologique l'outil permet de comparer l'enrichissement des protéines régulatrices de la transcription de ReMap pour les régions de l'utilisateur, à celui de l'enrichissement en protéines régulatrices de la transcription pour des régions aléatoires (Figure II.2).

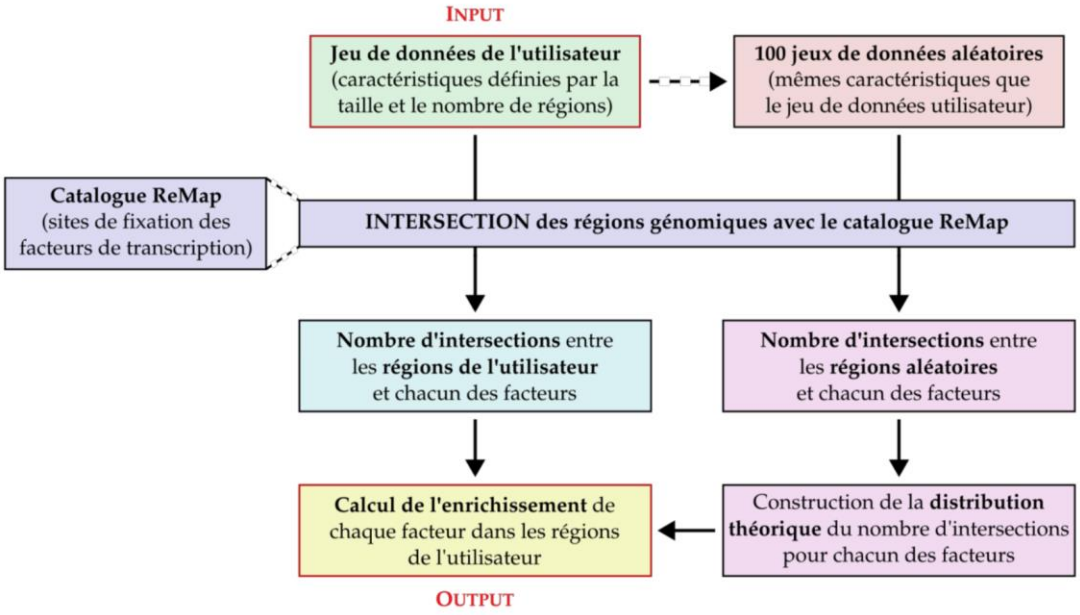


Figure II.2 : Pipeline de l'outil d'annotation des régions génomiques par rapport au catalogue ReMap.

ReMapEnrich permet, non seulement de corriger la p-valeur des tests d'enrichissement, mais aussi de choisir la fraction de chevauchement nécessaire pour qu'une région génomique soit considérée comme chevauchant une protéine régulatrice de la transcription de ReMap et inversement. La création du catalogue de régions aléatoires pour la distribution théorique est désormais dynamique et basée sur les caractéristiques des régions fournies par l'utilisateur (nombre, taille, etc). De plus, nous laissons à l'utilisateur le choix de limiter les régions du catalogue (=Univers) à des régions spécifiques. Ces régions peuvent être renseignées par l'utilisateur, mais nous fournissons également des univers précompilés à partir de catalogues de régions ouvertes (DNA-ase).

A la différence de sa précédente version, ReMapEnrich est un package R pouvant être téléchargé et utilisé hors ligne par les utilisateurs. Un GitHub a été créé pour le projet : https://github.com/remap-cisreg/ReMapEnrich. L'outil est disponible sur le site web de ReMap grâce à une interface R Shiny[a].

Figure II.2 : Etude de la complexité des éléments cis-régulateurs chez les mammifères en utilisant des approches à haut débit. Aurélien Griffon

a : : https://shiny.rstudio.com

D. Refonte du site

Afin d'améliorer l'accessibilité et l'utilisation des données produites par ReMap, le site web a totalement été remodelé. Il a été développé grâce au framework Laravel[a], outil open source basé sur PHP. Nous avons utilisé le Modèle-vue-contrôleur (MVC) qui s'appuie sur une architecture trois niveaux. Ce choix a permis le développement d'un site plus modulable et réactif. Les bases de données internes sont réunies sur une serveur maintenu grâce au système de gestion de bases de données relationnelles MySQL. Le site internet et ses outils web sont déployés sur un serveur HTTP Apache[b] contenu dans une image Docker. Le site web est donc portable et facilement déployable sur n'importe quel serveur HTTP possédant Docker. L'utilisation de conteneur Docker a grandement facilité l'intégration du module R Shiny de ReMapEnrich.

Nous avons également ajouté une interface REST (Representational state transfer) afin de permettre un accès programmatique aux données et métadonnées de ReMap. L'interface REST fournit dynamiquement des fichiers correspondant à des recherches complexes, ce qui n'était pas possible auparavant. L'implémentation d'une interface REST garantit le respect des principes FAIR. REST formalise et uniformise les requêtes favorisant ainsi l'interopérabilité entre les différents langages et systèmes d'exploitation. De plus, l'utilisation d'une interface facilite l'accessibilité et la recherche programmatique des données de ReMap. L'implémentation de cette ressource a permis d'améliorer la conformité de la base de données ReMap aux principes FAIR. Le nouveau site est présenté dans le figure II.3.

a : https://github.com/laravel/framework
b : https://httpd.apache.org

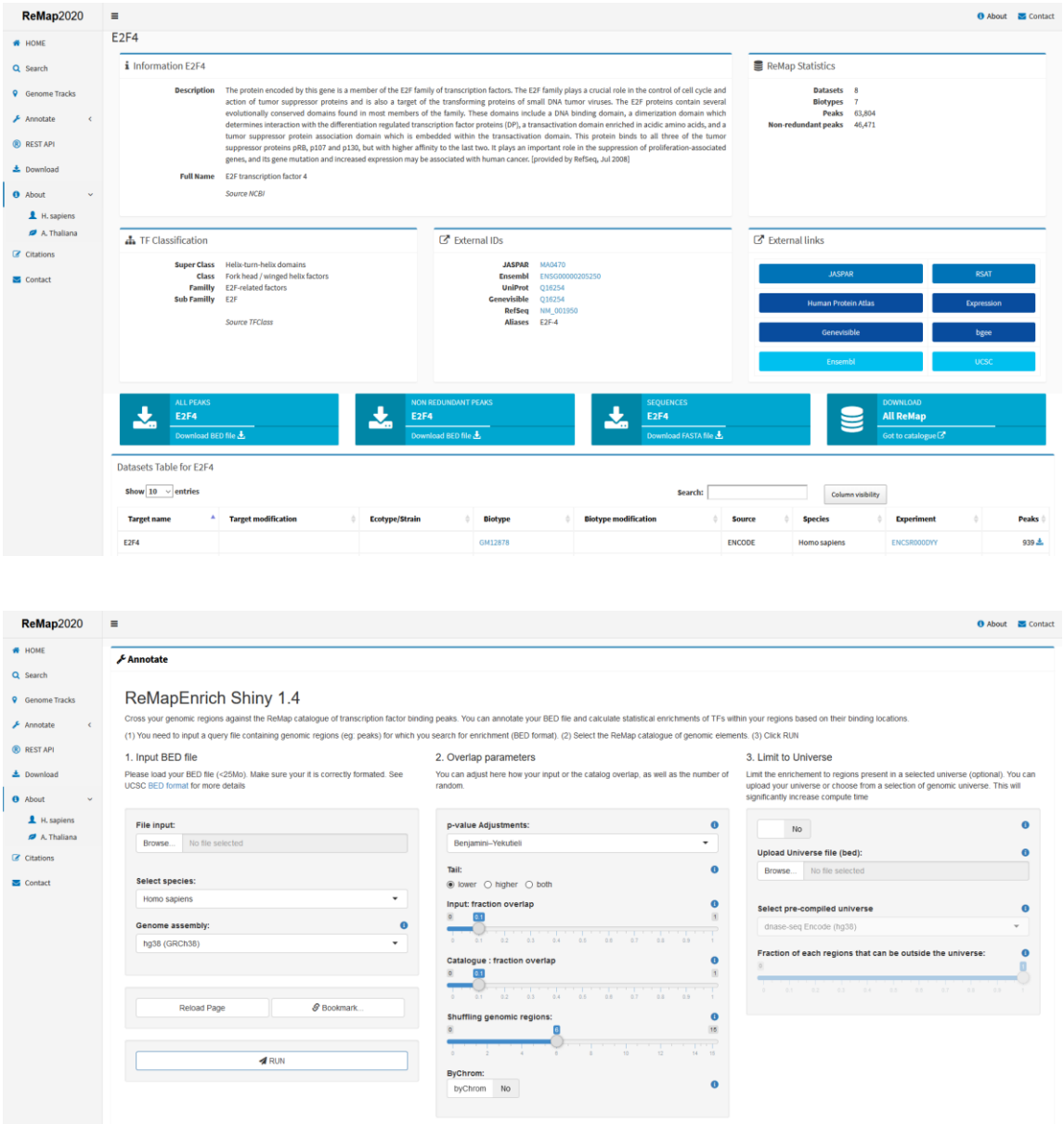


Figure II.3 : Nouveau portail ReMap2020.
Les métadonnées fournies pour chaque protéine sont similaires à celles fournies par ReMap2018. L'interface R Shiny de ReMapEnrich permet une modification des paramètres plus fine que celle proposée en 2018.

Nucleic Acids Research, 2019 **1**
doi: 10.1093/nar/gkz945

ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments

Jeanne Chèneby, Zacharie Ménétrier, Martin Mestdagh, Thomas Rosnet, Allyssa Douida, Wassim Rhalloussi, Aurélie Bergon, Fabrice Lopez and Benoit Ballester*

Aix Marseille Univ, INSERM, TAGC, Marseille, France

Received September 14, 2019; Revised October 07, 2019; Editorial Decision October 08, 2019; Accepted October 09, 2019

ABSTRACT

ReMap (http://remap.univ-amu.fr) aims to provide the largest catalogs of high-quality regulatory regions resulting from a large-scale integrative analysis of hundreds of transcription factors and regulators from DNA-binding experiments in Human and Arabidopsis (*Arabidopsis thaliana*). In this 2020 update of ReMap we have collected, analyzed and retained after quality control 2764 new human ChIP-seq and 208 ChIP-exo datasets available from public sources. The updated human atlas totalize 5798 datasets covering a total of 1135 transcriptional regulators (TRs) with a catalog of 165 million (M) peaks. This ReMap update comes with two unique Arabidopsis regulatory catalogs. First, a catalog of 372 Arabidopsis TRs across 2.6M peaks as a result of the integration of 509 ChIP-seq and DAP-seq datasets. Second, a catalog of 33 histone modifications and variants across 4.5M peaks from the integration of 286 ChIP-seq datasets. All catalogs are made available through track hubs at Ensembl and UCSC Genome Browsers. Additionally, this update comes with a new web framework providing an interactive user-interface, including improved search features. Finally, full programmatically access to the underlying data is available using a RESTful API together with a new R Shiny interface for a TRs binding enrichment analysis tool.

INTRODUCTION

The rapid accumulation of experiments capturing protein–DNA interactions in public databases provides a unique and valuable resource for thousands of protein occupancy maps. The development of high-throughput methods like chromatin immunoprecipitation followed by sequencing (ChIP-seq) (1), or ChIP-seq with DNA digestion steps (ChIP-exo) (2) as well as DNA affinity purification sequencing (DAP-seq) (3) has allowed to experimentally obtain genome-wide maps of binding regions across many cell types for a variety of DNA-binding proteins. Integrating these thousands of large-scale experiments would allow to explore the depth of the transcriptional regulatory repertoire. Unfortunately, the heterogeneous experimental metadata annotations deposited in data-warehouse, the inconsistency in target name convention, the different cell type or tissue names, the variety of bioinformatics methods and underlying file formats challenge a global analysis process and the underlying mapping of TF binding regions.

ReMap has been the first large scale integrative initiative with dedicated curation and uniform data processing pipeline to reveal the complex architecture of the human regulatory landscape (4). The core foundation of the ReMap project rely on the manual curation and annotation of experiments metadata. Each experiment metadata introduced in ReMap has been assessed and manually curated to ensure correct target and biotype annotation. The ReMap 2015 database (4) introduced a catalog of 13 million (M) DNA binding regions by compiling the genomic localization of 237 different transcriptional regulators (TRs) across 83 different human cell lines and tissue types based on the integration of 395 datasets from Gene Expression Omnibus (5) and ENCODE (6,7). For ReMap 2018 we updated the catalogue by processing 2829 quality controlled ChIP-seq datasets leading to a unique atlas of regulatory regions for 485 TRs across 346 cell types, for a total of 80M DNA binding regions (8).

Here, we describe ReMap 2020 which introduces two unique regulatory catalogs for Arabidopsis (*Arabidopsis thaliana*), and includes a major expansion of the human regulatory catalog, along with new user-interface features. The Arabidopsis regulatory catalog is the result of curation, annotation and integration of 179 quality controlled ChIP-seq and 330 DAP-seq datasets for transcription factors and general components of the transcriptional machinery. Those datasets have been mapped to the TAIR10 Arabidopsis assembly and analyzed with a uniform pipeline. This unified

*To whom correspondence should be addressed. Tel: +33 4 91 82 87 28; Fax: +33 4 91 82 87 01; Email: benoit.ballester@inserm.fr

integration of Arabidopsis datasets lead to a unique atlas of 342 TRs across 20 biotypes, 12 ecotypes, for a total of 2.6M DNA binding regions. Additionally, we introduce a catalog of 33 Arabidopsis histone modifications and variants across 4.5M peaks from the integration of 286 ChIP-seq datasets.

The human atlas has been updated with the curation, annotation and integration of 2969 quality controlled ChIP-seq and 208 ChIP-exo datasets mapped to the GRCh38/hg38 assembly and analysed with the ReMap pipeline. In this update, we propose a unique atlas of regulatory regions for 1135 TRs across 602 cell types, in 5798 datasets, for a total of 165M DNA binding regions.

The ReMap 2020 human update represents a 1.7-fold increase in the number of cell lines/tissue types, and a 2-fold increase in the number of DNA-binding proteins, number of processed datasets and number of identified peaks. While the eMap 2018 human catalog covers 19% (0.6 Gb) of the human genome with more than five peaks, the regulatory search space for ReMap 2020 covers 34% (1 Gb, 5+ peaks).

Finally, the fully redesigned ReMap web-interface gives the community richer options to navigate and search our data, to visualize and browse all catalogs with public track hubs integrated in Ensembl and UCSC Genome Browsers. Additionally we updated our ReMapEnrich tool with a new R Shiny interface, while also allowing programmatic access to the underlying data with a RESTful API.

This report presents the third ReMap release, which comes with two unique catalogs for Arabidopsis and an extensive data increase and regulatory catalog expansion of the human atlas as a result of our large-scale data integration and analysis efforts. The manual metadata curation engaged in the ReMap project offers a unique and unprecedented collection of DNA-binding regions for two species. This data expansion is supported by a range of new functionalities for better community access.

MATERIALS AND METHODS

Available Human and Arabidopsis datasets

New DNA-binding experiments such as ChIP-seq, ChIP-exo and DAP-seq were extracted from the NCBI Gene Expression Omnibus (GEO) (5) and ENCODE (6,7) databases. For GEO, the query 'Genome binding/occupancy profiling by high-throughput sequencing' AND 'homo sapiens'[organism] AND NOT 'ENCODE'[project]' was used to return a list of all potential datasets, which were then manually assessed and curated for further analyses. The same query with 'arabidopsis thaliana'[organism] was used to return all potential datasets. For each experiment ReMap metadata are manually curated and annotated with the official gene symbol. Materials and methods from published papers are often read when deposited metadata is insufficient. For human we used the HUGO Gene Nomenclature Committee (9) (www.genenames.org), BRENDA Tissue Ontologies (10) for cell lines (www.ebi.ac.uk/ols/ontologies/bto) as well as the Cellosaurus database (11) to homogenize cell names (e.g. MCF-7 not MCF7, Hep-G2, not HepG2, Hepg2 etc.). For Arabidopsis (*Arabidopsis thaliana*) we used the Ensembl Genome (12) gene symbols. Ecotypes and biotypes description were curated and homogenized when the information was available in the metadata or paper. Datasets involving polymerases (Pol2 and Pol3), and some mutated or fused TFs (e.g. KAP1 N/C terminal mutation, GSE27929) were filtered out. When multiple antibodies were pooled (eg: RUNX1 and RUNX3, GSE17954) we would name the target as RUNX1-3.

In ReMap, we define a dataset as a DNA-binding experiment in a given GEO/AE/ENCODE series (e.g. GSE37345), for a given TR (e.g. FOXA1), and in a particular biotype (e.g. LNCaP, K-562, Leaf, Seedling) in a given biological condition (e.g. 45 min DMSO, 21 days-wt-watered). Datasets were labeled with the concatenation of these pieces of information (e.g. GSE37345.FOXA1.LNCAP_45 min-DMSO). The core of ReMap data rely on ChIP-seq assays, but for this update we analysed a major human ChIP-exo experiment (GSE78099), and for Arabidopsis a major DAP-seq experiment (GSE60141).

A total of 7908 datasets were processed (Supplementary Table S1). Specifically, we analyzed 6498 human datasets deposited in public repositories from 1 July 2008 to 10 November 2018), and 1410 Arabidopsis datasets from 1 January 2009 to 2 February 2018 (full list of datasets in Supplementary Tables S5, S8, S11). For the ENCODE data in the 2020 update, we re-analyzed, starting from the raw data, all new ENCODE ChIP-seq experiments for TFs, transcriptional and chromatin regulators released since ReMap 2018 release (1 August 2016 to 5 February 2019), following the same processing pipeline. We retrieved the list of ENCODE data as FASTQ files from the ENCODE portal (https://www.encodeproject.org/) using the following filters: Assay: 'ChIP-seq', Organism: 'Homo sapiens', Target of assay: 'TF', Available data: 'fastq' on 5 February 2019. Metadata information in JSON format and FASTQ files were retrieved using the Python requests module. We processed 964 ENCODE datasets, 894 of whom passed our quality filters. We renamed TR ENCODE aliases into official HGNC identifiers (e.g. p65 into RELA, see Supplementary Table S6), and renamed cell lines to official BRENDA and Cellosaurus conventions (e.g. K562 into K-562, lost of modified names Supplementary Tables S6, S9, S12).

ChIP-seq processing

All human and arabidopsis ChIP-seq datasets were uniformly curated, processed and analyzed. Bowtie 2 (version 2.2.9) (13) with options -end-to-end -sensitive was used to align all reads on the human genome GRCh38/hg38 assembly and on the *A. thaliana* TAIR10 assembly. Adapters were removed using Trim Galore (https://github.com/FelixKrueger/TrimGalore), trimming reads up to 30 bp. Trim Galore is a wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files. Polymerase chain reaction duplicates were removed from the alignments with samtools rmdup (14). TR binding regions were identified using the MACS2 (15) peak-calling tool (version 2.1.1.2) to follow ENCODE ChIP-seq guidelines (13), with stringent thresholds (MACS2 default thresholds, *Q*-value: 1e–5, -g: with corresponding genome sizes). When available in the experiments, input datasets were used in the peak calling process. All peak-calling nar-

rowPeak or broadPeak files are available to download. For ChIP-seq, ChIP-exo and (amp)DAP-seq analyses, peak files containing <100 peaks were discarded.

Quality assessment

Data quality differs across experiments, as the data we process comes from various sources and are generated under different experimental conditions and platforms. Since the first release of ReMap 2015, our pipeline has assessed the quality of each dataset processed, unlike similar databases (Supplementary Tables S5, S8, S11). For ReMap 2020 the same quality pipeline and cutoffs were applied as in ReMap 2018 (8). Briefly, for both species, and all ChIP-seq and DAP-seq datasets processed for this update we computed a score based on the cross-correlation and the FRiP (fraction of reads in peaks) metrics developed by the ENCODE consortium (16) (Supplementary Figures S1–S3, ENCODE quality coefficients http://genome.ucsc.edu/ENCODE/qualityMetrics.html). Then our pipeline computes the normalized strand cross-correlation coefficient (NSC) as a ratio between the maximal fragment-length cross-correlation value and the background cross-correlation value, and the relative strand cross-correlation coefficient (RSC), as a ratio between the fragment-length cross-correlation and the read-length cross-correlation. Datasets not passing the QC were not included in catalogs or BED files available for download. Rejected datasets are listed in (Supplementary Tables S2, S7, S10). For the human ChIP-exo data (GSE78099) we applied three post-processing steps as described in (16) material and methods. In brief, we filtered out peaks that would meet any of these criteria: MACS2 score <80 (equivalent to a $P = 1 \times 10^{-8}$); ratio of forward versus reverse strand reads >4; <20 reads over 500 bp per 15 million reads; normalized read count was less than twofold over the control. For the arabidopsis DAP-seq (17) data (GSE60141), we applied our standard ChIP-seq ReMap pipeline. Our quality assessment protocol could not be applied to ChIP-exo data, as the specificity of ChIP-exo peaks (extremely narrow) would not allow the computing of FRiP/RSC/NCS scores.

Open ReMap pipeline

A common issue with bioinformatics workflow is that it normally evolves at a different speed than data is published in the literature. We are making the code of our ReMap pipeline available to GitHub in the ReMap Github organisation (https://github.com/remap-cisreg). As the ReMap project expands dramatically and regulatory catalogs for other species are requested, it becomes essential to enable joint efforts between the ReMap team and external teams. This is essential for future collaborative production efforts. Briefly, our pipelines uses SnakeMake either in a Conda or Singularity environment depending on the HPC resources, where Torque and Slurm managers are both supported. Details of the pipeline are published on the repository.

Genome coverage

Genome coverages were computed using the BedTools suite (17) (version 2.26.0) using the 'genomecov' function with the option -max 100 that combines all positions with a depth ≥100 binding locations. Full details of the ReMap 2018 and 2020 genome coverage are available in Supplementary Table S4 for both species. Genomic regions covered by at least five peaks were considered as regulatory (Figure 1E, K, blue/green and light blue/green), uncertain regulatory regions (Figure 1K, grey), or not covered (Figure 1K, light grey).

Non-redundant peak sets and CRMs

For each target, ReMap provides non-redundant binding regions, a specificity not found in other databases (Supplementary Table S3). As the number of redundant peaks increases, and to improve the genomic accuracy of these non-redundant peaks, we updated our method. For a given TR, all peak lengths were truncated to the median size of all peaks for this TR. Then, to find clusters of redundant peaks, we used BedTools to intersect overlapping truncated peaks across different datasets (with at least 25% overlap, both ways). Once the clusters of overlapping peaks identified, non-redundant peaks are computed by averaging start, end and summits coordinates of all peaks in a cluster using original ReMap peaks lengths. For a given factor across all experiments, the non-redundant peak set consists of the computed non-redundant peaks plus singletons, and are available for download from the ReMap website. *cis*-regulatory modules (CRMs) were obtained by merging regions of all non-redundant peaks using BedTools. Regions bound by several TRs are called CRMs, whereas regions bound by only one TR are labeled as singletons.

HUMAN REGULATORY CATALOG EXPANSION

This 2020 release of the ReMap human database exhibits significant growth in the number of datasets, the number of transcriptional regulators and overall in the number of binding regions integrated in our catalog. In this update, we curated processed and analyzed 3424 new human ChIP-seq and 222 ChIP-exo datasets against TRs from GEO and ENCODE. Since the first ReMap release, we ensure consistency and comparability across datasets by processing from the raw data, through our uniform ReMap pipeline that includes read filtering, read mapping, peak calling and quality assessment (see 'Materials and Methods' section). Unlike other databases, the core foundation of ReMap lie in the manual curation of metadata, involving reading materials and methods as submitted experiments annotation is heterogeneous. In addition, we run a critical data quality filtering step in our pipeline to address varying quality of DNA-binding experiments (18,19). After applying our quality filters we retained 2969 datasets (82%): 2767 ChIP-seq and 208 ChIP-exo datasets (Figure 1A and Supplementary Figure S1). This leads to an updated human regulatory atlas totalizing 5798 datasets. The uniform data processing contributes to a final ReMap 2020 human regulatory atlas of 164 732 372 peaks generated from 11 135 TRs (Figure 1B). Our analyses produced 163 741 896 peaks set across 927 TRs for ChIP-seq and 990 476 peaks set across 208 TRs for ChIP-exo.

This update shows a 2-fold increase in the number of TRs and number of peaks. The significant data growth is

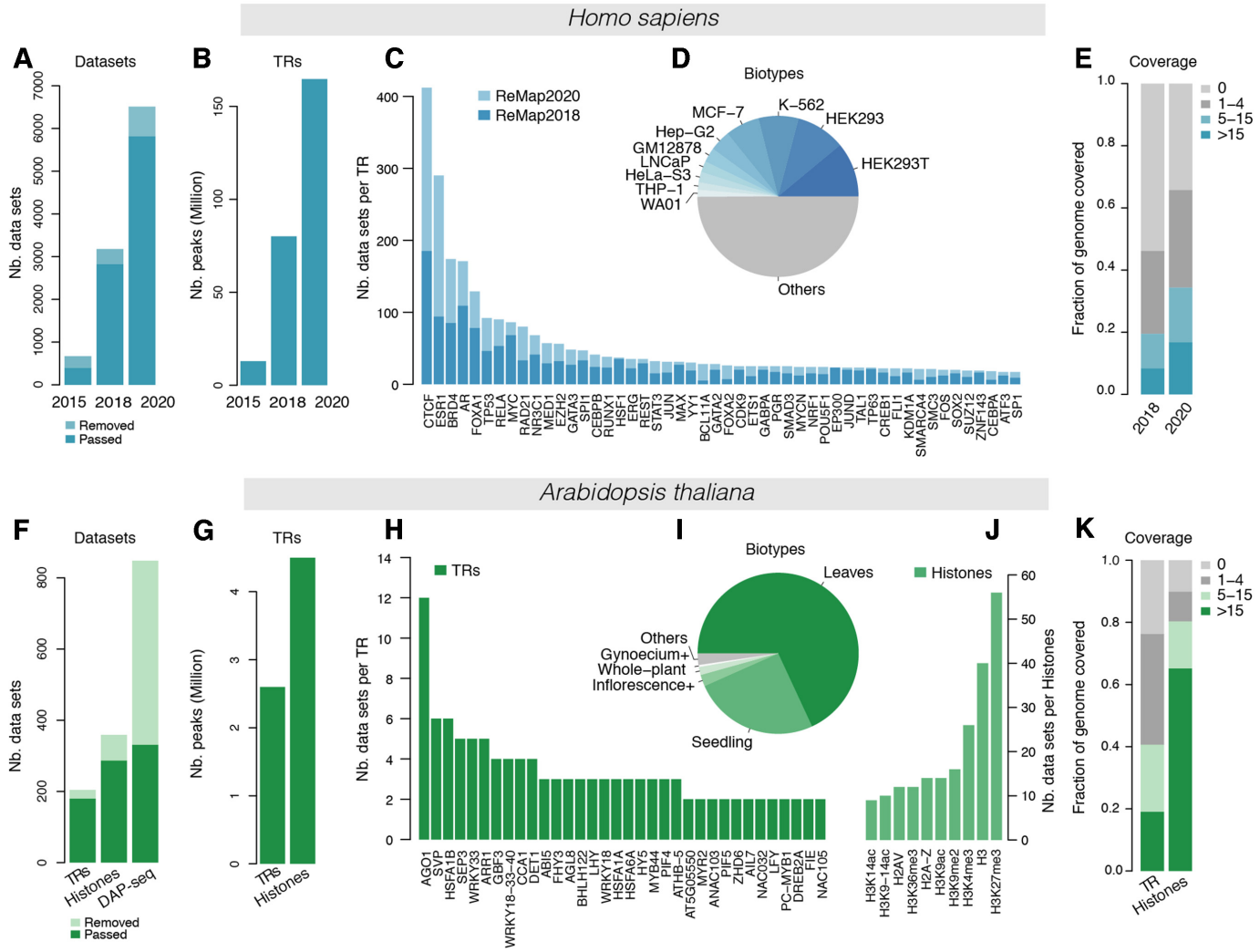


Figure 1. Overview of the ReMap database expansion. (**A**) Analyzed datasets growth in ReMap 2020 compared to 2018 and 2015 in human. (**B**) Transcriptional regulators (TRs) growth in ReMap 2020 compared to 2018 and 2015. (**C**, **D**) Evolution of the number of datasets across the top 50 TRs between ReMap 2020 and 2018. (**E**) Genome coverage fraction of each ReMap atlas by genomic regions covered by at least five peaks considered as regulatory (blue, light blue), potentially regulatory (grey), or not covered (light grey). (**F**) Analyzed TRs and Histones datasets in Arabidopsis. (**G**) Transcriptional regulators (TRs) in ReMap Arabidopsis. (**H, I**) Number of datasets for the top 30 TRs and top 5 biotypes. (**J**) Number of datasets for the top 10 Histone modifications and variants. (**K**) Genome coverage fraction of TRs and Histone ReMap catalogs, regions covered by at least five peaks considered regulatory (green, light green), potentially regulatory (grey), or not covered (light grey).

spread across almost all TRs when compared to ReMap 2018 (Figure 1C, light blue bars). More specifically, we observe that key TFs (e.g. ESR1, AR, FOXA1, TP53), transcriptional repressors (e.g. CTCF) and CRFs (e.g. BRD4) display larger data expansion than other DNA-binding proteins. Nevertheless, most of the top 50 TRs show additional datasets integrated in ReMap 2020 (Figure 1C, light blue bars). The top 10 most frequent biotypes correspond to the most common cell lines used in genomics (Figure 1D). The constant integration of a wide spectrum of cell lines and tissues will allow for a greater definition of the regulatory space across the genome. Indeed, ∼34.4% (106 Gb) of the human genome is covered by at least five features or more, and 17% (516 Mb) covered by 15 features or more (Figure 1E and Supplementary Table S4). As comparison, the ReMap 2018 catalog covered 19% (601 Mb) of the genome by at least five features, and 8% (257 Mb) with at least 15 or more features.

In this update, we expand the regulatory panorama revealing dense co-localized regulatory regions at unprecedented depth (Figure 2A). Indeed, the ReMap database shows an unprecedented landscape of the human regulatory abundance and complexity constituted by 165M binding regions forming 1.7M CRMs. The genomic organization of our atlas reveals dense co-localizations of peaks forming tight clusters of heterogeneous binding regions with variable TRs complexity (Figure 2A, Supplementary Figure S4). Since 2015 we highlight this regulatory complexity by observing the vicinity of the human ELAC1 promoter illustrating ReMap 2015, 2018 and 2020 catalog growth (n = 229; 1037; 2546 peaks respectively). We observe three clusters of peaks, one large at the promoter followed by two

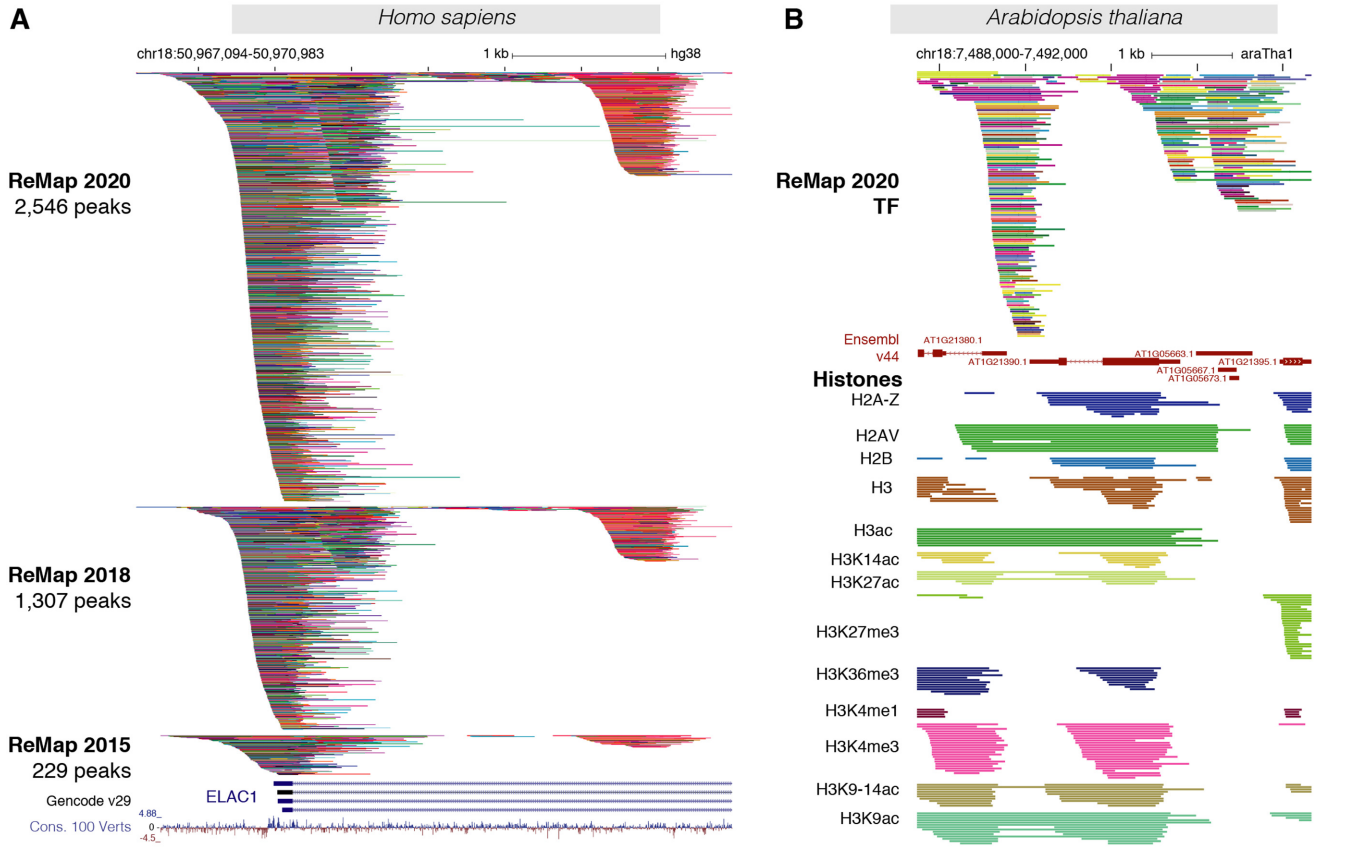


Figure 2. Genome browser views of both ReMap 2020 atlas. (**A**) ReMap 2020 human DNA–protein binding pattern of 5798 datasets. A genome browser example of the DNA-binding peak depth of the ReMap 2020 atlas compared to ReMap 2018 and 2015 at the vicinity of the ELAC1 promoter (hg38 chr18:50 967 094–50 970 983). The tracks displayed are compacted to thin lines so the depth of ReMap 2020 bindings can be compared to 2015. Uncompacted screenshot is available as Supplementary Figures S5. In this region ReMap 2020 displays 2546 peaks, 1307 peaks for ReMap2018, 229 peak for ReMap 2015 (lifted to GRCh38/hg38). The following genome tracks correspond to the GENCODE v29 annotation and the 100 vertebrates base-wise conservation showing sites predicted to be conserved (positive scores in blue). (**B**) A genome browser view of the first ReMap 2020 Arabidopsis TF and Histones modifications catalogs at the vicinity of the AT1G21390 gene (araTha1 chr1:7 488 000–7 492 000). The annotation genome track correspond to the Ensembl Genes v44 TAIR10 annotation. All peaks have been compacted for rendering and 13 (out of 33) histone modifications displayed.

clusters located at about +500 bp and +2 kb from the transcription start site. This third cluster has been detailed in our previous ReMap publications (4,8) to illustrate how integrating data from different sources improves genome annotations. Indeed this cluster contains two ENCODE peaks for FOXA1 in the 2020 update, and only one in previous versions. This update further consolidate the binding location with 93 FOXA1 peaks (60 peaks in 2018, 15 in 2015) across different cells, antibodies, and laboratories (Supplementary Figure S4). The summit of each peak is represented by a vertical bar, which when aggregated closely gives information about the putative location of the DNA binding site. This FOXA1 clustering shows overlapping peaks, and does not reveal the discrete repertoire of binding regions in the human genome. Therefore, to address redundancy between datasets, we merged peaks for the same TR, resulting in a catalog of 76M non-redundant peaks. The genomic accuracy of these non-redundant peaks have been improved with a new method to reduce peaks redundancy (see 'Materials and Methods' for details). Non-redundant peaks and CRMs annotations are computed across all ReMap datasets and biotypes, thus representing a multi-cellular multi-tissue regulatory map. Taken together, the 2020 update of the ReMap human catalog provides a unique opportunity to identify complex regulatory architectures in our genome, each containing hundreds or thousands of bound regulators. By adding more experiments and more DNA-binding proteins to the atlas, we increased the genome regulatory space (Figure 1), dramatically increased its depth (Figure 2), and refined current annotations of bound regions (Supplementary Figure S4).

FIRST ARABIDOPSIS THALIANA REGULATORY CATALOGS

Arabidopsis transcriptional regulators atlas

This ReMap release comes with the first large scale Arabidopsis regulatory atlas for transcription factors and general components of the transcriptional machinery. Indeed, the extent of the regulatory space in the Arabidopsis genome has not yet been fully apprehended with systematic integration of public DNA-binding assays. Thus, to enable genome-wide identification of Arabidopsis regulatory ele-

ments we have collected, curated, uniformly processed and analysed 204 ChIP-seq and 848 DAP-seq datasets against TRs from GEO (Figure 1F and Supplementary Figure S2). We retained 509 datasets after quality control: 179 ChIP-seq and 330 DAP-seq datasets leading to a final Arabidopsis regulatory atlas of 2 645 004 peaks for 372 TRs (Figure 1F and G). While a study (20) was carried out by integrating TF peaks from few ChIP-seq experiments in flowers development, our ReMap atlas is the first to provide a global view of all detected TRs binding in a wide variety of biological contexts and variety of experiments (Figure 2B). The top three most represented TRs are Argonaute protein AGO1 (mRNA and chromatin binding), MADS-box protein SVP (Transcription repressor), Heat stress transcription factor HSFA1B (Figure 1H), while the two most represented tissues are leaves and seedlings (Figure 1I). About 40% (49 Mb) of the Arabidopsis genome is covered by at least five features or more, and 19% (22 Mb) covered by 15 features or more (Figure 1K and Supplementary Table S4). We present here a unique transcription factors occupancy map forming complex architecture in the plant genome revealed by the first large scale integration of public Arabidopsis DNA-binding experiments.

Arabidopsis histone modifications atlas

The ReMap Arabidopsis integration efforts have been complemented with the first release of histone modifications catalogue as a result of the uniform integration of ChIP-seq datasets. Histone modifications and variants regulate gene expression by remodelling the chromatin structure thus acting on chromatin accessibility (21–23). The integration of the histone modifications assays can be used to attribute functional properties to genomic regions as histone modification positioning often correlate with genomic features (6,24). We have collected, curated, processed and analysed 358 ChIP-seq datasets against histone modifications (Figure 1F and Supplementary Figure S3). We retained 286 datasets after quality control leading to a final atlas of 4 528 203 peaks covered by 33 histone modifications and variants (Figures 1F, G, 2B, Supplementary Figure S7). The top three most represented histones modifications and variants are H3K27me3, H3 and H3K4me3 (Figure 1J), with seedlings and leaves being most common represented tissues (Supplementary Figure S8). As histone modifications remodel large and broad chromatin regions, 80% (96 Mb) of the Arabidopsis genome is covered by at least 5 features or more (Figure 1K and Supplementary Table S4). This ReMap update comes with two unique Arabidopsis regulatory catalogues, one providing 372 transcription factors and general actors of the transcriptional machinery and a second catalogue of 33 of known histone modifications and variants.

A NEWLY DESIGNED WEB PORTAL

A new web interface

With this ReMap 2020 update, a new web interface has been designed to incorporate many of the features and characteristics present in contemporary websites. The new interactive website is designed with Laravel, a PHP web application framework for the development of web applications following the model–view–controller (MVC) architectural pattern. As template we used AdminLTE, a responsive HTML template based on the CSS framework Bootstrap 3. We used MySQL as a backend database to store ReMap datasets, targets, biotypes informations and external metadata. The site can be divided into six main sections: Search, Target and Biotype pages, Browsing data, the ReMapEnrich R Shiny tool, RESTful API and About pages. We have greatly improved the search functionalities of ReMap (Figure 3A). The search bar with advanced search options has been introduced into the main page for direct access to the data. The new search interface searches among all the main metadata, target synonyms, biotype synonyms to identify potential matches. Search results are presented in a paginated table which can be further filtered by a filter box. We have improved the Target and Biotype pages data visibility and usability by organising collections of blocks. Each Target and Biotype page contains an interactive paginated table for downloading separate dataset as narrow or broad peak files from macs2 (Figure 3B). The 'Browsing' page and browsing vignettes assist in the visual exploration of the regulatory catalogues, allowing researchers to navigate data in genome browsers. Finally, the About pages provide detailed information for each species on the experiments, pre-processing steps, quality results, as well as basic statistics of the catalogues. We believe that the new web interface and features incorporated in ReMap 2020 have improved our users experience. ReMap 2020 can be accessed at http://remap.univ-amu.fr or http://remap.cisreg.eu.

New genomic enrichment interface

Since ReMap 2015 an annotation tool has been deployed allowing users to query their regions of interests against our catalogue to identify enrichments of TRs. However a better web interface and statistics to compute genomic region enrichment analyses against ReMap catalogs was much needed. Associated with this release we introduced the ReMapEnrich R Shiny interface (Figure 3C) to identify significantly enriched regions from user defined regions against ReMap catalogs. This interface facilitates the interpretation of functional genomics, epigenomics and genomics data by providing common statistical functions and plot the enrichment analyses of ReMap TRs.

RESTful API

We have developed a RESTful web service API to provide programmatic access to ReMap metadata as well as additional resources such as target, biotype informations, and BED files (Figure 3D). The API has been implemented for querying the database with a batch list of TR, and/or biotypes. The RESTful API enables various bioinformatics tasks through a broad range of client software to access ReMap data programmatically. This web services is a convenient solution to integrate in an automated manner heterogeneous genomic information in complex work-

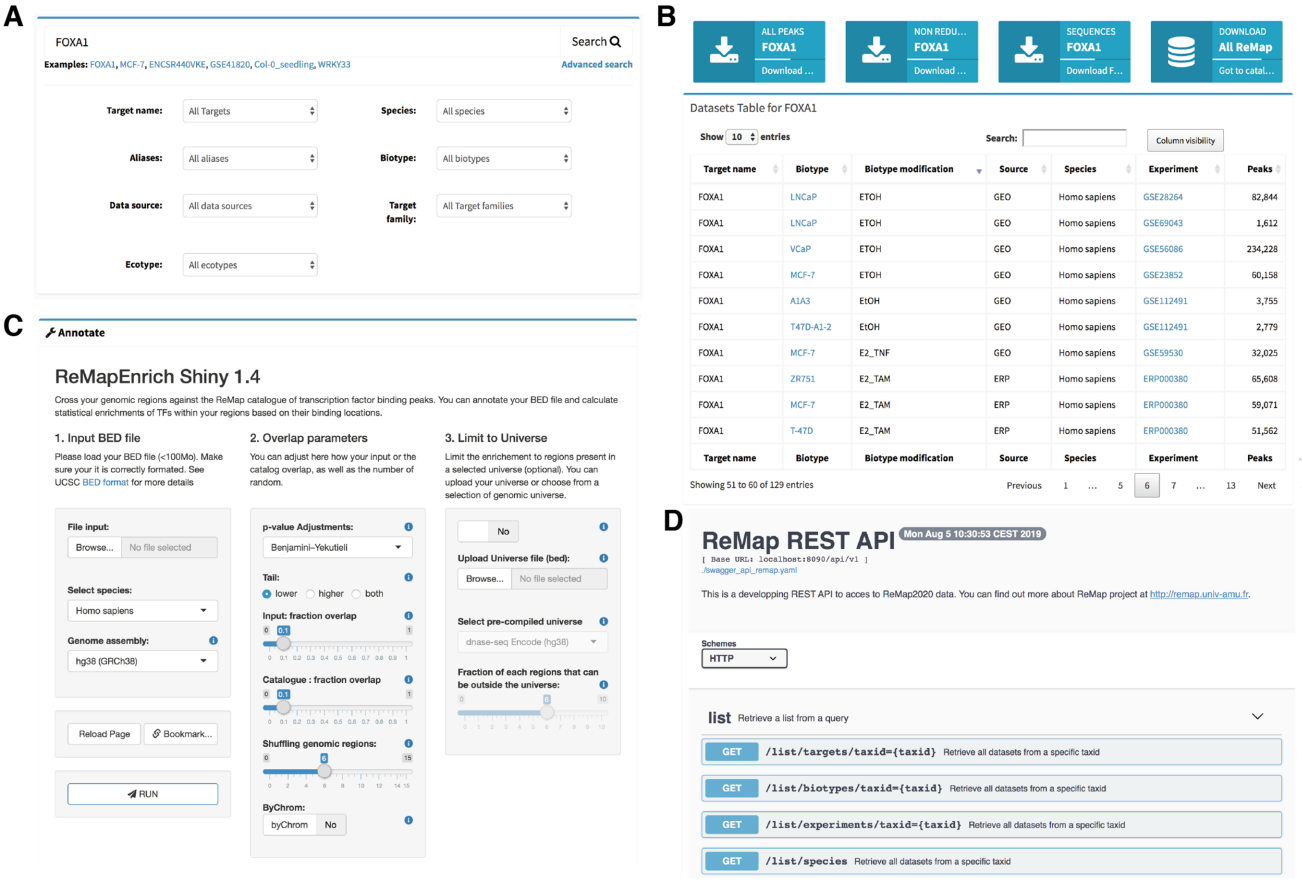


Figure 3. Overview of the ReMap 2020 new web interface with interactive searching activity. (**A**) A full search bar on the homepage and search page with advanced features. (**B**) For each targets or biotype page a responsive table allowing further filtering. (**C**) The ReMapEnrich R Shiny interface for plotting enrichment analyses. (**D**) List of queries available for the ReMap RESTful API.

flows without the need of local databases installations. The RESTful API is natively implemented in the Laravel Web Framework of ReMap, it returns the data in the convenient JSON format, and is accessible at http://remap.univ-amu.fr/rest_page.

Data download and genome browsers

Most users may prefer to download our atlas directly and query it locally, for instance to integrate with internal data or use in workflows. Each of the three ReMap database releases are available from the download page, where the entire catalogs are accessible in BED format in a variety of configurations: 'all' peaks, non-redundant peaks, CRM peaks. In addition, the download page is the entry point for searching and directly accessing BED files for specific targets, biotypes or datasets centric tables.

Since the 2018 update we provided data navigation options for visual exploration of the ReMap regulatory catalogs combined with public or user-specific genome-wide annotations. As the ReMap catalogs expand in size and complexity, and researchers routinely generate large scale genome-wide genomic data, it became crucial to provide navigation flexibility to centralise these data. The content of the different ReMap atlas can be browsed across major Genome Browsers such as Ensembl (25), Ensembl Genomes (12) and the UCSC Genome Browser (26) within public sessions or public hubs (Figure 2, Supplementary Figures S5–7). Additionally, for each ReMap species, track hubs have been deposited to the Track Hub Registry (https://trackhubregistry.org/) for open data integration in various platforms or browsers (27). Our goal is to facilitate researchers to discover the abundance and complexity of ReMap catalogs in combination with other biological tracks.

CONCLUSION AND FUTURE DIRECTIONS

The 2020 release of ReMap maintains the long-term focus of providing the research community with the largest catalogs of high-quality regulatory regions by integrating all available DNA-binding assays. The ReMap 2020 update comes with (i) a significant expansion of the human regulatory atlas; (ii) the first Arabidopsis regulatory atlas for transcriptional regulators; (iii) the first Arabidopsis catalog for histone modifications; (iv) a new website with improved user experience; (v) a new annotation tool with an R Shiny interface; (vi) a programmatic access to ReMap with a RESTful API; (vii) an updated genome browsing experience with Track Hubs data integration for different platforms. The

quality of ReMap data is illustrated in the latest version of the JASPAR database (28), for which peaks were used to calculate TF link profiles. We believe that our ReMap catalogs will help in better understanding the regulation processes in Human and Arabidopsis.

In the future, as new datasets are constantly added to repositories, we would like to engage the scientific community in the curation process to increase our capacity to introduce new regulatory profiles for different species and different DNA-binding assays. As a long term goal, we would like to call the regulatory community with the curation and production processes to provide the best annotated regulatory repertoire in different species.

FEEDBACK

The ReMap team welcomes your feedback on the catalogs, use of the website and use of the downloadable files. We thank our users for past and future feedback to make ReMap useful for the community. Please contact benoit.ballester@inserm.fr for development requests.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Lionel Spinelli for in depth scientific discussions in pipeline implementation and his expertise on Docker and Singularity usage with the HPC resource. We would like to thank Marius Gheorghe, Aziz Khan and Anthony Mathelier from NCMM Norway for constant scientific feedback, the UCSC Genome informatics groups for help with tack hubs, the Ensembl and Ensembl Plant group for help with the Homo sapiens and Arabidopsis thaliana track hub. This work was granted access to the HPC resources of Aix-Marseille Université financed by the project Equip@Meso [ANR-10-EQPX-29-01] of the program 'Investissements d'Avenir' supervised by the Agence Nationale de la Recherche.

FUNDING

French Ministry of Higher Education and Research (MESR) PhD Fellowship (to J.C.); Funding for open access charge: Institut National de la Santé et de la Recherche Médicale (INSERM).
Conflict of interest statement. None declared.

REFERENCES

1. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
2. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive Genome-wide Protein-DNA interactions detected at single nucleotide resolution. *Cell*, **147**, 1408–1419.
3. Bartlett,A., O'Malley,R.C., Huang,S.-S.C., Galli,M., Nery,J.R., Gallavotti,A. and Ecker,J.R. (2017) Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat. Protoc.*, **12**, 1659–1672.
4. Griffon,A., Barbier,Q., Dalino,J., van Helden,J., Spicuglia,S. and Ballester,B. (2015) Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Res.*, **43**, e27.
5. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2012) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**, D991–D995.
6. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
7. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Gabdank,I., Hilton,J.A., Jain,K., Baymuradov,U.K., Narayanan,A.K. *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.
8. Chèneby,J., Gheorghe,M., Artufel,M., Mathelier,A. and Ballester,B. (2018) ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.*, **46**, D267–D275.
9. Braschi,B., Denny,P., Gray,K., Jones,T., Seal,R., Tweedie,S., Yates,B. and Bruford,E. (2019) Genenames.org: the HGNC and VGNC resources in 2019. *Nucleic Acids Res.*, **47**, D786–D792.
10. Chang,A., Schomburg,I., Placzek,S., Jeske,L., Ulbrich,M., Xiao,M., Sensen,C.W. and Schomburg,D. (2015) BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Res.*, **43**, D439–D446.
11. Bairoch,A. (2018) The Cellosaurus, a cell-line knowledge resource. *J. Biomol. Tech. JBT*, **29**, 25–38.
12. Kersey,P.J., Allen,J.E., Allot,A., Barba,M., Boddu,S., Bolt,B.J., Carvalho-Silva,D., Christensen,M., Davis,P., Grabmueller,C. *et al.* (2018) Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.*, **46**, D802–D808.
13. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
14. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
15. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nussbaum,C., Myers,R.M., Brown,M. and Li,W. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
16. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
17. Quinlan,A.R. (2014) BEDTools: The Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinforma.*, **47**, doi:10.1002/0471250953.bi1112s47.
18. Mendoza-Parra,M.A., Saleem,M.-A.M., Blum,M., Cholley,P.-E. and Gronemeyer,H. (2016) NGS-QC Generator: A quality control system for ChIP-Seq and related deep sequencing-generated datasets. *Methods Mol. Biol.*, **1418**, 243–265.
19. Marinov,G.K., Kundaje,A., Park,P.J. and Wold,B.J. (2014) Large-scale quality analysis of published ChIP-seq data. *G3*, **4**, 209–223.
20. Chen,D., Yan,W., Fu,L.-Y. and Kaufmann,K. (2018) Architecture of gene regulatory networks controlling flower development in Arabidopsis thaliana. *Nat. Commun.*, **9**, 4534.
21. Bannister,A.J. and Kouzarides,T. (2011) Regulation of chromatin by histone modifications. *Cell Res.*, **21**, 381–395.
22. Venkatesh,S. and Workman,J.L. (2015) Histone exchange, chromatin structure and the regulation of transcription. *Nat. Rev. Mol. Cell Biol.*, **16**, 178–189.
23. Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.
24. Roadmap Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
25. Cunningham,F., Achuthan,P., Akanni,W., Allen,J., Amode,M.R., Armean,I.M., Bennett,R., Bhai,J., Billis,K., Boddu,S. *et al.* (2019) Ensembl 2019. *Nucleic Acids Res.*, **47**, D745–D751.
26. Haeussler,M., Zweig,A.S., Tyner,C., Speir,M.L., Rosenbloom,K.R., Raney,B.J., Lee,C.M., Lee,B.T., Hinrichs,A.S., Gonzalez,J.N. *et al.*

(2019) The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.*, **47**, D853–D858.
27. Raney,B.J., Dreszer,T.R., Barber,G.P., Clawson,H., Fujita,P.A., Wang,T., Nguyen,N., Paten,B., Zweig,A.S., Karolchik,D. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.
28. Fornes,O., Castro-Mondragon,J.A., Khan,A., Lee,R. van der, Zhang,X., Richmond,P.A., Modi,B.P., Correard,S., Gheorghe,M., Baranašić,D. *et al.* (2019) JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, doi:10.1093/nar/gkz1001.

SUPPLEMENTAL INFORMATION:

ReMap 2020: A database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments

Supplemental tables

Table S1 Overview of the content growth in ReMap 2020 compared to 2018 and 2015.

	ReMap 2015 Homo sapiens TRs	ReMap 2018 Homo sapiens TRs	ReMap 2018 Homo sapiens TRs	ReMap 2020 Arabidopsis thaliana TRs	ReMap 2020 Arabidopsis thaliana Histones
Experiment type	ChIP-seq	ChIP-seq	ChIP-seq ChIP-exo	ChIP-seq DAP-seq	ChIP-seq
Total targets	237	486	1,135	372	33
Datasets (selected / processed)	396 / 668	2,818 / 3,180	5,798 / 6,498	509 / 1,050	286 /358
Total peaks (million)	13	80	165	2.6	4.5
Merged peaks (million)	8.8	35.5	76	1.8	1.3
CRMs (million)	none	1.6	1.7	50,993	none

Table S2 Rejected Homo sapiens datasets.

In Excel spreadsheet

Table S3 Evaluation of Databases of ChIP-seq analyses.

We evaluated ChIP-seq databases similar to ReMap from a Homo sapiens and Arabidopsis thaliana perspective, using the comparison table 1 from GTRD as a starting point (Yevshin et al. 2017). *Datasets may not be defined similarly in different databases, some may define them as "experiments" or representing a single sample. Definitions differs, thus datasets numbers cannot be compared.

Database	**ChIPBase** http://rna.sysu.edu.cn/chipbase/	**ReMap** http://remap.cisreg.eu	**Cistrome** http://dc2.cistrome.org/#/	**ChIP-Atlas** http://chip-atlas.org/	**GTRD** http://gtrd.biouml.org/
Release dates	2013, 2017	2015, 2018, 2020	2017	2018	2017, 2019
Homo sapiens assemblies	hg38?	hg38, hg19	hg38	hg19	hg38
Homo sapiens datasets*	3549	5801	13976	6869	10675
Homo sapiens TFs	480 TFs	960 TFs and non-TFs	260 TFs and non-TFs	not stated	1158 TFs
Arabidopsis	none	tair10	none	none	tair10

Thaliana assemblies					
Arabidopsis thaliana datasets*	none	509	none	none	424
Arabidopsis thaliana TFs	none	342 TFs and non-TFs	none	none	120 TFs
Arabidopsis thaliana Histones	none	38	none	none	none
Sources	GEO, ENCODE	GEO, ARRAYEXPRESS, ENA, ENCODE	GEO, ENCODE, SRA, ENA	GEO, ENCODE, SRA, ENA, RoadMap	GEO, SRA, ENCODE
QC filtering	no	yes	yes	no	no
Manual annotation	no	yes	no	no	no
Peak callers	Multiple (import)	MACS2	MACS2	MACS2	MACS, SISSRs, GEM, PICS
Atlas provided	no	yes	no	no	yes
Non-Redundant analysis	no	yes	no	no	metacluster
Clustering analyses	no	CRMs	no	no	metacluster
Uniform data processing	no (import)	yes	yes	yes	yes
Displays, Genome browser	Self developed	UCSC, ENSEMBL, Track Hub, IGV	UCSC	IGV	Self developed
Annotation tool	no	yes	no	no	no
Restful API	no	yes	no	no	no
Standard format	no	BED	BED	BED	no
Bulk Download	no	yes	yes	yes	yes

Table S4 Genome Coverage of ReMap 2018 and ReMap 2020.

For ReMap 2018 and 2020, the genomic coverage is computed for peaks overlapping the regions at a depth of 1 to 4 peaks, 5 to 15 peaks, >15 peaks.

	ReMap 2018 Homo sapiens TR				**ReMap 2020 Homo sapiens TR**			
	Depth 1-4	Depth 5-15	Depth >15	Total	Depth 1-4	Depth 5-15	Depth >15	Total
Percent	26.7	11.1	8.3	46.1	31.4	17.6	16.7	65.7
Bases (Mb)	823	344	257	1.4Gb	969	543	516	2Gb

	ReMap 2020 Arabidopsis thaliana TR				**ReMap 2020 Arabidopsis thaliana Histones**			
	Depth 1-4	Depth 5-15	Depth >15	Total	Depth 1-4	Depth 5-15	Depth >15	Total
Percent	35.6	21.6	19.0	76.2	9.6	15.1	65.1	89.9
Bases (Mb)	42.6	25.8	22.8	91.2	11.4	18.1	78	107.5

Table S5 List of ReMap 2020 datasets.

In Excel spreadsheet

Table S6 Modified name of transcriptional regulator and cell line.

In Excel spreadsheet

Table S7 Rejected datasets Arabidopsis thaliana.

In Excel spreadsheet

Table S8 List of ReMap 2020 Arabidopsis thaliana TF datasets.

In Excel spreadsheet

Table S9 Modified name of Arabidopsis thaliana transcriptional regulator and Biotype.

In Excel spreadsheet

Table S10 Rejected Histone datasets Arabidopsis thaliana analyses.

In Excel spreadsheet

Table S11 List of ReMap 2020 Arabidopsis thaliana Histones datasets.

In Excel spreadsheet

Table S12 Modified name of Arabidopsis thaliana Histone, TF and Biotype.

In Excel spreadsheet

Supplemental figures

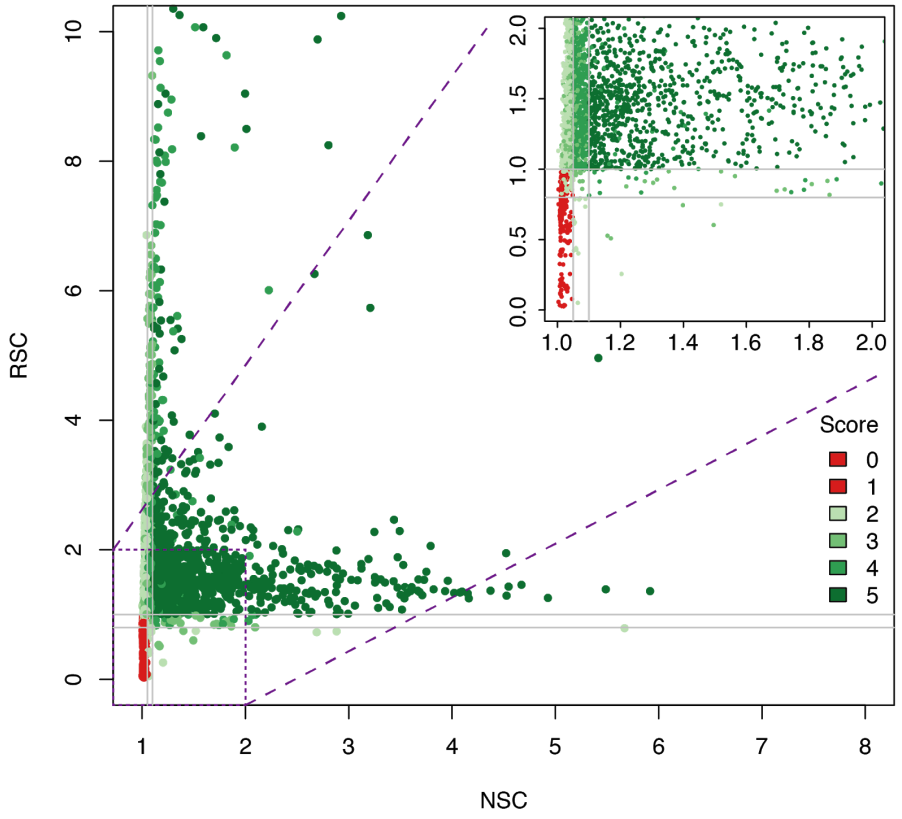


Figure S1 Quality assessment of Homo sapiens datasets.

Each analysed dataset (n=6,498) is represented by a colored dot according to its assigned score. To assess the quality of public datasets, we computed a score based on the cross-correlation and the FRiP (fraction of reads in peaks) metrics developed by the ENCODE Consortium and the phantompeak tools. This score is computed as follow. Two thresholds based on ENCODE studies were defined for each of the two cross-correlation ratios (Grey lines; Normalized Strand Coefficient: 1.05 and 1.10 on the x-axis; Relative Strand Coefficient: 0.8 and 1.0 on the y-axis, see M&M for details). A basal score ranging from 0 to 4 was assigned to each dataset corresponding to the number of thresholds it exceeds for NSC and RSC (2 thresholds for each score). Finally, this basal score was incremented by one if the FRiP is equal or higher than 1%. We observe that datasets having a minimum score of 2 exceed at least one threshold of RSC or NSC, which are both scores independent of peak calling procedures. Thus for our analyses, datasets having a score less than or equal to 1 (red dots), as well as datasets with fewer than 100 identified peaks were discarded for further downstream analyses. Reds dots within accepted thresholds are datasets with less than 100 peaks. Datasets with scores greater to 1 were kept for further analyses (green dots, n=5,801).
Score = 0 : the dataset has NSC and RSC outside thresholds
Score = 1 : the dataset has one (RSC or NSC) above 1st threshold
Score = 2 : : the dataset has one (RSC or NSC) above optimum threshold, OR two RSC or NSC above 1st threshold
Score = 3 : the dataset has one (RSC or NSC) above optimum threshold, and one RSC or NSC above 1st threshold

Score = 4 : the dataset has two (RSC + NSC) above optimum threshold
Score = 5 the dataset has two (RSC + NSC) above optimum threshold + 1 extra point if the FRiP is equal or higher that 1% (see our M&M)
In short, we give a score from 0 to 4 according to NSC and RSC, and then add 1 point according to FRiP. Datasets with a score below 2 are discarded.

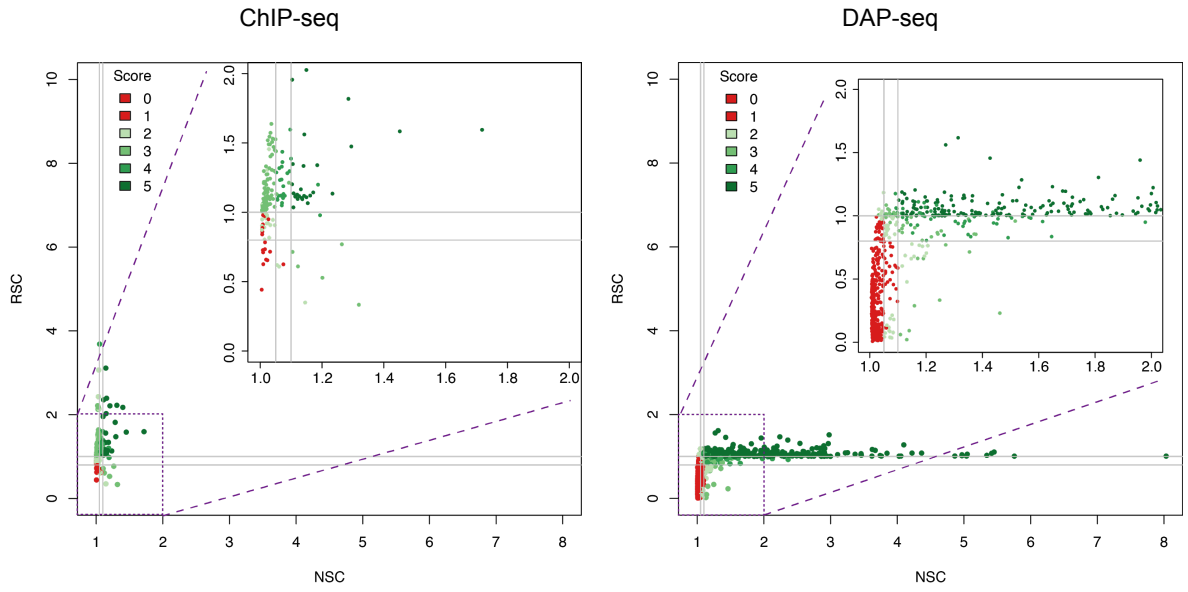


Figure S2 Quality assessment of Arabidopsis thaliana TRs datasets (ChIP-seq, DAP-seq).

Each analysed dataset (n=203 for ChIP-seq, n=847 for DAP-seq) is represented by a colored dot according to its assigned score. To assess the quality of public datasets, we computed a score based on the cross-correlation and the FRiP (fraction of reads in peaks) metrics developed by the ENCODE Consortium and the phantompeak tools. This score is described in figure S1.

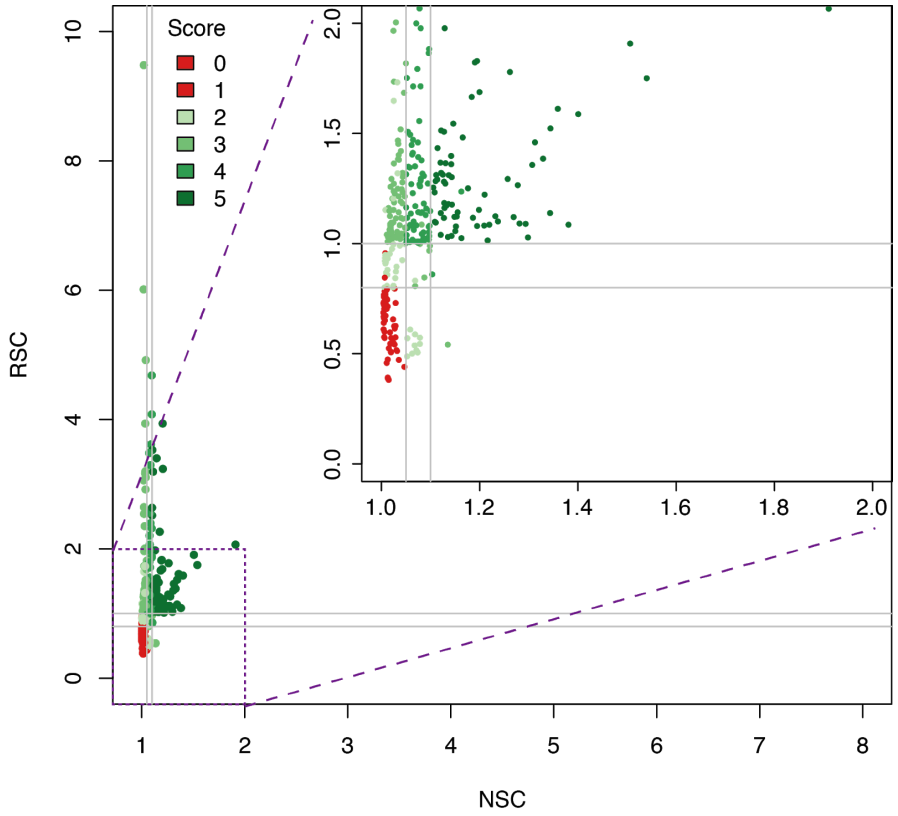


Figure S3 Quality assessment of Arabidopsis thaliana Histones datasets.

Each analysed dataset (n=358) is represented by a colored dot according to its assigned score. To assess the quality of public datasets, we computed a score based on the cross-correlation and the FRiP (fraction of reads in peaks) metrics developed by the ENCODE Consortium and the phantompeak tools. This score is described in figure S1.

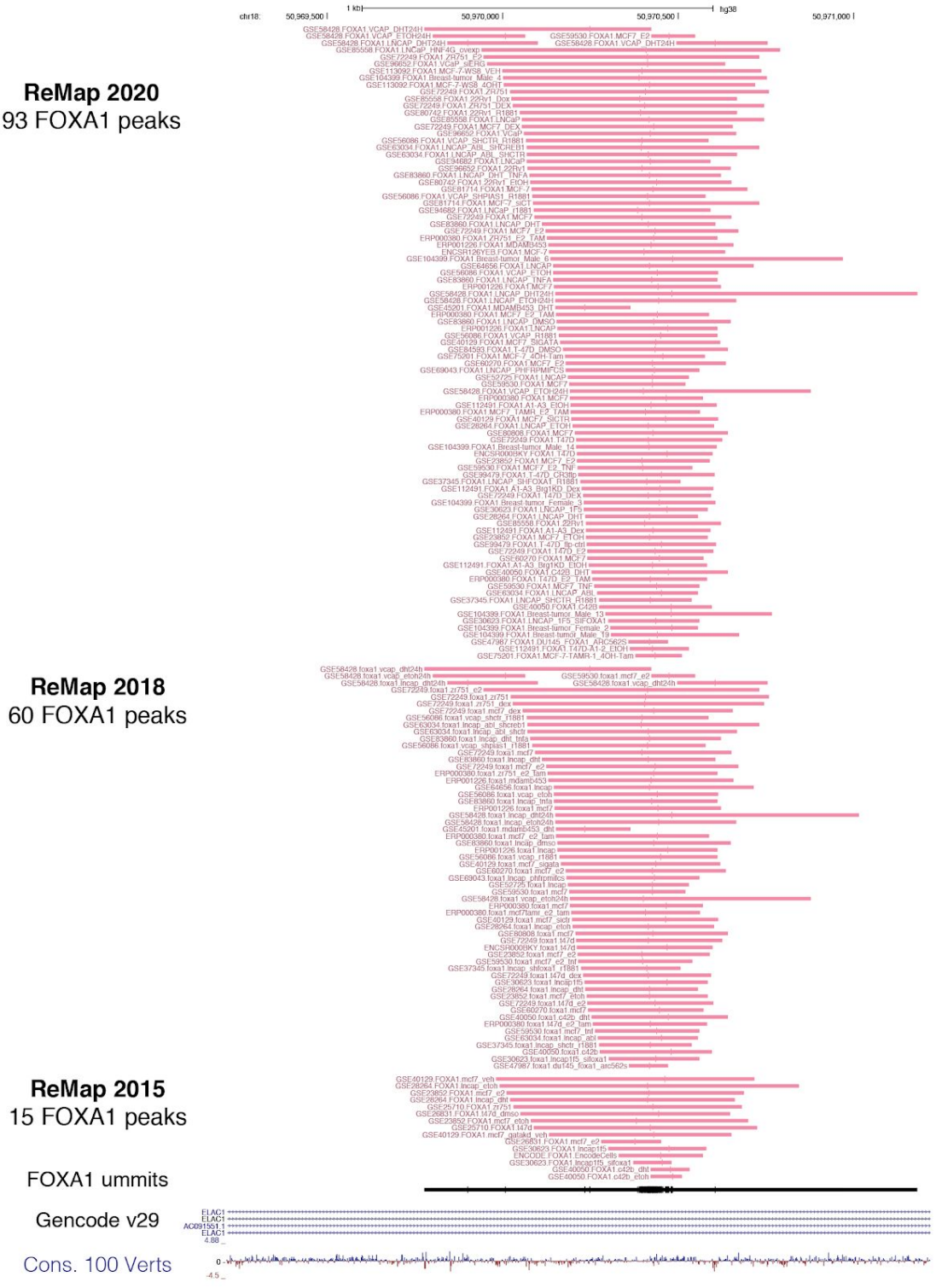


Figure S4 FOXA1 ChIP-seq peaks pattern evolution across ReMap versions.

The detailed view of the FOXA1 peaks present in ReMap 2020 (93 peaks) is compared to the FOXA1 peaks in ReMap 2018 (60 peaks) and 2015 (15 peaks). It shows FOXA1 binding pattern evolution across ReMap versions found at the genomic location chr18:50,969,638-50,970,931 in the first intron of the ELAC1 gene. Additionally, it can be noted that the peak summits (vertical bars) of each peak aggregate closely from each other, defining precisely the DNA binding location. Those aggregations of the 93 FOXA1 summits are an illustration of what is globally observed for peaks of different TFs across the genome.

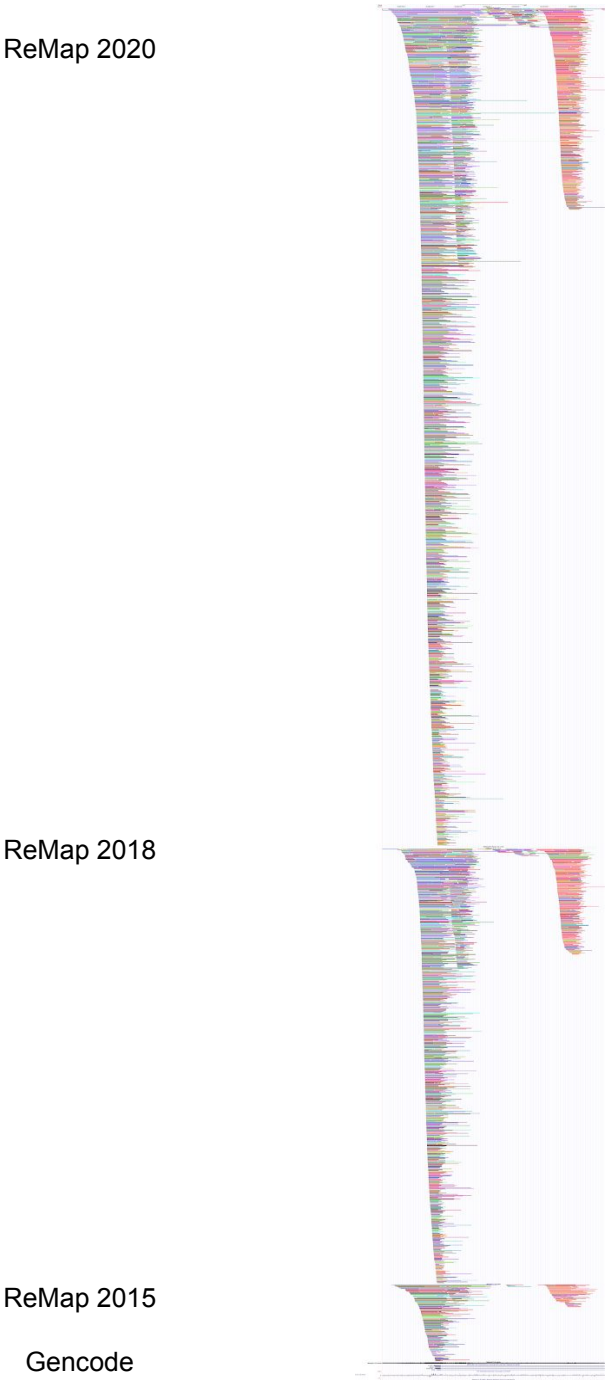


Figure S5 ReMap 2020 visualization with data on UCSC Genome Browser.

This is a non "*compacted*" view of the ReMap atlas on UCSC Genome Browser. The ReMap catalogue can be viewed through the Track hubs on both Ensembl and UCSC Genome Browser. Public sessions have been also created on mirrors of UCSC Genome Browser allowing the display of larger regions and fester rendering of the data.

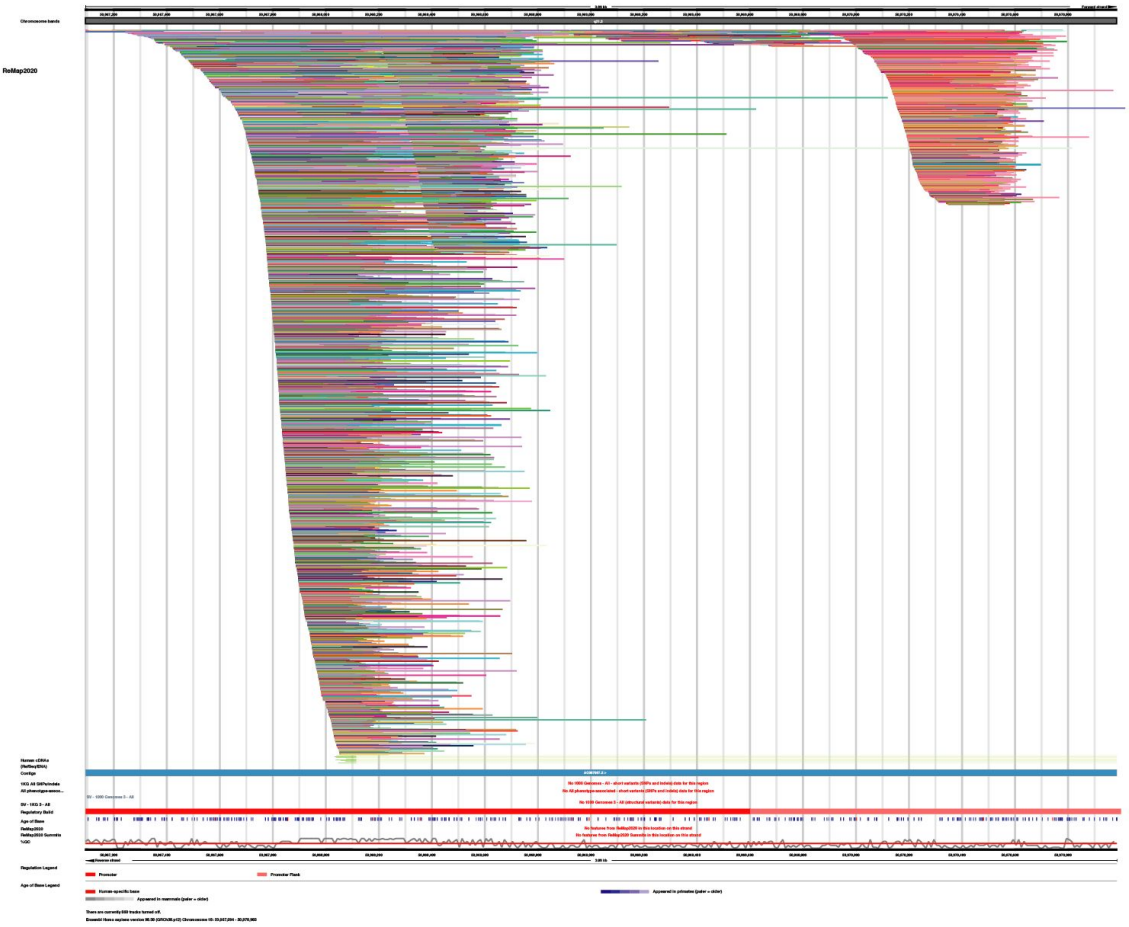


Figure S6 ReMap 2020 visualization with Track Hub on Ensembl Genome Browser.

The ReMap 2020 catalogue can be viewed through a public track hub. Here is a screenshot example of the ReMap 2020 track on the Ensembl Genome Browser. Users may choose to switch on or off any of the Ensembl track as desired, and or change the rendering of the tracks (Compact, Full, Stacked, etc…).

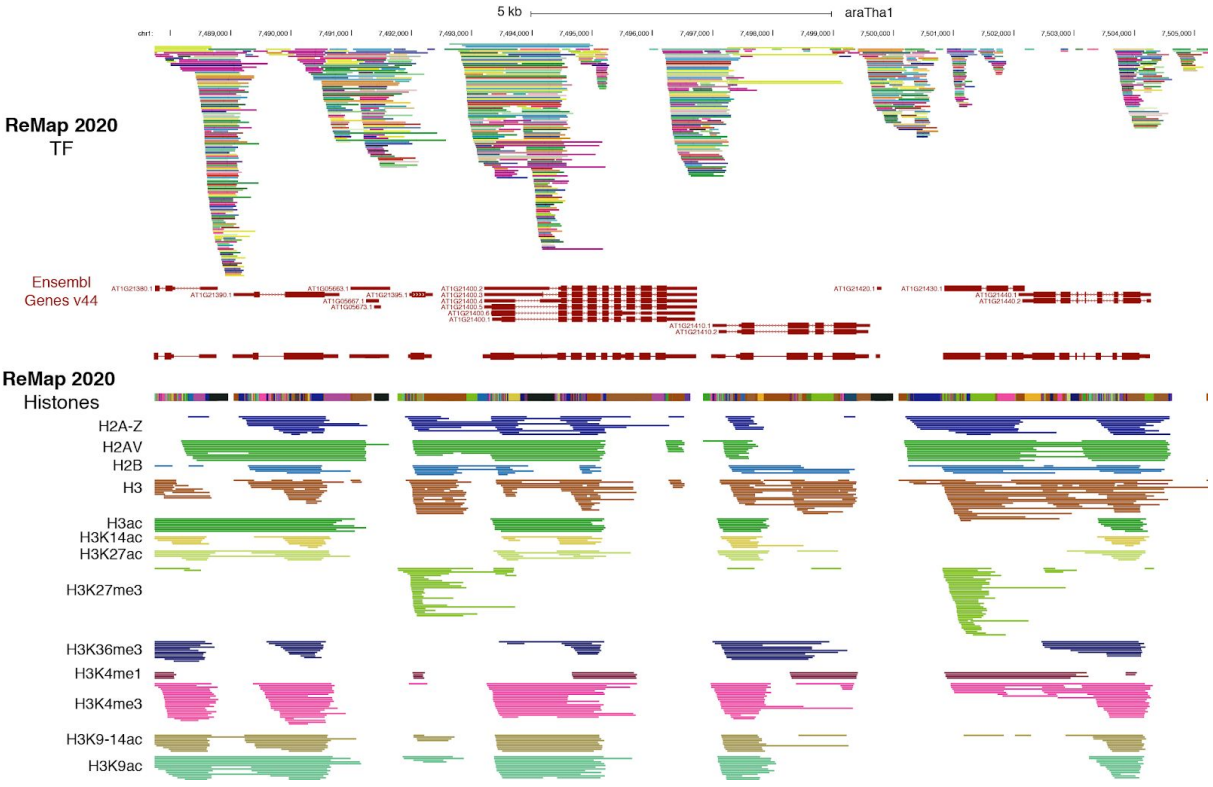


Figure S7 ReMap 2020 Arabidopsis thaliana histone modifications on UCSC Genome Browser.

The ReMap 2020 Arabidopsis thaliana Histone modifications catalogue can be viewed through a public track hub. Here is a screenshot example of the track on the USCS Genome Browser with a selected number of Histone tracks (n=13/38 total). Users may choose to switch on or off any of the Histone track as desired, and or change the rendering of the tracks (Compact, Full, Stacked, etc…).

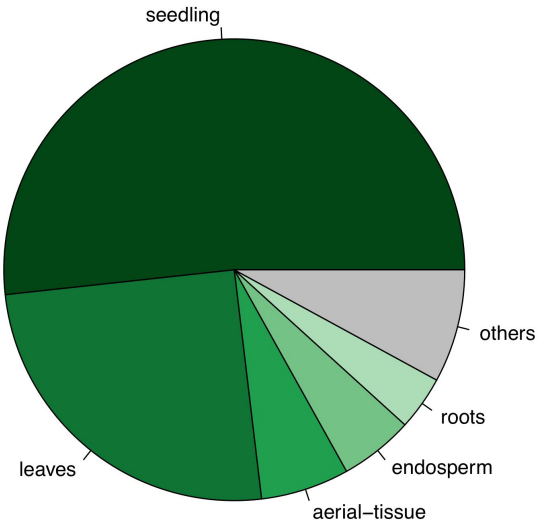


Figure S8 ReMap 2020 Arabidopsis thaliana datasets numbers per biotypes.

Repartition of datasets for the top 5 biotypes across the ReMap 2020 Arabidopsis thaliana histone catalog. The top two most represented dataset are seedling and leaves.

E. Conclusion

1. Résultats

a. Comparaison avec ReMap2018

Pour la mise à jour de ReMap2020 chez l'homme, nous avons traité 3424 nouvelles expériences ChIP-seq provenant d'ENCODE et de GEO. A ce chiffre, nous pouvons ajouter 222 expériences ChIP-exo traitées grâce au nouveau workflow de ReMap. Après contrôle qualité, nous gardons 2969 ChIP-seq et 208 ChIP-exo, ce qui représente 82% des expériences traitées. Ces chiffres sont comparables à ceux de ReMap 2018.

Au final, nous sommes passés de 2829 expériences de fixation de protéines à l'ADN à 5798 pour le catalogue ReMap2020 humain. Le catalogue présente désormais 164 millions (M) de pics pour 1135 protéines régulatrices de la transcription différentes et 602 différentes lignées cellulaires et tissus ce qui représente une multiplication par deux du nombre de pics (80M en 2018), de protéines régulatrices de la transcription (485 en 2018) et de lignées cellulaire et tissus (346 en 2018). Il est intéressant de noter que nous sommes passés de 1,6M de CRE en 2018 à 1,7M. Le nombre de CRE semble se stabiliser.

ReMap 2020 possède aussi le premier catalogue d'annotations de régions régulatrices et de marques d'histones chez *A. thaliana*. Le catalogue de régions régulatrices est constitué grâce au traitement de 204 expériences de ChIP-seq et de 848 DAP-seq. Au final, nous ne gardons que 509 expériences afin de composer un catalogue de 2,6M de régions de fixations de protéines régulatrices de la transcription. Nous avons des informations pour 372 protéines régulatrices de la transcription différentes, pour 20 biotypes et 12 écotypes différents. Nous avons donc pu annoter 50 000 CRE chez *A. thaliana* ce qui couvre environ 40% du génome.

Le catalogue de modification d'histone provient du retraitement de 358 expériences de ChIP-seq, dont 286 de ces expériences passent le contrôle qualité ce qui a permis de créer un catalogue de 4,5M de régions pour 33 marques d'histones. Ce catalogue comprend 8 écotypes et 14 biotypes.

Les résultats sont présentés dans la figure II.4.

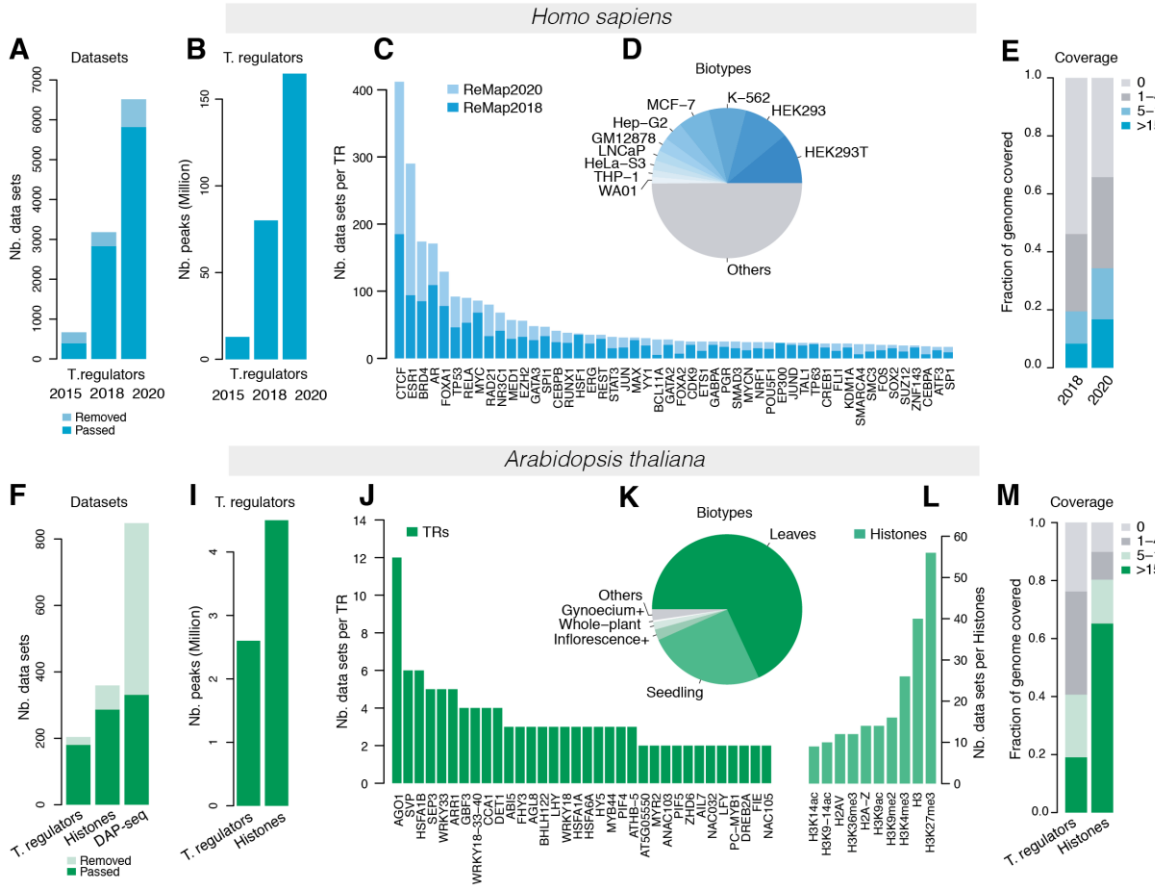


Figure II.4 : Aperçu des données disponibles pour la mise à jour ReMap2020.

F. Projet AtyPeak

Le catalogue ReMap2020 couvre 34% du génome avec 5 pics ou plus et 65% du génome avec un pic ou plus. Ce résultat n'est pas surprenant car nous intégrons des données très hétérogènes au niveau du protocole, du type de protéines et du biotype. Cette particularité n'explique pas, à elle seule, ce résultat. En effet les expériences ChIP-seq produisent des faux positifs[173–175]. Cette erreur est multipliée par le nombre d'expériences ChIP-seq dans ReMap d'où notre choix d'intégrer des données issues d'un grand nombre d'expériences biologiques, afin de ne conserver que les CREs confirmés par plusieurs expériences indépendantes.

L'annotation actuelle des CREs ReMap se base uniquement sur l'information de position des pics non redondants. De fait, nous perdons des informations sur la nature de ces pics (protéine concernée, lignée cellulaire) et sur la redondance des pics issus d'expériences similaires. Par exemple, on dénombre 129 expériences ayant pour cible le facteur de transcription FOXA1. Un pic confirmé par ces 129 expériences a plus de probabilité d'être un vrai positif qu'un pic ne recoupant aucun autre pic de FOXA1.

Afin d'affiner l'annotation des CREs et d'identifier les pics inattendus nous avons décidé de développer une méthode d'annotation de chaque pic pour chaque lignée cellulaire en fonction de sa typicalité. Il s'agit donc d'identifier, pour une même protéine d'une même cellule, les pics dont les positions sont corrélées entre elles. Cette corrélation est étendue aux pics concernant différentes protéines. Les protéines régulatrices collaborent entre elles (I.A.4 Coopération entre protéines régulatrices). On s'attend donc à ce que les pics concernant des protéines collaborant entre elles soient colocalisés, donc corrèlent au niveau de leur position. Cette corrélation peut être calculée par des statistiques classiques telles que la corrélation de Pearson. L'inconvénient de cette méthode est qu'elle permet uniquement de comparer les éléments deux par deux. De plus, elle détermine seulement l'absence ou la présence de pics. La lignée cellulaire d'origine n'est pas prise en compte. Or cette information est extrêmement importante dans la mesure où la fixation des TFs varie en fonction de la cellule et de son environnement. Pour toutes ces raisons, nous nous sommes intéressés à des méthodes de machine learning (ML).

Au sein du laboratoire nous collaborons avec d'autres chercheurs développant des méthodes de deep learning (DL) appliquées à la biologie afin d'élaborer cette nouvelle approche. Le DL est approprié car nous voulons analyser une grande quantité de données. Nous utilisons des tenseurs pour représenter l'information des pics (position, protéines, expérience d'origine ou dataset, figure II.5). L'utilisation de tenseurs sont des matrices à n dimensions, permettant d'ajouter autant d'informations qu'il est nécessaire. Chaque tenseur contient les pics de fixation pour un CRE, et a pour dimensions la position des pics, les différents datasets et la protéine (Figure II.5).

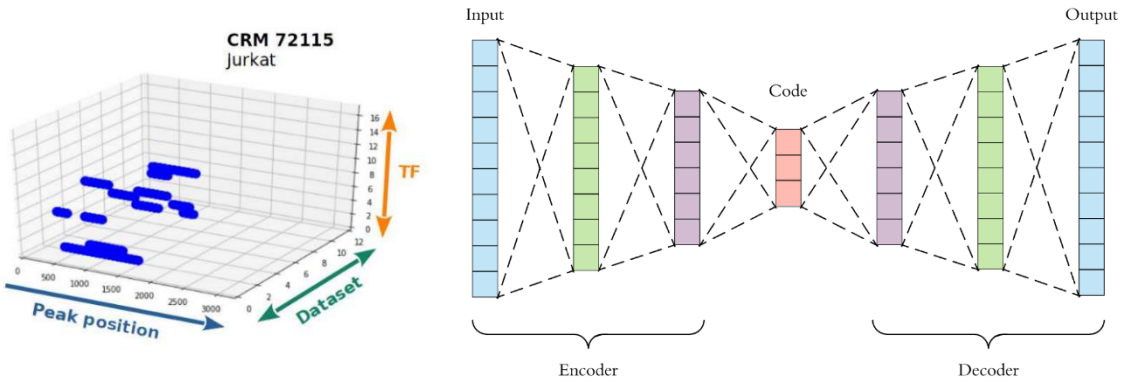


Figure II.5 : Représentation schématique des tenseurs et autoencodeurs.
La première figure montre la représentation des tenseurs pour un CRE (CRM). La deuxième représente l'organisation des couches réseaux de neurones d'un autoencodeur. Chaque neurone apprend une information simple. Le budget d'information détermine la quantité d'information pouvant être conservé. A chaque étape de l'encodage, les neurones ne contenant pas d'informations permettant la reconstitution de l'information d'origine sont enlevés selon le budget allouée. Le code contient l'information minimale. Il doit être décodé afin de se rapprocher de l'information de départ.

Figure II.5: towardsdatascience.com/generating-images-with-autoencoders-77fd3a8dd368

Ces tenseurs sont ensuite analysés par des autoencodeurs. Ces méthodes réalisent des compressions avec perte, suivies de reconstitution de l'information (Figure II.5). Leur but est de réduire autant que possible l'information d'origine et de la reconstituer à partir de cette information minimum. L'objectif, au final, est de retrouver une information aussi proche que possible que celle d'origine. En fonction du budget d'apprentissage, les informations retenues permettent d'éliminer les relations et corrélations rares. Appliquée à nos données, cette méthode permet donc de ne conserver que les pics ayant des relations des corrélations typiques entre jeux de données. L'intérêt est, pour nous, de pouvoir comparer en même temps plusieurs expériences (n-wise) de ChIP-seq concernant plusieurs protéines pour une lignée cellulaire précise. Au final, cette méthode associe un score de typicalité pour chaque pic. L'outil permettant de réaliser ce test s'appelle atyPeak. Un exemple de résultats est présenté dans la figure II.6).

Nous validons notre approche par l'utilisation de données artificiellement générées. Ces données permettent aussi de calibrer les modèles de DL utilisés. De plus, nos avons comparé le score des pics de paires TFs collaborant *in vivo* en présence et en absence de colocalisation (Figure II.6).

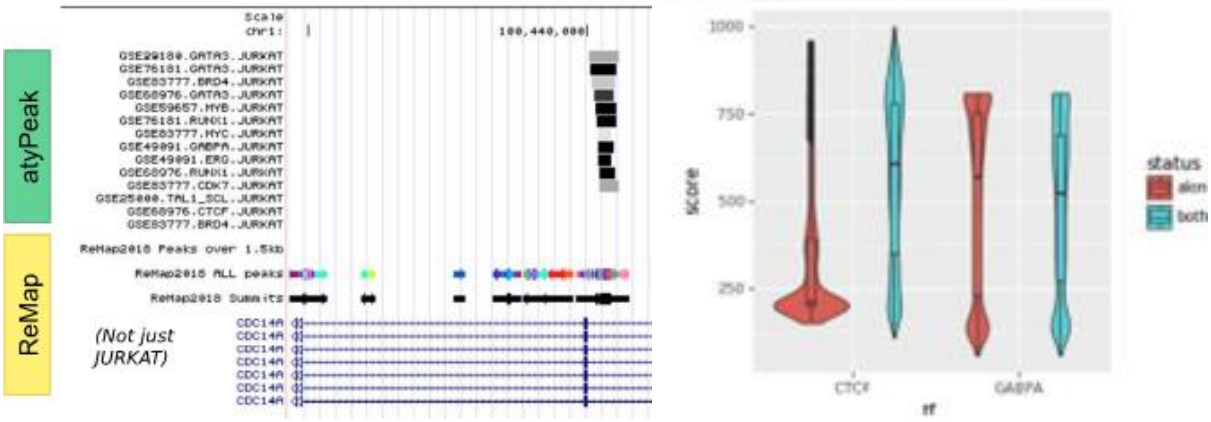


Figure II.6 : Représentation des données AtyPeak.
La première figure présente le resultat d'AtyPeak dans un navigateur de génome. Pour chaque pic le score est représenté par un niveau de gris. Plus le pic est foncé, plus le pic est corrélé avec des pics d'autres expériences concernant le même TF ou corrélé avec des pics d'autres TFs. La deuxième figure représente le score des TFs CTCF et GABPA en fonction de la présence, ou non, de l'autre TF. CTCF à une relation coopérative unilatérale avec GABPA[176] *et nous observons que le score des pics de CTCF est faible en l'absence de GABPA.*

Figure II.6 : .Ferré Q., Chèneby J., Puthier D., Caponni C. Ballester B. : A*tyPeak* : Unsupervised outlier detection in genomic peak databases using multi-view convolutional autoencoders. (Article en preparation)

III JASPAR 2018-2020

A. Introduction

L'identification des éléments régulateurs le long du génome passe par l'identification de la fixation des protéines régulatrices sur l'ADN. Les expériences biologiques permettent d'identifier ces fixations, mais sont soumises aux limites propres à l'expérimentation, à savoir, les erreurs humaines, les limites techniques et surtout le coût des expériences. Pour ces différentes raisons, des méthodes bioinformatiques ont été développées afin d'identifier les sites de fixation des TFs. De nouvelles bases de données, listant les motifs de fixation des TFs, ont contribué à faciliter la recherche de ces sites de fixation.

B. La base de données JASPAR

JASPAR, créée en 2006[177], est une base de données populaire, concernant les profils de fixation des TFs chez les Métazoaires, Plantes et Fungi. Ces profils de fixation sont modélisés par des matrices qui peuvent être converties en Matrices Poids-Position (matrice poids-position). Ces matrices sont utilisées pour scanner le génome à la recherche de sites de fixation. La base de données JASPAR est disponible dans un site dédié : http://jaspar.genereg.net.

Pour constituer son catalogue, JASPAR utilise des données issues d'expériences de capture de fixation de protéines à l'ADN[178]. Pour chaque TF, une matrice poids-position est générée à partir de listes de séquences identifiées par des expériences de fixation de protéines à l'ADN. On peut, par exemple, citer, l'évolution systématique de ligands par enrichissement exponentiel[179] (SELEX) et le ChIP-seq (Figure III.1). Les matrices poids-position issues des données ChIP-seq sont générées grâce à l'outil "peak-motifs' de RSAT[180]. Seuls les motifs proches du centre du sommet des pics de ChIP-seq[170] sont sélectionnés. Les profils redondants sont retirés.

La précision de la matrice poids-position générée dépend donc de la qualité des expériences utilisées. Dans le cas des expériences de ChIP-seq, les régions faussement capturées par des interactions transientes de l'anticorps produisent du bruit lors de la création des profils de fixation. Il en est de même, quand les régions capturées sont larges et recouvrent des séquences ne jouant pas de rôle dans la fixation des TFs. De plus, les données issues d'expériences séquencées en bulk engendrent une variabilité des résultats du fait de l'hétérogénéité des cellules (Partie III.2. Séquençage à haut débit). Toutes les données sont présentées sous le format JASPAR, TRANSFAC et MEME.

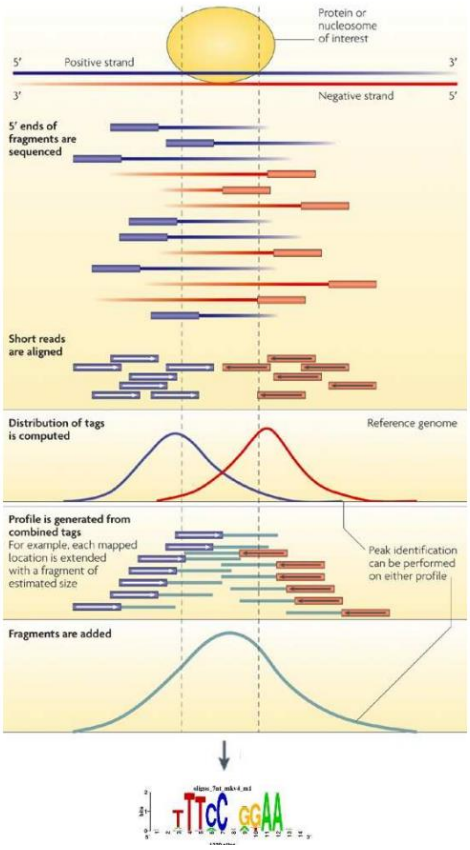


Figure III.1 *: Extraction des profils de fixation à partir de données ChIP-seq.*

1. Mise à jour 2018 et 2020

JASPAR a été mis à jour en 2018 et en 2020. La base de données propose une extension de son catalogue des sites de fixation des protéines et une refonte du site internet.

Les mises à jour de JASPAR 2018 et JASPAR 2020 utilisent les données produites par ReMap 2018 et ReMap 2020. Les données ReMap ont permis de compléter le catalogue des profils de fixation chez l'Homme ainsi que d'élargir le catalogue d'*Arabidopsis. Thaliana*. ReMap a été sélectionné car il fournit des données dont la qualité a été reconnue et validée. Toutes les données sont présentées sous plusieurs formats afin de faciliter l'interopérabilité avec des données externes. JASPAR propose les profils de fixation sous les formats les plus populaires : JASPAR, TRANSFAC et MEME.

Figure III.1: Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669–680 (2009).

Entre autres nouveautés, la version 2018 de JASPAR offre la possibilité de comparer les profils de fixation de protéines proches. En effet, les protéines ayant des domaines de fixation similaires ont souvent des préférences de fixation similaires[181,182]. JASPAR propose également une liste de prédiction des sites de fixation des différentes protéines le long du génome de l'organisme étudié. Cette liste peut être consultée sur les navigateurs de génomes tels que UCSC. Le site internet de JASPAR 2018 a été développé grâce au framework Django basé sur Python. Django est similaire à Laravel et se base sur le modèle, Modèle-vue-contrôleur. Le moteur de recherche d'informations a été amélioré afin de pouvoir filtrer avec de nouveaux critères (organisme, qualité, etc.). De plus, JASPAR 2018 propose une interface REST favorisant l'accessibilité, la consultabilité et l'interopérabilité des données (Résultats I.C. Conclusion ReMap2018).

La base de données JASPAR propose différents outils disponibles en ligne. Le service web "Profile Inference" permet de prédire le profil de fixation d'une protéine à partir de sa séquence. La méthode est basée sur la ressemblance de la séquence de la protéine entrée avec celle des protéines présentes dans la base de données. "Matrix Align" permet de comparer directement un profil de protéine à la base de données JASPAR. JASPAR propose également une liste d'outils pour aider à l'analyse de la fixation de protéines. La mise à jour propose aussi des outils simples permettant de trouver, à partir d'une séquence nucléotidique, des profils de fixation compatibles, de permuter les colonnes d'une même matrice poids-position, ou de permuter les colonnes de plusieurs matrices poids-position ainsi que de créer des matrices poids-position aléatoires.

La mise à jour JASPAR 2020 propose une nouvelle extension de son catalogue ainsi qu'un nouveau catalogue de profils de fixation de protéine dit non-validé par des sources scientifiques externes ("unvalidated"). Ces profils ont passé le seuil de sélection et sont disponibles sur le site internet. L'outil de classification hiérarchique a, lui aussi, été mis à jour.

a : www.djangoproject.com

D260–D266 *Nucleic Acids Research, 2018, Vol. 46, Database issue* *Published online 13 November 2017*
doi: 10.1093/nar/gkx1126

JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework

Aziz Khan[1,†], Oriol Fornes[2,†], Arnaud Stigliani[3,†], Marius Gheorghe[1], Jaime A. Castro-Mondragon[1], Robin van der Lee[2], Adrien Bessy[3], Jeanne Chèneby[4,5], Shubhada R. Kulkarni[6,7,8], Ge Tan[9,10], Damir Baranasic[9,10], David J. Arenillas[2], Albin Sandelin[11,*], Klaas Vandepoele[6,7,8], Boris Lenhard[9,10,12,*], Benoît Ballester[4,5], Wyeth W. Wasserman[2,*], François Parcy[3] and Anthony Mathelier[1,13,*]

[1]Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway, [2]Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, BC Children's Hospital Research Institute, University of British Columbia, 950 28th Ave W, Vancouver, BC V5Z 4H4, Canada, [3]University of Grenoble Alpes, CNRS, CEA, INRA, BIG-LPCV, 38000 Grenoble, France, [4]INSERM, UMR1090 TAGC, Marseille, F-13288, France, [5]Aix-Marseille Université, UMR1090 TAGC, Marseille, F-13288, France, [6]Ghent University, Department of Plant Biotechnology and Bioinformatics, Technologiepark 927, 9052 Ghent, Belgium, [7]VIB Center for Plant Systems Biology, Technologiepark 927, 9052 Ghent, Belgium, [8]Bioinformatics Institute Ghent, Ghent University, Technologiepark 927, 9052 Ghent, Belgium, [9]Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London W12 0NN, UK, [10]Computational Regulatory Genomics, MRC London Institute of Medical Sciences, London W12 0NN, UK, [11]The Bioinformatics Centre, Department of Biology and Biotech Research & Innovation Centre, University of Copenhagen, DK2200 Copenhagen N, Denmark, [12]Sars International Centre for Marine Molecular Biology, University of Bergen, N-5008 Bergen, Norway and [13]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, 0310 Oslo, Norway

Received September 25, 2017; Revised October 17, 2017; Editorial Decision October 18, 2017; Accepted October 27, 2017

ABSTRACT

JASPAR (http://jaspar.genereg.net) is an open-access database of curated, non-redundant transcription factor (TF)-binding profiles stored as position frequency matrices (PFMs) and TF flexible models (TFFMs) for TFs across multiple species in six taxonomic groups. In the 2018 release of JASPAR, the CORE collection has been expanded with 322 new PFMs (60 for vertebrates and 262 for plants) and 33 PFMs were updated (24 for vertebrates, 8 for plants and 1 for insects). These new profiles represent a 30% expansion compared to the 2016 release. In addition, we have introduced 316 TFFMs (95 for vertebrates, 218 for plants and 3 for insects). This release incorporates clusters of similar PFMs in each taxon and each TF class per taxon. The JASPAR 2018 CORE vertebrate collection of PFMs was used to predict TF-binding sites in the human genome. The predictions are made available to the scientific community through a UCSC Genome Browser track data hub. Finally, this update comes with a new web framework with an interactive and responsive user-interface, along with new features. All the underlying data can be retrieved programmatically using a RESTful API and through the JASPAR 2018 R/Bioconductor package.

INTRODUCTION

Transcription factors (TFs) are sequence-specific DNA-binding proteins involved in the transcriptional regulation of gene expression (1). TFs bind to DNA through their DNA-binding domain(s) (DBDs), which are used for TF classification (2). DNA regions at which TFs bind are defined as TF-binding sites (TFBSs) and can be identified

*To whom correspondence should be addressed. Tel: +47 228 40 561; Email: anthony.mathelier@ncmm.uio.no
Correspondence may also be addressed to Albin Sandelin. Tel: +45 2245 6668; Fax: +45 3532 2128; Email: albin@binf.ku.dk
Correspondence may also be addressed to Boris Lenhard. Tel: +44 20 8383 8353; Email: b.lenhard@imperial.ac.uk
Correspondence may also be addressed to Wyeth W. Wasserman. Tel: +1 604 875 3812; Fax: +1 604 875 3840; Email: wyeth@cmmt.ubc.ca
†These authors contributed equally to the paper as first authors.

Table 1. Overview of the growth of the number of PFMs in the JASPAR 2018 CORE collection compared to the JASPAR 2016 CORE collection

Taxonomic group	Non-redundant PFMs in JASPAR 2016	New non-redundant PFMs in JASPAR 2018	Updated PFMs in JASPAR 2018	Total PFMs (non-redundant) in JASPAR 2018	Total PFMs (all versions) in JASPAR 2018
Vertebrates	519	60	24	579	719
Plants	227	262	8	489	501
Insects	133	0	1	133	140
Nematodes	26	0	0	26	26
Fungi	176	0	0	176	177
Urochordata	1	0	0	1	1
Total	**1082**	**322**	**33**	**1404**	**1564**

in vivo by methods such as chromatin immunoprecipitation (ChIP) or *in vitro* by methods based on binding of large pools of DNA fragments (e.g. Systematic evolution of ligands by exponential enrichment (SELEX) or protein-binding microarrays (PBM)) (reviewed in (3)). Analysis of TFBSs for a given TF provides models for its specific DNA-binding preferences, which in turn can be used to predict TFBSs in DNA sequences (4). This is important as experiments can only identify TFBSs that are bound in the cell and state analyzed.

The computational representation of TF binding preferences has evolved over the years, from simple consensus sequences to position frequency matrices (PFMs). A PFM summarizes experimentally determined DNA sequences bound by an individual TF by counting the number of occurrences of each nucleotide at each position within aligned TFBSs. Such matrices can be converted into position weight matrices (PWMs), also known as position-specific scoring matrices, which are probabilistic models that can be used to predict TFBSs in DNA sequences (reviewed in (5)).

PFMs/PWMs have been the standard models for describing binding preferences of TFs for many years. The JASPAR database is among the most popular and longest maintained databases for PFMs and a standard resource in the field. In particular, the JASPAR CORE collection of the database, which is the most used, stores non-redundant TF binding profiles, providing a single representative DNA binding model per TF decided by expert curators. Exceptionally, multiple TF-binding profiles are associated to a TF when it is known to interact with DNA with multiple distinct sequence preferences, due to differential splicing for example (6,7). JASPAR was created and persists under three guiding principles: (i) unrestricted open-access; (ii) manual curation and non-redundancy of profiles; and (iii) ease-of-use. The 2016 release of the JASPAR CORE collection stored 1082 non-redundant and manually curated TF-binding profiles as PFMs for TFs from six different taxonomic groups (vertebrates, plants, insects, nematodes, fungi and urochordata) (8).

An intrinsic limitation to PFMs/PWMs is that they ignore inter-nucleotide dependencies within TFBSs (9–13). TF–DNA interaction data derived from next-generation sequencing assays has improved the computational modeling of TF binding (14–19). For example, the TF flexible models (TFFMs) (14), based on first-order hidden Markov models, capture dinucleotide dependencies within TFBSs and were introduced in the 2016 release of the JASPAR database.

In this report, we describe the seventh release of JASPAR (8,20–24), which comes with a major expansion and update of the CORE collection of TF-binding profiles as PFMs and TFFMs. These models have been manually assessed by expert curators who reconciled recent high-throughput data with available literature and linked the models to the classification of their TF DBDs from TFClass (2). The CORE collection expansion is supported by a range of new functionalities and resources, including PFM clustering, genome-wide UCSC tracks of predicted TFBSs and fully redesigned user and programming interfaces.

EXPANSION AND UPDATE OF THE JASPAR CORE COLLECTION

In this 2018 release of the JASPAR database, we added 355 new PFMs for TFs from plants (270), vertebrates (84) and insects (1) to the JASPAR CORE collection (Table 1). Specifically, we added 322 PFMs (262 for plants, a 118% increase and 60 for vertebrates, an 11% increase) for TF monomers and dimers that were not previously present in JASPAR and updated 33 (8 in plants, 3% of JASPAR 2016, 24 in vertebrates, 5% of JASPAR 2016 and 1 in insects). The PFMs were manually curated using independent external literature supporting the candidate TF-binding preferences, as previously described in (23). The curated PFMs were derived from ChIP-seq (from ReMap (25) and (26–30)), DAP-seq (31), SMiLE-seq (32), PBM (33) and HT-SELEX (34) experiments. The JASPAR CORE collection now includes 1404 non-redundant PFMs (579 for vertebrates, 489 for plants, 176 for fungi, 133 for insects, 26 for nematodes and 1 for urochordata) (Table 1).

We continued with the incorporation of TFFM models, initiated in JASPAR 2016. In this release of JASPAR, we introduced 316 new TFFMs for vertebrates (95), plants (218) and *Drosophila* (3), which represents a 243% increase in the number of non-redundant TFFMs stored in the JASPAR CORE collection.

HIERARCHICAL CLUSTERING OF TF-BINDING PROFILES

While the non-redundancy of binding profiles is one of the guiding principles of JASPAR, TFs with similar DBDs often have similar binding preferences (35,36). To facilitate the exploration of similar profiles in the JASPAR CORE collection, we performed hierarchical clustering of PFMs using the RSAT matrix-clustering tool (37). Specifically, the tool was applied to PFMs in each taxon independently as

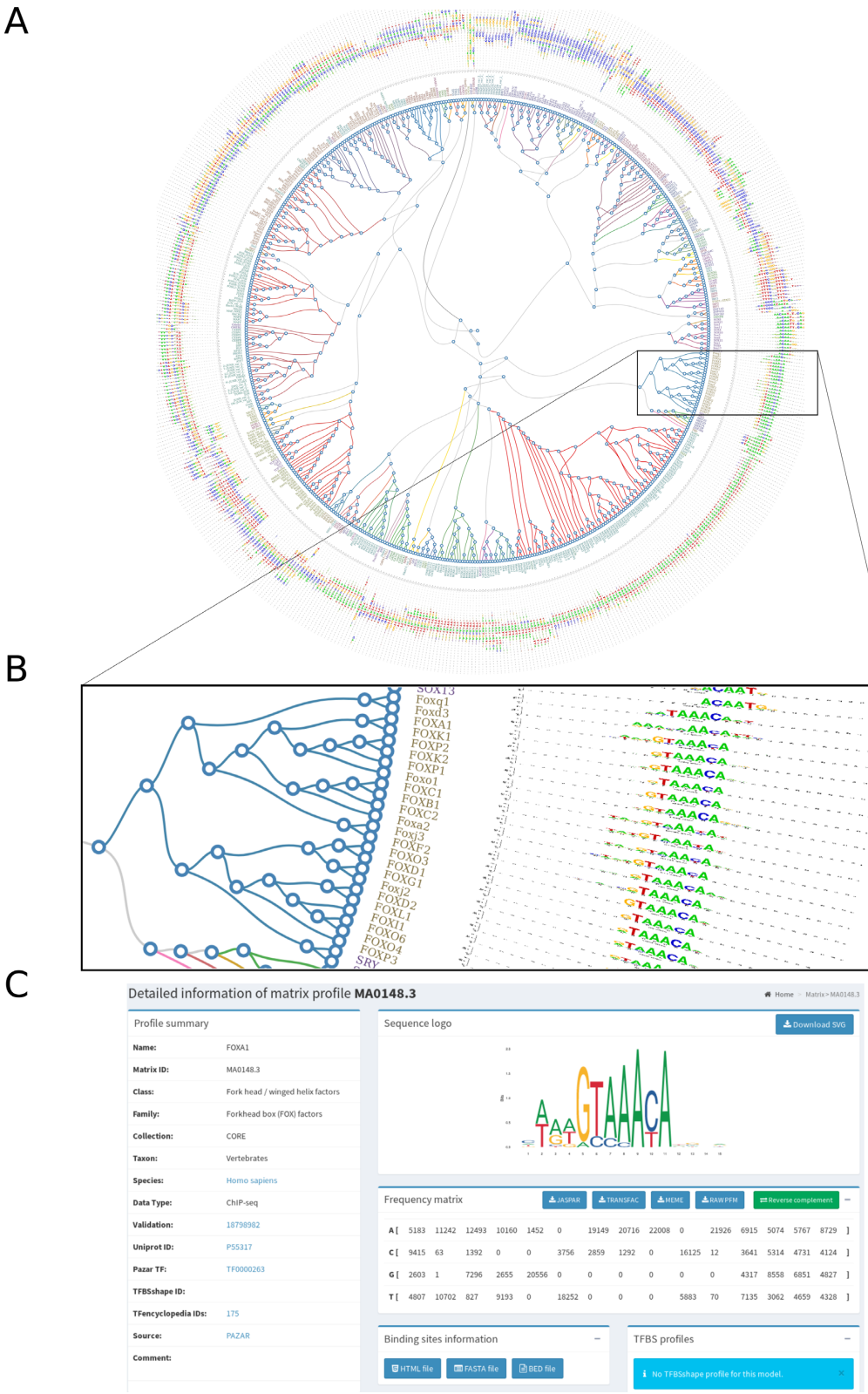


Figure 1. JASPAR PFM clustering. (**A**) Radial tree representing the clusterization of the JASPAR CORE vertebrate PFMs. (**B**) Zoom in view of the radial tree where the predicted clusters are highlighted at the branches and the TF classes are indicated with different colors at the leaves. (**C**) Clicking on a leaf in the radial tree will open a link to the corresponding motif description page on the JASPAR website (the MA0148.3 profile associated to FOXA1 is provided here as an example).

Figure 2. Overview of the JASPAR 2018 new web interface with interactive searching activity. (**A**) A quick and detailed search feature on the homepage. (**B**) A responsive table lists the searched profile(s), which can be further selected and added to the cart listed on the right panel for users to perform their own analyses. (**C**) A detailed page for the GATA3 matrix profile, which is divided into sub-panels including the profile summary, sequence logo, PFM, TF-binding information, external links, version information, ChIP-seq centrality, TFFM and other details. (**D**) The PFM for the GATA3 profile (MA0037.2) is downloaded in MEME format using the RESTful API.

well as in each TF class per taxon. The clustering results are provided as radial trees (Figure 1), which can further be explored through dedicated web pages (http://jaspar.genereg.net/matrix-clusters).

JASPAR UCSC TRACKS FOR GENOME-WIDE ANALYSES OF TFBSs

A typical application of JASPAR TF-binding profiles in gene regulation studies is the identification of TFBSs in DNA sequences for further analyses. Although, we recognize that genome-wide PWM-based predictions contain a high number false positives, we believe that they are a powerful resource for the research community in the context of a variety of genomic information, including transcription start site activity, DNA accessibility, histone marks, evolutionary conservation or *in vivo* TF binding (38–46). To facilitate such integrative analyses, we have performed TFBS predictions on the human genome using the JASPAR CORE vertebrate PFMs (see Supplementary Data for details on the computation). The predicted TFBSs are publicly available through a UCSC Genome Browser data hub (47) containing tracks for the human genome assemblies hg19 and hg38 (http://jaspar.genereg.net/genome-tracks/).

A NEW, POWERFUL AND USER-FRIENDLY WEB INTERFACE

A new web interface

The JASPAR 2018 release comes with a completely redesigned web interface that meets modern web standards. This interactive web framework is implemented using Django, a model-view-controller based web-framework for Python. We used MySQL as a backend database to store profile metadata and Bootstrap as a frontend template engine. We have greatly improved the visibility and usability of existing functionality, created easier navigation with semantic URLs, and enhanced browsing and searching. On the homepage, we provide a dynamic tour of JASPAR 2018, walking users through the main features of the new website. A video of the tour is available at http://jaspar.genereg.net/tour. The database can be browsed for individual collections by using the navigation links on the left sidebar. Moreover, it can be searched for each of the six different taxonomic groups included in the JASPAR CORE collection using the tabs available on the homepage (Figure 2). TF-binding profiles can be further filtered through the case insensitive search option available on the homepage. In addition, through the 'Advanced Options', the search criteria can be further restricted (Figure 2A). Search results are presented in a responsive and paginated table along with sequence logos of the PFMs, which can be selected for download or to perform a variety of analyses available on the right panel (Figure 2B). All information in the tables can be downloaded as comma-separated value files. Profile IDs and sequence logos can be clicked to view the detailed profile pages (Figure 2C). PFMs can be downloaded in several formats including JASPAR, TRANSFAC and MEME (Figure 2D). Furthermore, we have incorporated new features to the web interface, such as 'Add to Cart', where users can add TF profiles of interest for download or further analyses (Figure 2B). Finally, we have introduced semantic URLs to facilitate external linking to the detailed pages of individual profiles (e.g. http://jaspar.genereg.net/matrix/MA0059.1/). We have implemented a URL redirection mechanism to correctly direct the links pointing to previous JASPAR URL patterns from external resources.

RESTful API

In previous releases, the underlying data could be retrieved as flat files or by using programming language-specific modules. Associated with this release, we introduced a RESTful API to access the JASPAR database programmatically (see https://www.biorxiv.org/content/early/2017/07/06/160184 for details). The RESTful API enables programmatic access to JASPAR by most programming languages and returns data in seven widely used formats: JSON, JSONP, JASPAR, MEME, PFM, TRANSFAC and YAML. Further, it provides a browsable interface and access to the JASPAR motif inference tool for bioinformatics tool developers. The RESTful API is implemented in Python using the Django REST Framework and is freely accessible at http://jaspar.genereg.net/api/. The source code for the website and RESTful API are freely available at https://bitbucket.org/CBGR/jaspar under GPL v3 license.

CONCLUSION AND PERSPECTIVES

In this seventh release of the JASPAR database, we continue our commitment to provide the research community with high-quality, non-redundant TF-binding profiles for TFs in six taxa. As in previous releases, we have greatly expanded the number of available profiles in the database, both for PFMs and TFFMs. We also greatly improved user experience through a new easy-to-use website and a RESTful API that grants universal programmatic access to the database. Moreover, for the PFMs in the JASPAR CORE collection, we provide a hierarchical clustering and genome-wide TFBS predictions for the hg19 and hg38 human genome assemblies as UCSC tracks.

During the curation process, hundreds of PFMs were discarded because our curators failed to find any support from existing literature. As new experiments and data become available, binding preferences for these TFs will be considered for JASPAR incorporation. For instance, we re-examined data from (34) to incorporate seven previously excluded PFMs into JASPAR 2018. In the future, we would like to engage the scientific community in the curation process to increase our capacity to introduce new TF-binding profiles in JASPAR. We plan to dedicate a specific section of the website to hosting the profiles that were not introduced into JASPAR, to encourage researchers to perform experiments and/or point us to literature that our curators missed in order to support these profiles. We believe that the engagement of the scientific community to support JASPAR will further improve our capacity to expand the collection of high quality TF-binding profiles.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the scientific community for performing experimental assays of TF–DNA interactions and for publicly releasing the data. We thank Georgios Magklaras and his team for IT support. We thank José Manuel Franco for sharing the plant PBM data, Jens De Ceukeleire for help with plant ChIP-seq data processing and José Luis Villanueva-Cañas for sharing the *Drosophila* TFFMs prior to publication. We thank Rachelle Farkas for proofreading the manuscript.

FUNDING

Norwegian Research Council, Helse Sør-Øst, and University of Oslo through the Centre for Molecular Medicine Norway (NCMM) (to A.M., M.G., A.K.); Genome Canada and Canadian Institutes of Health Research (On-Target Grants) [255ONT and BOP-149430 to W.W.W., O.F., R.v.d.L., D.J.A.]; Natural Sciences and Engineering Research Council of Canada (Discovery Grant) [RGPIN-2017–06824 to W.W.W.]; Weston Brain Institute [20R74681 to O.F.]; Agence Nationale de la Recherche [ANR-10-LABX-49–01 to F.P., A.S.]; IDEX graduate schoool (to A.S.); CNRS (to A.B., F.P.); Research Foundation–Flanders Grant [G001015N to S.R.K.]; French Ministry

of Higher Education and Research (MESR) PhD Fellowship (to J.A.C.-M.); Lundbeck Foundation (to A.S.); Independent Research Fund Denmark (to A.S.); Innovation Fund Denmark (to A.S.); Elixir Denmark (to A.S.); Wellcome Trust [106954 to G.T., D.B., B.L.]; Biotechnology and Biological Sciences Research Council [BB/N023358/1 to G.T., D.B., B.L.]; Medical Research Council UK [MC_UP_1102/1 to G.T., D.B., B.L.]. The open access publication charge for this paper has been waived by Oxford University Press - *NAR* Editorial Board members are entitled to one free paper per year in recognition of their work on behalf of the journal.
Conflict of interest statement. None declared.

REFERENCES

1. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
2. Wingender,E., Schoeps,T., Haubrock,M. and Dönitz,J. (2015) TFClass: a classification of human transcription factors and their rodent orthologs. *Nucleic Acids Res.*, **43**, D97–D102.
3. Xie,Z., Hu,S., Qian,J., Blackshaw,S. and Zhu,H. (2011) Systematic characterization of protein-DNA interactions. *Cell. Mol. Life Sci.*, **68**, 1657–1668.
4. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
5. Stormo,G.D. (2013) Modeling the specificity of protein-DNA interactions. *Quant. Biol.*, **1**, 115–130.
6. Stormo,G.D. (2015) DNA motif databases and their uses. *Curr. Protoc. Bioinformatics*, **51**, 1–6.
7. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
8. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.-Y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.
9. Man,T.K. and Stormo,G.D. (2001) Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.*, **29**, 2471–2478.
10. Bulyk,M.L., Johnson,P.L.F. and Church,G.M. (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–1261.
11. Zhou,Q. and Liu,J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20**, 909–916.
12. Tomovic,A. and Oakeley,E.J. (2007) Position dependencies in transcription factor binding sites. *Bioinformatics*, **23**, 933–941.
13. Chin,F. and Leung,H.C.M. (2008) DNA motif representation with nucleotide dependency. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **5**, 110–119.
14. Mathelier,A. and Wasserman,W.W. (2013) The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, **9**, e1003214.
15. Zellers,R.G., Drewell,R.A. and Dresch,J.M. (2015) MARZ: an algorithm to combinatorially analyze gapped n-mer models of transcription factor binding. *BMC Bioinformatics*, **16**, 1–14.
16. Eggeling,R., Roos,T., Myllymäki,P. and Grosse,I. (2015) Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data. *BMC Bioinformatics*, **16**, 1–15.
17. Siebert,M. and Söding,J. (2016) Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.*, **44**, 6055–6069.
18. Mathelier,A., Xin,B., Chiu,T.-P., Yang,L., Rohs,R. and Wasserman,W.W. (2016) DNA shape features improve transcription factor binding site predictions in vivo. *Cell Syst.*, **3**, 278–286.
19. Omidi,S., Zavolan,M., Pachkov,M., Breda,J., Berger,S. and van Nimwegen,E. (2017) Automated incorporation of pairwise dependency in transcription factor binding site prediction using dinucleotide weight tensors. *PLoS Comput. Biol.*, **13**, e1005176.
20. Sandelin,A., Alkema,W., Engström,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
21. Vlieghe,D., Sandelin,A., De Bleser,P.J., Vleminckx,K., Wasserman,W.W., van Roy,F. and Lenhard,B. (2006) A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.*, **34**, D95–D97.
22. Bryne,J.C., Valen,E., Tang,M.-H.E., Marstrand,T., Winther,O., da Piedade,I., Krogh,A., Lenhard,B. and Sandelin,A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.
23. Portales-Casamar,E., Thongjuea,S., Kwon,A.T., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W.W. and Sandelin,A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D1010.
24. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C.-Y., Chou,A., Ienasescu,H. *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D1427.
25. Chèneby,J., Gheorghe,M., Artufel,M., Mathelier,A. and Ballester,B. (2017) ReMap 2018: An updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seqexperiments. *Nucleic Acids Res.*, doi:10.1093/nar/gkx1092.
26. Eveland,A.L., Goldshmidt,A., Pautler,M., Morohashi,K., Liseron-Monfils,C., Lewis,M.W., Kumari,S., Hiraga,S., Yang,F., Unger-Wallace,E. *et al.* (2014) Regulatory modules controlling maize inflorescence architecture. *Genome Res.*, **24**, 431–443.
27. Verkest,A., Abeel,T., Heyndrickx,K.S., Van Leene,J., Lanz,C., Van De Slijke,E., De Winne,N., Eeckhout,D., Persiau,G., Van Breusegem,F. *et al.* (2014) A generic tool for transcription factor target gene discovery in Arabidopsis cell suspension cultures based on tandem chromatin affinity purification. *Plant Physiol.*, **164**, 1122–1133.
28. Li,C., Qiao,Z., Qi,W., Wang,Q., Yuan,Y., Yang,X., Tang,Y., Mei,B., Lv,Y., Zhao,H. *et al.* (2015) Genome-wide characterization of cis-acting DNA targets reveals the transcriptional regulatory framework of opaque2 in maize. *Plant Cell*, **27**, 532–545.
29. Cui,X., Lu,F., Qiu,Q., Zhou,B., Gu,L., Zhang,S., Kang,Y., Cui,X., Ma,X., Yao,Q. *et al.* (2016) REF6 recognizes a specific DNA sequence to demethylate H3K27me3 and regulate organ boundary formation in Arabidopsis. *Nat. Genet.*, **48**, 694–699.
30. Birkenbihl,R.P., Kracher,B. and Somssich,I.E. (2017) Induced genome-wide binding of three Arabidopsis WRKY transcription factors during early MAMP-triggered immunity. *Plant Cell*, **29**, 20–38.
31. O'Malley,R.C., Huang,S.-S.C., Song,L., Lewsey,M.G., Bartlett,A., Nery,J.R., Galli,M., Gallavotti,A. and Ecker,J.R. (2016) Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell*, **165**, 1280–1292.
32. Isakova,A., Groux,R., Imbeault,M., Rainer,P., Alpern,D., Dainese,R., Ambrosini,G., Trono,D., Bucher,P. and Deplancke,B. (2017) SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nat. Methods*, **14**, 316–322.
33. Franco-Zorrilla,J.M., López-Vidriero,I., Carrasco,J.L., Godoy,M., Vera,P. and Solano,R. (2014) DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 2367–2372.
34. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
35. Weirauch,M.T., Yang,A., Albu,M., Cote,A., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.

36. Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.*, **338**, 207–215.
37. Castro-Mondragon,J.A., Jaeger,S., Thieffry,D., Thomas-Chollier,M. and van Helden,J. (2017) RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res.*, **45**, e119.
38. Kwon,A.T., Arenillas,D.J., Worsley Hunt,R. and Wasserman,W.W. (2012) oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. *G3*, **2**, 987–1002.
39. Mathelier,A., Lefebvre,C., Zhang,A.W., Arenillas,D.J., Ding,J., Wasserman,W.W. and Shah,S.P. (2015) Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. *Genome Biol.*, **16**, 1–17.
40. Verfaillie,A., Imrichova,H., Janky,R. and Aerts,S. (2015) iRegulon and i-cistarget: reconstructing regulatory networks using motif and track enrichment. *Curr. Protoc. Bioinformatics*, **52**, 1–39.
41. Arenillas,D.J., Forrest,A.R.R., Kawaji,H., Lassmann,T. and FANTOM ConsortiumFANTOM Consortium, Wasserman,W.W. and Mathelier,A. (2016) CAGEd-oPOSSUM: motif enrichment analysis from CAGE-derived TSSs. *Bioinformatics*, **32**, 2858–2860.
42. Shi,W., Fornes,O., Mathelier,A. and Wasserman,W.W. (2016) Evaluating the impact of single nucleotide variants on transcription factor binding. *Nucleic Acids Res.*, **44**, 10106–10116.
43. Arner,E., Daub,C.O., Vitting-Seerup,K., Andersson,R., Lilje,B., Drabløs,F., Lennartsson,A., Rönnerblad,M., Hrydziuszko,O., Vitezic,M. *et al.* (2015) Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science*, **347**, 1010–1014.
44. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest,A.R.R., Kawaji,H., Rehli,M., Baillie,J.K., de Hoon,M.J.L., Haberle,V., Lassmann,T., Kulakovskiy,I.V., Lizio,M. *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.
45. Neph,S., Vierstra,J., Stergachis,A.B., Reynolds,A.P., Haugen,E., Vernot,B., Thurman,R.E., John,S., Sandstrom,R., Johnson,A.K. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**, 83–90.
46. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
47. Raney,B.J., Dreszer,T.R., Barber,G.P., Clawson,H., Fujita,P.A., Wang,T., Nguyen,N., Paten,B., Zweig,A.S., Karolchik,D. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.

JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework

Aziz Khan, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A. Castro-Mondragon, Robin van der Lee, Adrien Bessy, Jeanne Chèneby, Shubhada R. Kulkarni, Ge Tan, Damir Baranasic, David J. Arenillas, Albin Sandelin, Klaas Vandepoele, Boris Lenhard, Benoît Ballester, Wyeth W. Wasserman, François Parcy, and Anthony Mathelier

SUPPLEMENTARY TEXT

ChIP-seq data processing

Human ChIP-seq data

Peak summit locations for human (hg38) TF ChIP-seq data were retrieved from the uniformly processed ChIP-seq datasets stored in the ReMap 2018 database at http://remap.cisreg.eu (1). Each dataset was analyzed to predict PFMs from the ChIP-seq peak summits using the following protocol:

1. Extract genomic regions of +/- 50bp around the peak summits using BedTools (2).
2. Apply the RSAT *peak-motifs* tool (3) to discover de novo motifs with the option '-disco local_words,positions'
3. For each of the discovered motifs:
    a. Compute the corresponding PWM
    b. Extract +/- 250bp around peak summits
    c. Predict TFBSs from these sequences using a 85% threshold on the relative scores of the PWM
    d. Compute a centrality p-value (following (4))
4. For each TF, select the TF binding profile from all associated ChIP-seq datasets providing the best centrality p-value.

The TF binding profiles obtained were then manually curated for inclusion into JASPAR 2018.

Plant ChIP-seq data

Raw reads from the original publications were collected (5–9). Reads were aligned using Bowtie2 (10) to the genomes of *Arabidopsis thaliana* (TAIR10) and *Zea mays* (5b.60). ChIP-seq peaks were called using MACS2 (11). Genomic regions around peak summits were analyzed the same way as for human ChIP-seq data (see above), except the option '-disco local_words,positions,oligos' was used for the RSAT *peak-motifs* tool.

TFFM computation

ChIP-seq data from ReMap 2018 (1) have been collected to construct TFFMs. JASPAR 2018 TF binding profiles were assigned to the ChIP-seq datasets wherever possible. These profiles

were used to initialize TFFMs that were trained on genomic regions of -/+ 50bp around the corresponding peak summits. Centrality enrichment p-values were computed using genomic regions +/- 250bp on each side of the peak summits. TFFMs providing a centrality p-value < -200 were further assessed for manual curation.

Sequence logos

In this release of JASPAR, we have regenerated all sequence logos as SVG files using the R package ggseqlogo (12).

Matrix clustering

TF binding profile clusterizations were obtained using the RSAT *matrix-clustering* tool (13) with the following parameters:
-hclust_method average -calc sum -metric_build_tree Ncor -lth w 5 -lth cor 0.6 -lth Ncor 0.4 -label_in_tree name -return json -radial_tree_only

UCSC Genome Browser track data hubs

We generated a custom UCSC Genome Browser track data hub (14) containing genome-wide TFBS predictions from PFMs in the JASPAR CORE vertebrates collection. Specifically, for each profile, the human genome assemblies hg19 and hg38 were scanned in parallel using the TFBS Perl module (15) and FIMO (16), as distributed within the MEME suite (version 4.11.2) (17). For scanning the human genome with the BioPerl TFBS module, we converted profiles to PWMs and kept matches with a relative score ≥ 0.8. For the FIMO scan, profiles were reformatted to MEME motifs and matches with a *p*-value < 0.05 were kept. TFBS predictions that were not consistent between the two methods (TFBS Perl module and FIMO) were filtered out. The remaining TFBS predictions were converted to genome tracks and colored according to their FIMO *p*-value (scaled between 0-1000, where 0 corresponds to a *p*-value of 1 and 1000 to a *p*-value ≤ 10^{-10}) to allow for comparison of prediction confidence between different profiles. The tracks are collected as a data hub that can be visualized in the UCSC Genome Browser or downloaded for custom analysis (http://jaspar.genereg.net/genome-tracks/). Code and data used to create the UCSC tracks are available at https://github.com/wassermanlab/JASPAR-UCSC-tracks. The underlying BED files and individual matches for each TF binding profile on the human genome (hg19 and hg38 genome assemblies) are available at http://expdata.cmmt.ubc.ca/JASPAR/downloads/UCSC_tracks/2018/.

References

1. Chèneby,J., Gheorghe,M., Artufel,M., Mathelier,A. and Ballester,B. (2018) ReMap 2018: An updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.*, **Submitted**.

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

3. Thomas-Chollier,M., Herrmann,C., Defrance,M., Sand,O., Thieffry,D. and van Helden,J. (2012) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res.*, **40**, e31.

4. Bailey,T.L. and Machanick,P. (2012) Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.*, **40**, e128.

5. Eveland,A.L., Goldshmidt,A., Pautler,M., Morohashi,K., Liseron-Monfils,C., Lewis,M.W., Kumari,S., Hiraga,S., Yang,F., Unger-Wallace,E., *et al.* (2014) Regulatory modules controlling maize inflorescence architecture. *Genome Res.*, **24**, 431–443.

6. Verkest,A., Abeel,T., Heyndrickx,K.S., Van Leene,J., Lanz,C., Van De Slijke,E., De Winne,N., Eeckhout,D., Persiau,G., Van Breusegem,F., *et al.* (2014) A generic tool for transcription factor target gene discovery in Arabidopsis cell suspension cultures based on tandem chromatin affinity purification. *Plant Physiol.*, **164**, 1122–1133.

7. Li,C., Qiao,Z., Qi,W., Wang,Q., Yuan,Y., Yang,X., Tang,Y., Mei,B., Lv,Y., Zhao,H., *et al.* (2015) Genome-wide characterization of cis-acting DNA targets reveals the transcriptional regulatory framework of opaque2 in maize. *Plant Cell*, **27**, 532–545.

8. Cui,X., Lu,F., Qiu,Q., Zhou,B., Gu,L., Zhang,S., Kang,Y., Cui,X., Ma,X., Yao,Q., *et al.* (2016) REF6 recognizes a specific DNA sequence to demethylate H3K27me3 and regulate organ boundary formation in Arabidopsis. *Nat. Genet.*, **48**, 694–699.

9. Birkenbihl,R.P., Kracher,B. and Somssich,I.E. (2017) Induced Genome-Wide Binding of Three Arabidopsis WRKY Transcription Factors during Early MAMP-Triggered Immunity. *Plant Cell*, **29**, 20–38.

10. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

11. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W., *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.

12. Wagih,O. (2017) ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics*, 10.1093/bioinformatics/btx469.

13. Castro-Mondragon,J.A., Jaeger,S., Thieffry,D., Thomas-Chollier,M. and van Helden,J. (2017) RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res.*, **45**, e119.

14. Raney,B.J., Dreszer,T.R., Barber,G.P., Clawson,H., Fujita,P.A., Wang,T., Nguyen,N., Paten,B., Zweig,A.S., Karolchik,D., *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.

15. Lenhard,B. and Wasserman,W.W. (2002) TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.

16. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

17. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–8.

Nucleic Acids Research, 2019 ***1–6***
doi: 10.1093/nar/gkz1001

JASPAR 2020: update of the open-access database of transcription factor binding profiles

Oriol Fornes[1,†], Jaime A. Castro-Mondragon[2,†], Aziz Khan[2,†], Robin van der Lee[1], Xi Zhang[1], Phillip A. Richmond[1], Bhavi P. Modi[1], Solenne Correard[1], Marius Gheorghe[2], Damir Baranašić[3,4], Walter Santana-Garcia[5], Ge Tan[6], Jeanne Chèneby[7], Benoit Ballester[7], François Parcy[8], Albin Sandelin[9], Boris Lehnard[3,4], Wyeth W. Wasserman[1] and Anthony Mathelier[2,10,*]

[1]Centre for Molecular Medicine and Therapeutics, Department of Medical Genetics, BC Children's Hospital Research Institute, University of British Columbia, 950 W 28th Ave, Vancouver, BC V5Z 4H4, Canada, [2]Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway, [3]Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London W12 0NN, UK, [4]Computational Regulatory Genomics, MRC London Institute of Medical Sciences, London W120NN, UK, [5]Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France, [6]Functional Genomics Centre Zurich, ETH Zurich, Zurich, Switzerland, [7]Aix Marseille Univ, INSERM, TAGC, Marseille, France, [8]CNRS, Univ. Grenoble Alpes, CEA, INRA, IRIG-LPCV, 38000 Grenoble, France, [9]The Bioinformatics Centre, Department of Biology and Biotech Research & Innovation Centre, University of Copenhagen, DK2200 Copenhagen N, Denmark and [10]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, 0310 Oslo, Norway

Received September 15, 2019; Revised October 15, 2019; Editorial Decision October 16, 2019; Accepted October 16, 2019

ABSTRACT

JASPAR (http://jaspar.genereg.net) is an open-access database of curated, non-redundant transcription factor (TF)-binding profiles stored as position frequency matrices (PFMs) for TFs across multiple species in six taxonomic groups. In this 8th release of JASPAR, the CORE collection has been expanded with 245 new PFMs (169 for vertebrates, 42 for plants, 17 for nematodes, 10 for insects, and 7 for fungi), and 157 PFMs were updated (125 for vertebrates, 28 for plants and 3 for insects). These new profiles represent an 18% expansion compared to the previous release. JASPAR 2020 comes with a novel collection of unvalidated TF-binding profiles for which our curators did not find orthogonal supporting evidence in the literature. This collection has a dedicated web form to engage the community in the curation of unvalidated TF-binding profiles. Moreover, we created a Q&A forum to ease the communication between the user community and JASPAR curators. Finally, we updated the genomic tracks, inference tool, and TF-binding profile similarity clusters. All the data is available through the JASPAR website, its associated RESTful API, and through the JASPAR2020 R/Bioconductor package.

INTRODUCTION

Transcription factors (TFs) are proteins involved in the regulation of gene expression at the transcriptional level (1). They interact with DNA in a sequence-specific manner through their DNA-binding domains (DBDs), which are used to classify TFs into structural families (2). The genomic locations where TFs bind to DNA are known as TF binding sites (TFBSs), which are typically short (6–20 bp) and exhibit sequence variability (3). Genome-wide identification of TFBSs is key to understanding transcriptional regulation. As it is not possible to identify all TFBSs for every cell type and cellular condition experimentally, computational modeling of TF-binding specificities has been instrumental to predict TFBSs in the genome. These computational models aim at representing the complex interplay between nucleotide and/or DNA shape readout at TFBSs (4), and can be used to predict not only the precise location where TFs interact in the genome (5), but also TFs with enriched TFBSs in a set of sequences (6), or the impact of mutations on TF binding (7,8), amongst others.

From the plethora of existing computational models (9), position frequency matrices (PFMs) (10) are one of the sim-

*To whom correspondence should be addressed. Email: anthony.mathelier@ncmm.uio.no
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

plest and (still) most commonly used, although more complex models, for instance based on hidden Markov models or deep learning (11–13), are becoming more common. A PFM is a TF-binding profile that models the DNA-binding specificity of a TF by summarizing the frequencies of each nucleotide at each position from observed TF-DNA interactions. These interactions are usually derived from *in vitro* assays (e.g. SELEX (14) or protein binding microarrays (15)), which assess the binding affinity of TFs to DNA sequences, or from ChIP-based experiments (e.g. ChIP-seq (16), ChIP-exo (17), or ChIP-nexus (18)), which capture TF-DNA interactions *in vivo*, by looking for over-represented DNA sequences in regions bound by the ChIP'ed TF.

With the advent of high-throughput sequencing more than a decade ago, the number of PFMs derived from *in vivo and in vitro* experiments has increased dramatically, leading to the creation of multiple databases storing PFMs or more complex TF-binding profiles such as JASPAR (19), CIS-BP (20) and HOCOMOCO (21) (see (22) for a comprehensive review). The JASPAR database (http://jaspar.genereg.net/) is one of the most popular databases of TF-binding profiles, and has been maintained for over 15 years (23). As such, many computational tools dedicated to the study of gene regulation incorporate profiles from JASPAR (*e.g.* TFBSshape (24), RSAT (25), MEME (26) or i-cisTarget (6)). At the heart of JASPAR is its CORE collection, which contains TF-binding profiles that are: (i) manually curated (meaning that orthogonal supporting evidence from the literature is required for each profile); (ii) non-redundant (one profile per TF with the exception of TFs with multiple DNA-binding sequence preferences (27)); (iii) associated with TFs from one of six taxa (vertebrates, nematodes, insects, plants, fungi, and urochordata) and (iv) freely available to the community through a user-friendly web interface, a RESTful API (28), and a dedicated R/Bioconductor data package ('JASPAR2020').

Here, we present the 8th release of JASPAR, which comes with a major expansion and update of its CORE collection. Moreover, we introduce a new collection of unvalidated profiles, which stores quality-controlled PFMs for which our curators could not find orthogonal support. This collection has a dedicated web interface to engage the community of users in the curation of TF-binding profiles. Finally, we have updated the hierarchical clusters of TF-binding profiles, the genomic tracks of predicted TFBSs (now available for 8 genomes), and the profile inference tool.

EXPANSION AND UPDATE OF THE JASPAR CORE COLLECTION

For this 8th release of JASPAR, we added to the CORE collection 245 new TF-binding profiles for TFs in the following taxa: vertebrates (169 profiles, corresponding to an expansion of 29% for this taxon), plants (42 profiles, 9% expansion), nematodes (17 profiles, 65% expansion), insects (10 profiles, 8% expansion) and fungi (7 profiles, 4% expansion). We updated 157 profiles (Table 1). The new PFMs were derived from HT-SELEX (29), PBMs (20), ChIP-seq and DAP-seq experiments (data sourced from CistromeDB (30), ReMap (31,32), GTRD (33), ChIP-atlas (34) and ModERN

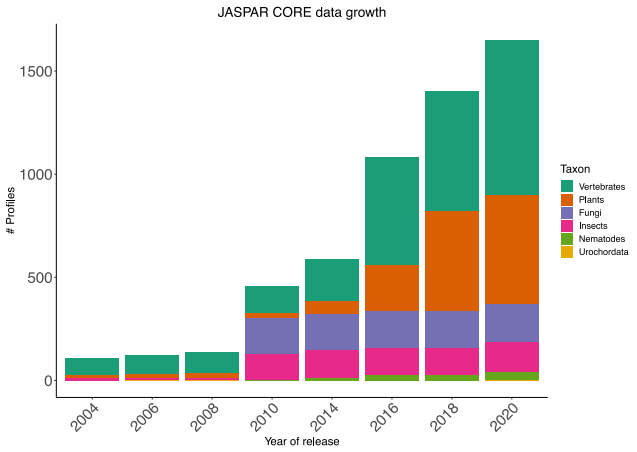


Figure 1. JASPAR CORE growth. The number of profiles in each taxon and overall (see legend) through all JASPAR releases. Inset: pie chart of the percentages of profiles on each taxa in the JASPAR 2020 CORE collection.

(35), see Supplementary Text for method details). As previously described, the newly introduced profiles were manually curated to be supported by an orthogonal reference from the literature, which is provided in the metadata of the profiles. Moreover, the TF DBD class and family (following the TFClass classification (2)), the TF UniProt ID (36), and links to the TFBSshape (37,38), ReMap (31,32) and UniBind (39) databases are provided in the profiles metadata (whenever possible). Finally, the profiles previously associated with ID2, ID4 and TRB2 were removed from the CORE collection as these proteins are not TFs (1).

Overall, the JASPAR 2020 CORE collection includes 1647 non-redundant PFMs (746 for vertebrates, 530 for plants, 183 for fungi, 143 for insects, 43 for nematodes and 1 for urochordates) (Table 1; Figure 1). Moreover, we continued with the incorporation of novel transcription factor flexible models (TFFMs), which are hidden Markov-based models capturing dinucleotide dependencies in TF–DNA interactions (11). We introduced new TFFMs for 217 TFs (136 for vertebrates, 38 for plants, 21 for insects, 17 for nematodes, and 5 for fungi) and updated TFFMs for 20 vertebrates TFs, which represents a 50% increase in the number of TFFMs available. All data is available on the JASPAR website, its associated RESTful API, and through the JASPAR2020 R/Bioconductor package.

A NEW COLLECTION OF UNVALIDATED PROFILES FOR COMMUNITY ENGAGEMENT

We introduced a novel 'unvalidated' collection to store high-quality (i.e. passing multiple quality controls, see Supplementary Text) TF-binding profiles for which no independent support was found in the literature by our curators. This collection contains 337 PFMs. As these profiles are not yet supported by an orthogonal evidence, we recommend users to use this collection with caution. We encourage the community to engage in the curation of these profiles by providing the JASPAR curators with supporting complementary evidence (from their own work or others) whenever possible. This is facilitated by the availability of an individ-

Table 1. Overview of the growth of the number of PFMs in the JASPAR 2020 CORE and unvalidated collections compared to the JASPAR 2018 CORE collection

Taxonomic Group	Non-redundant PFMs in JASPAR 2018	New non-redundant PFMs in JASPAR 2020	Removed profiles	Updated PFMs in JASPAR 2020	Total PFMs (non-redundant) in JASPAR 2020	Total PFMs (all versions) in JASPAR 2020
Vertebrates	579	169	2	125	746	1011
Plants	489	42	1	28	530	572
Insects	133	10	0	3	143	153
Nematodes	26	17	0	0	43	43
Fungi	176	7	0	0	183	184
Urochordata	1	0	0	0	1	1
Total CORE	**1404**	**245**	**3**	**157**	**1647**	**1964**
unvalidated					**337**	**337**

ual submission form for each profile in the 'unvalidated' collection (Figure 2).

Further, we started a Q&A forum (https://groups.google.com/forum/#!forum/jaspar) to ease the communication between JASPAR curators and the community; we welcome the community to send us their questions and suggestions, or to report errors in JASPAR.

CLUSTERED PROFILES, GENOMIC TRACKS AND PROFILE INFERENCE TOOL

In the previous releases, we introduced novel features such as hierarchical clustering of TF-binding profiles in the CORE collection to visualize profile similarities, genomic tracks of predicted TFBSs, and an inference tool to predict TF-binding profiles likely recognized by TFs not available in the JASPAR CORE. We improved the profile inference tool using our own implementation of a recently described similarity regression method (20). We updated the generation of genomic tracks that are publicly available through the UCSC Genome Browser data hub (40) for 7 organisms: human (hg19, hg38), mouse (mm10), zebrafish (danRer11), *Drosophila melanogaster* (dm6), *Caenorhabditis elegans* (ce10), *Arabidopsis thaliana* (araTha1) and baker's yeast (sacCer3). For more details on the updated genomic tracks and inference tool, refer to the Supplementary Text. Finally, we generated the hierarchical clusters of available TF-binding profiles for each taxon with RSAT *matrix-clustering* (41). Users can explore the CORE/unvalidated collection through the trees and access directly the corresponding profiles by clicking on the TF name.

CONCLUSIONS AND PERSPECTIVES

Similar to previous releases, we substantially expanded the CORE collection of the JASPAR database. For this 8th release, we processed more than 18,000 ChIP-seq datasets. As a large number of the obtained high-quality TF-binding profiles were not supported with orthogonal supporting evidence, it motivated us to create the novel 'unvalidated' collection of profiles. We expect that upcoming experiments and publications will provide additional supporting evidence to some profiles to be incorporated into the JASPAR CORE collection. Meanwhile, we would like to extend our invitation to the research community to 1) help us curate these unvalidated profiles (e.g. by pointing us to supporting literature), and 2) send us their own novel profiles (e.g. determined experimentally) for incorporation in the next release of JASPAR.

The JASPAR CORE vertebrates collection now contains 746 profiles, 637 of which are associated with human TFs with known DNA-binding profiles (1), which corresponds to a 58% of the 1,107 reported by Lambert *et al.* (1). While this is an impressive collective achievement by the field (the original JASPAR database only contained 81 profiles, a ∼7% coverage for human TFs), it suggests that targeted experimental efforts to find the binding preferences for remaining TFs will be important. Although computational approaches can be used to infer missing TF-binding profiles (20,42), especially for non-model organisms, the JASPAR approach is conservative, including profiles supported by at least two experiments in the literature. This is very important as we stand by the reliability of our data. Since its initial publication in 2004 (23), the JASPAR database has been committed to provide the research community with high-quality, manually curated, non-redundant TF-binding profiles.

Lastly, although PFMs have dominated the field of gene regulation for decades, new profile representations have emerged. For example, profiles with expanded alphabets to represent methylated bases (43,44), modelling binding energy (45) or derived from deep learning importance scores (46). Depending on how the field evolves and how popular these profiles become, we will consider them for inclusion in JASPAR in the future.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

We thank the user community for useful input and the scientific community for performing experimental assays of TF–DNA interactions and for publicly releasing the data. We thank Giovanna Ambrosini for her help with PWMScan, the UCSC Genome Browser Project Team for their assistance with the genome tracks, WestGrid (https://www.westgrid.ca), Compute Canada (https://www.computecanada.ca), Georgios Magklaras and Georgios Marselis for their IT support, Jacques van Helden and Adam Handel for contacting us to add and validate TF

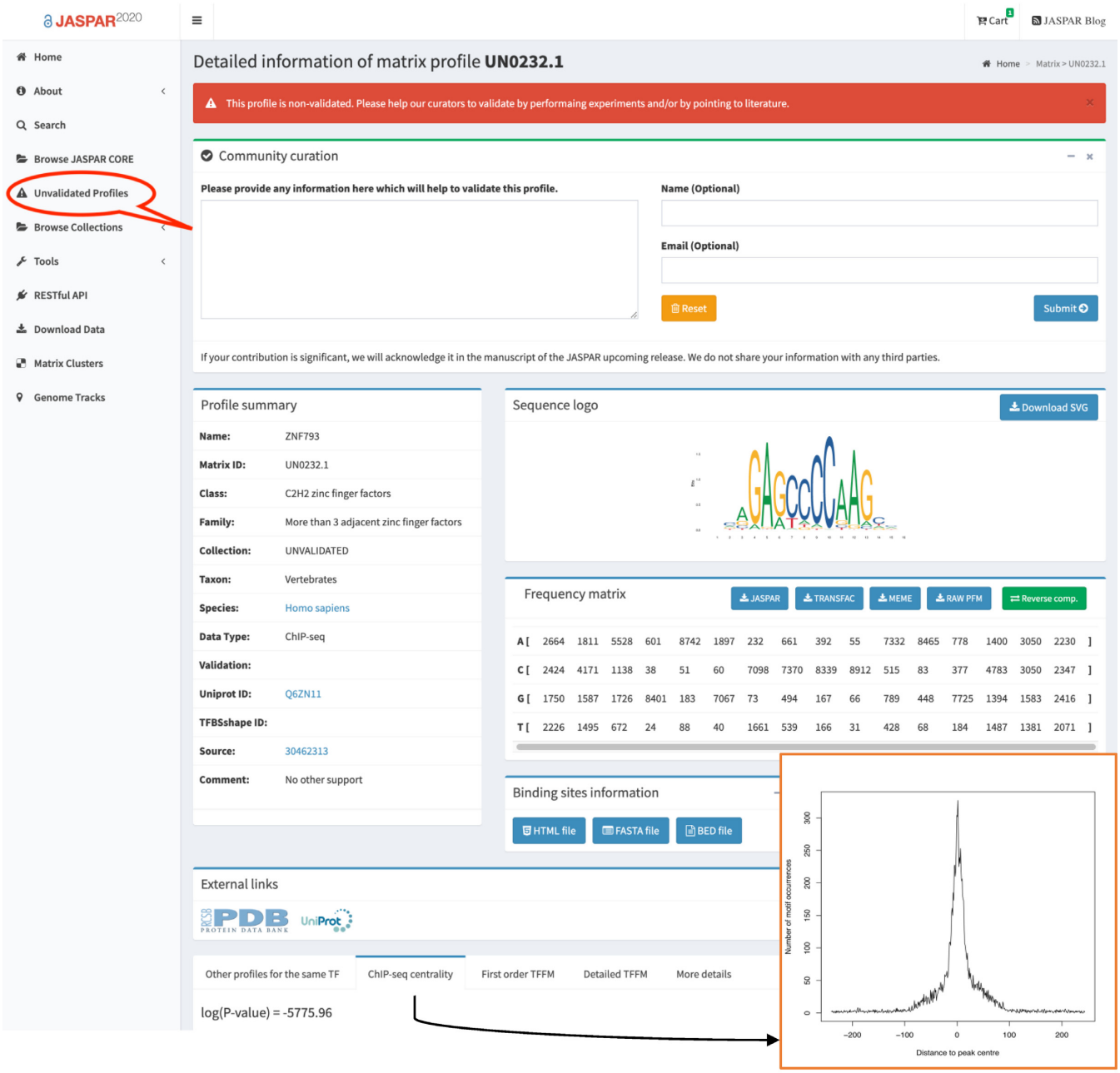


Figure 2. Unvalidated TF-binding profile collection. Example with the ZNF793 profile. This high-quality PFM was derived from a ChIP-seq experiment and was built from thousands of potential TFBSs. Further, the TFBSs are enriched around the ChIP-seq peak summits. However, no orthogonal evidence supporting this profile was found by our curators. Users can upload relevant information about the profile in the unvalidated collection through the '*Community curation*' box.

binding profiles, and Dora Pak and Ingrid Kjelsvik for administrative support.

FUNDING

Norwegian Research Council [187615]; Helse Sør-Øst; University of Oslo through the Centre for Molecular Medicine Norway (NCMM) (to A.M., J.A.C.-M., A.K., M.G.); Norwegian Research Council [288404 to J.A.C.-M. and Mathelier group]; The Norwegian Cancer Society [197884 to Mathelier group]; O.F., X.Z., P.A.R., S.C. and W.W.W. were supported by grants from the Canadian Institutes of Health Research [BOP-149430 and PJT-162120]; Genome Canada and Genome British Columbia [255ONT and 275SIL]; Michael Smith Foundation for Health Research [17746]; Natural Sciences and Engineering Research Council of Canada Discovery Grant [RGPIN-2017-06824]; CREATE programs; Weston Brain Institute [20R74681]; BC Children's Hospital Foundation and Research Institute; Netherlands Organization for Scientific Research [Rubicon fellowship to R.v.d.L., 452172015]; Genome British Columbia

[SIP007 to B.P.M.]; A.S. was supported by grants from the Lundbeck Foundation, the Danish Cancer Foundation, the Danish Innovation Fund and the Danish Council for Independent Research. F.P. was supported by the French National Agency for Research [FloPiNet ANR-16-CE92-0023-01; GRAL, ANR-10-LABX-49-01]; D.B. is a recipient of a Rutherford Fund Fellowship.
Conflict of interest statement. None declared.

REFERENCES

1. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The human transcription factors. *Cell*, **172**, 650–665.
2. Wingender,E., Schoeps,T., Haubrock,M., Krull,M. and Dönitz,J. (2018) TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Res.*, **46**, D343–D347.
3. Patikoglou,G.A., Kim,J.L., Sun,L., Yang,S.H., Kodadek,T. and Burley,S.K. (1999) TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes Dev.*, **13**, 3217–3230.
4. Slattery,M., Zhou,T., Yang,L., Dantas Machado,A.C., Gordân,R. and Rohs,R. (2014) Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.*, **39**, 381–399.
5. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
6. Imrichová,H., Hulselmans,G., Atak,Z.K., Potier,D. and Aerts,S. (2015) i-cisTarget 2015 update: generalized cis-regulatory enrichment analysis in human, mouse and fly. *Nucleic Acids Res.*, **43**, W57–W64.
7. Fornes,O., Gheorghe,M., Richmond,P.A., Arenillas,D.J., Wasserman,W.W. and Mathelier,A. (2018) MANTA2, update of the Mongo database for the analysis of transcription factor binding site alterations. *Sci. Data*, **5**, 180141.
8. Fu,Y., Liu,Z., Lou,S., Bedford,J., Mu,X.J., Yip,K.Y., Khurana,E. and Gerstein,M. (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.*, **15**, 480.
9. Weirauch,M.T., Cote,A., Norel,R., Annala,M., Zhao,Y., Riley,T.R., Saez-Rodriguez,J., Cokelaer,T., Vedenko,A., Talukder,S. *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.
10. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
11. Mathelier,A. and Wasserman,W.W. (2013) The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, **9**, e1003214.
12. Kiesel,A., Roth,C., Ge,W., Wess,M., Meier,M. and Söding,J. (2018) The BaMM web server for de-novo motif discovery and regulatory sequence analysis. *Nucleic Acids Res.*, **46**, W215–W220.
13. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
14. Jolma,A., Kivioja,T., Toivonen,J., Cheng,L., Wei,G., Enge,M., Taipale,M., Vaquerizas,J.M., Yan,J., Sillanpää,M.J. *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, **20**, 861–873.
15. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. 3rd and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
16. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
17. Pugh,B.F. and Franklin Pugh,B. (2012) Ultra-high resolution mapping of protein-genome interactions using ChIP-exo. *BMC Proc.*, **6**, O27.
18. He,Q., Johnston,J. and Zeitlinger,J. (2015) ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat. Biotechnol.*, **33**, 395–401.
19. Khan,A., Fornes,O., Stigliani,A., Gheorghe,M., Castro-Mondragon,J.A., van der Lee,R., Bessy,A., Chèneby,J., Kulkarni,S.R., Tan,G. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D1284.
20. Lambert,S.A., Yang,A.W.H., Sasse,A., Cowley,G., Albu,M., Caddick,M.X., Morris,Q.D., Weirauch,M.T. and Hughes,T.R. (2019) Similarity regression predicts evolution of transcription factor sequence specificity. *Nat. Genet.*, **51**, 981–989.
21. Kulakovskiy,I.V., Vorontsov,I.E., Yevshin,I.S., Sharipov,R.N., Fedorova,A.D., Rumynskiy,E.I., Medvedeva,Y.A., Magana-Mora,A., Bajic,V.B., Papatsenko,D.A. *et al.* (2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.*, **46**, D252–D259.
22. Stormo,G.D. (2015) DNA motif databases and their uses. *Curr. Protoc. Bioinformatics*, **51**, 2.15.1–6.
23. Sandelin,A., Alkema,W., Engström,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.
24. Yang,L., Dror,I., Zhou,T., Mathelier,A., Wasserman,W.W., Gordân,R. and Rohs,R. (2015) 15 TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *J. Biomol. Struct. Dyn.*, **33**, 9–9.
25. Nguyen,N.T.T., Contreras-Moreira,B., Castro-Mondragon,J.A., Santana-Garcia,W., Ossio,R., Robles-Espinoza,C.D., Bahin,M., Collombet,S., Vincens,P., Thieffry,D. *et al.* (2018) RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Res.*, **46**, W209–W214.
26. Bailey,T.L., Johnson,J., Grant,C.E. and Noble,W.S. (2015) The MEME suite. *Nucleic Acids Res.*, **43**, W39–W49.
27. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
28. Khan,A. and Mathelier,A. (2018) JASPAR RESTful API: accessing JASPAR data from any programming language. *Bioinformatics*, **34**, 1612–1614.
29. Yin,Y., Morgunova,E., Jolma,A., Kaasinen,E., Sahu,B., Khund-Sayeed,S., Das,P.K., Kivioja,T., Dave,K., Zhong,F. *et al.* (2017) Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, **356**, eaaj2239.
30. Zheng,R., Wan,C., Mei,S., Qin,Q., Wu,Q., Sun,H., Chen,C.-H., Brown,M., Zhang,X., Meyer,C.A. *et al.* (2019) Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res.*, **47**, D729–D735.
31. Chèneby,J., Gheorghe,M., Artufel,M., Mathelier,A. and Ballester,B. (2017) ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.*, **46**, D267–D275.
32. Chèneby,J., Ménétrier,J., Mestdagh,M., Rosnet,T., Douida,A., Rhalloussi,W., Bergon,A., Lopez,F. and Ballester,B. (2019) ReMap 2020: A database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res.*, **46**, D267–D275.
33. Yevshin,I., Sharipov,R., Kolmykov,S., Kondrakhin,Y. and Kolpakov,F. (2019) GTRD: a database on gene transcription regulation—2019 update. *Nucleic Acids Res.*, **47**, D100–D105.
34. Oki,S., Ohta,T., Shioi,G., Hatanaka,H., Ogasawara,O., Okuda,Y., Kawaji,H., Nakaki,R., Sese,J. and Meno,C. (2018) ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep.*, **19**, e46255.
35. Kudron,M.M., Victorsen,A., Gevirtzman,L., Hillier,L.W., Fisher,W.W., Vafeados,D., Kirkey,M., Hammonds,A.S., Gersch,J., Ammouri,H. *et al.* (2018) The ModERN Resource: Genome-Wide binding profiles for hundreds of drosophila and caenorhabditis elegans transcription factors. *Genetics*, **208**, 937–949.
36. Consortium,T.U. and The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
37. Chiu,T.P., Xin,B., Markarian,N., Wang,Y. and Rohs,R. (2019) TFBSshape v2.0: an expanded motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res.*, **42**, D148–D155.
38. Yang,L., Zhou,T., Dror,I., Mathelier,A., Wasserman,W.W., Gordân,R. and Rohs,R. (2014) TFBSshape: a motif database for

DNA shape features of transcription factor binding sites. *Nucleic Acids Res.*, **42**, D148–D155.
39. Gheorghe,M., Sandve,G.K., Khan,A., Chèneby,J., Ballester,B. and Mathelier,A. (2019) A map of direct TF-DNA interactions in the human genome. *Nucleic Acids Res.*, **47**, e21.
40. Raney,B.J., Dreszer,T.R., Barber,G.P., Clawson,H., Fujita,P.A., Wang,T., Nguyen,N., Paten,B., Zweig,A.S., Karolchik,D. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.
41. Castro-Mondragon,J.A., Jaeger,S., Thieffry,D., Thomas-Chollier,M. and van Helden,J. (2017) RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res.*, **45**, e119.
42. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.-Y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.
43. Viner,C., Johnson,J., Walker,N., Shi,H., Sjöberg,M., Adams,D.J., Ferguson-Smith,A.C., Bailey,T.L. and Hoffman,M.M. Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet. 10.1101/043794.
44. Chang,Y.K., Granas,D. and Stormo,G.D. (2017) Measuring quantitative effects of methylation on transcription factor–DNA binding affinity. *Science.*, **3**, eaao1799.
45. Ruan,S., Swamidass,S.J. and Stormo,G.D. (2017) BEESEM: estimation of binding energy models using HT-SELEX data. *Bioinformatics*, **33**, 2288–2295.
46. Shrikumar,A., Tian,K., Shcherbina,A., Avsec,Ž., Banerjee,A., Sharmin,M., Nair,S. and Kundaje,A. (2018) TF-MoDISco v0.4.2.2-alpha: Technical Note.

C. Conclusion

L'utilisation des données issues de méthodes à haut débit, comme le ChIP-seq, a conduit à l'augmentation des données JASPAR sur les profils de liaison des TFs (Tableau III.1). De plus, l'ajout de nouvelles données a permis d'affiner les profils de fixation déjà définis. Pour la mise à jour 2020, JASPAR utilise des données issues d'expériences de ChIP-exo. Ces données définissent une région de fixation plus précise. En réduisant le bruit des données utilisées, elles permettent de calculer les matrices poids-position des profils de fixation des TFs. Une autre possibilité, serait d'utiliser des données issues de single cell ChIP-seq[183]. Les méthodes de séquençage single cell permettant d'identifier les régions de fixation pour chaque cellule, cette méthode pourrait identifier la variabilité cellulaire.

Tableau III.1 : Evolution des profils de liaison disponibles sur JASPAR de 2016 à 2020

Taxonomic group	Non-redundant PFMs in JASPAR 2016	Total PFMs (non-redundant) in JASPAR 2018	Total PFMs (non-redundant) in JASPAR 2020	Total PFMs (all versions) in JASPAR 2020
Vertebrates	519	579	735	929
Plants	227	489	498	511
Insects	133	133	142	152
Nematodes	26	26	42	42
Fungi	176	176	183	184
Urochordata	1	1	1	1
Total	1082	1404	1601	1819

IV Etude de la fixation directe des protéines dans les expériences ChIP-seq

A. Introduction

Les expériences de ChIP capturent les interactions protéine/ADN. Toutefois, la capture des fragments d'intérêt par les anticorps n'est pas exempte de faux positifs. Il arrive que soient capturées des interactions protéine/ADN, indirectes ou non, spécifiques/passagères[175,184]. La capture des interactions indirectes peut être intéressante pour identifier les cofacteurs qui ne se fixent pas directement à l'ADN. En revanche, si on s'intéresse uniquement au TF, ces artefacts de ChIP peuvent fausser les analyses. En particulier lors de l'étude de l'influence des modifications de la séquence nucléotidique sur la fixation de TFs, ou encore si on souhaite modéliser les profils de fixation de TF comme le fait JASPAR. Les méthodes ChIP étant très répandues, il est intéressant de trouver une méthode permettant d'identifier les fixations directes dans les données ChIP.

B. ChIP-eat

Le pipeline ChIP-eat utilise des modèles de calcul de liaison TF sur des pics ChIP afin de prédire automatiquement les interactions directes TF/ADN. C'est donc une méthode à posteriori d'identification directe.

Le logiciel utilise des listes de "pics de fixation" capturés par une expérience de ChIP-seq et un modèle de fixation matrice poids-position (PPM). Le but est d'extraire, pour tous les pics, le meilleur site potentiel de fixation. Dans un premier temps, un score est défini pour chaque position ce qui permet de noter la similarité entre la région et le profil théorique de fixation de la protéine. Dans un deuxième temps, le logiciel se base sur l'observation suivante. Plus le site de fixation est proche du sommet du pic, plus il est probable que la fixation TF/ADN observée soit directe[169–171]. En résumé, les sites de fixation ayant un score élevé et proches du sommet des pics ont plus de probabilité de représenter des interactions protéine/ADN directes.

La difficulté repose sur l'attribution d'un seuil pour chacune de ces deux métriques et pour chaque pic. Au final, il a été décidé d'utiliser une approche heuristique[169] et donc un algorithme non paramétrique, basé sur l'entropie. A l'origine, cet algorithme a été conçu pour séparer l'arrière-plan (le bruit) de l'avant-plan (le signal), dans le traitement d'images[185]. Cet algorithme est appliqué à la distribution des scores de site et à la distance des sites par rapport aux sommets des pics. Cette démarche a permis de séparer les interactions directes entre l'ADN et la TF des autres sous-types de liaison et des artefacts ChIP-seq. Le pipeline est présenté dans la figure IV.1.

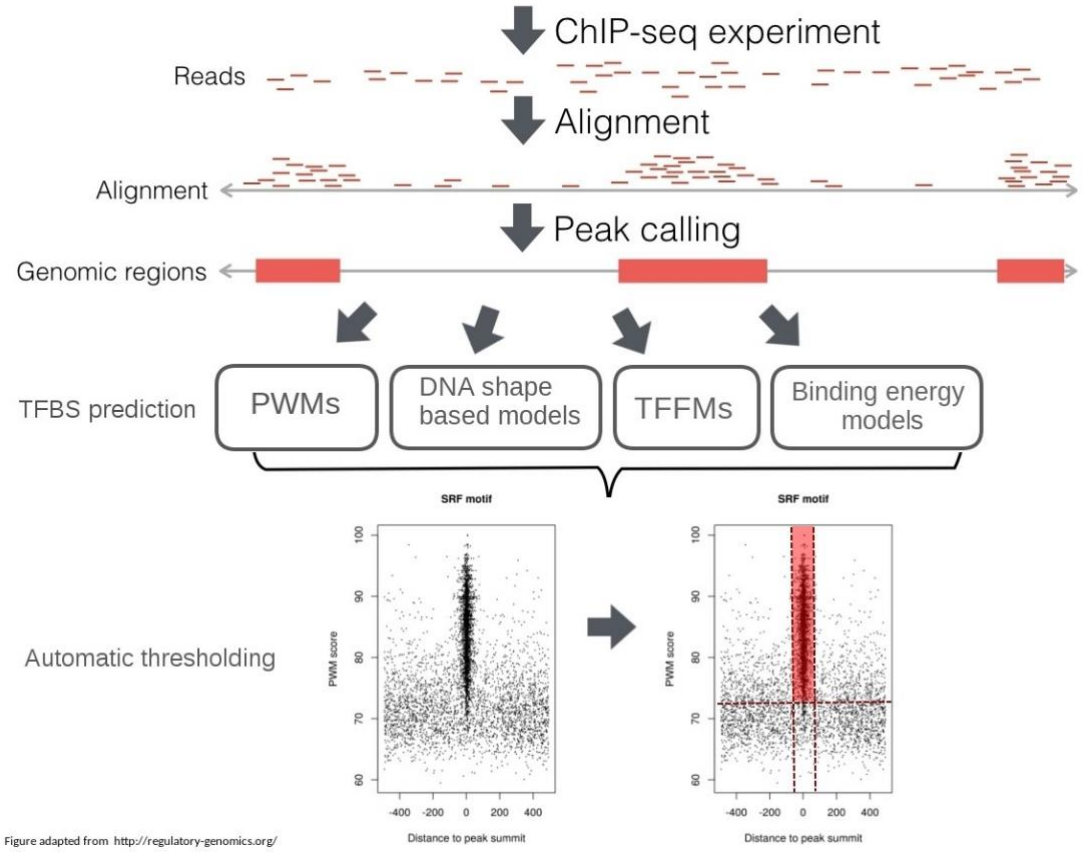


Figure IV.1 : Pipeline du modèle ChIP-eat.

Quatre types de modèles de site de fixation de TF models: DiMO-optimized PWMs, BEMs, TFFMs, and DNAshapedTFBS.

Afin de vérifier cette méthode plusieurs évaluations ont été effectuées :

- Des résultats d'expériences de microarray provenant d'UniProbe[186] ont été utilisés pour valider les interactions protéines/ADN prédites.
- La cohérence des résultats a été vérifiée en comparant les sites prédits pour les mêmes TFs
- Nous avons émis l'hypothèse que le signal ChIP-seq des pics contenant une interaction directe était susceptible d'être plus élevé que le signal des autres pics.

Cette méthode a aussi été testée sur des expériences de ChIP-exo. Le code source est disponible sur internet (https://bitbucket.org/CBGR/chip-eat/src/master).

C. Base de données Unibind

Au final, une liste d'interactions directes prédites par ChIP-eat est disponible. Ces résultats sont présentés sous la forme de fichiers BED. Ils sont aussi disponibles par TF et par lignée cellulaire sur le site UniBind disponible à l'adresse : https://unibind.uio.no. Le site Unibind est similaire à celui de JASPAR. Il a été développé à l'aide de Django. Une fonction de recherche permet d'en parcourir les résultats.

La liste des interactions directes de protéine/ADN a permis l'annotation d'éléments cis-régulateurs (CRE). Suite à l'observation montrant que les CREs possédaient des clusters de TFBS, les régions possédant un grand nombre d'interactions directes ont été annotées comme CRE. L'accès à une liste des sommets des pics de ChIP-seq est également proposé. Cette liste permet de visualiser les régions susceptibles de contenir des interactions directes n'ayant pas été identifiées avec ChIP-eat.

Published online 4 December 2018

Nucleic Acids Research, 2019, Vol. 47, No. 4 ***e21***
doi: 10.1093/nar/gky1210

A map of direct TF–DNA interactions in the human genome

Marius Gheorghe [1], Geir Kjetil Sandve [2], Aziz Khan [1], Jeanne Chèneby[3], Benoit Ballester [3] and Anthony Mathelier [1,4,*]

[1]Centre for Molecular Medicine Norway (NCMM), University of Oslo, Oslo, Norway, [2]Department of Informatics, University of Oslo, Oslo, Norway, [3]Aix Marseille Université, INSERM, TAGC, Marseille, France and [4]Department of Cancer Genetics, Institute for Cancer Research, Radiumhospitalet, Oslo, Norway

Received August 18, 2018; Revised October 31, 2018; Editorial Decision November 18, 2018; Accepted November 20, 2018

ABSTRACT

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is the most popular assay to identify genomic regions, called ChIP-seq peaks, that are bound *in vivo* by transcription factors (TFs). These regions are derived from direct TF–DNA interactions, indirect binding of the TF to the DNA (through a co-binding partner), nonspecific binding to the DNA, and noise/bias/artifacts. Delineating the *bona fide* direct TF–DNA interactions within the ChIP-seq peaks remains challenging. We developed a dedicated software, ChIP-eat, that combines computational TF binding models and ChIP-seq peaks to automatically predict direct TF–DNA interactions. Our work culminated with predicted interactions covering >2% of the human genome, obtained by uniformly processing 1983 ChIP-seq peak data sets from the ReMap database for 232 unique TFs. The predictions were *a posteriori* assessed using protein binding microarray and ChIP-exo data, and were predominantly found in high quality ChIP-seq peaks. The set of predicted direct TF–DNA interactions suggested that high-occupancy target regions are likely not derived from direct binding of the TFs to the DNA. Our predictions derived co-binding TFs supported by protein-protein interaction data and defined *cis*-regulatory modules enriched for disease- and trait-associated SNPs. We provide this collection of direct TF–DNA interactions and *cis*-regulatory modules through the UniBind web-interface (http://unibind.uio.no).

INTRODUCTION

The transcription of DNA into RNA is mainly regulated through a complex interplay between proteins and the chromatin at *cis*-regulatory regions such as promoters and enhancers. Transcription factors (TFs) are key proteins specifically binding short DNA sequences, known as TF binding sites (TFBSs), to ensure transcription at appropriate rates in the correct cell types (1). Therefore, genome-wide identification of TFBSs is a critical step to decipher transcriptional regulation, and how this process is altered in diseases (2).

Classically, genome-wide *in vivo* TF binding regions are identified through the chromatin immunoprecipitation followed by sequencing (ChIP-seq) assay (3). The genomic regions obtained with ChIP-seq, the so-called ChIP-seq peaks, are usually a few hundred base pairs (bp)-long and should encompass the TFBSs (∼10 bp-long), where direct TF–DNA interactions occur. However, ChIP-seq peaks derive from either direct TF–DNA interactions, protein-protein interactions with other regulators such as co-factors, or unspecific binding. Moreover, ChIP-seq experiments are prone to artifacts and delineating *bona fide* TF-bound regions is still an ongoing challenge (4–6) (Wreczycka *et al.*, bioRxiv, 10.1101/107680).

As TFs specifically recognize DNA sequence motifs, computational tools have been instrumental in the prediction and characterization of direct TF–DNA interactions (7). TFBSs are commonly modelled with position weight matrices (PWMs), which represent the probability of each nucleotide to be present at each position within bona fide TFBSs (7). While PWMs work well (8), more sophisticated approaches have recently been designed to model complex features of TF–DNA interactions captured by next-generation sequencing data (e.g. (9–13)). However, the best performing model varies for different TFs or TF families (8,14,15).

While multiple resources collecting TF binding regions derived from ChIP-seq exist (16–19), a limited number store genome-wide identification of TFBSs (17,20,21). The TFBS Conserved Track of the UCSC Genome Browser combined phylogenetic sequence conservation and PWMs to identify TFBSs (22) while the MANTA resource (23) integrated ChIP-seq peaks from ReMap (16) with PWMs from JASPAR (24) for TFBS predictions. A strong limitation of these approaches is that they use the same pre-defined score

*To whom correspondence should be addressed. Tel: +47 228 40 561; Email: anthony.mathelier@ncmm.uio.no

thresholds for all PWMs and all data sets. The ORegAnno database provides TFBSs obtained through literature curation (21), but the number of TFBSs available for human is limited to ∼8000.

A previous study showed that ChIP-seq data sets fall within one of three categories: (i) data sets enriched for the TF canonical binding motif close to the ChIP-seq peak summit (where the highest number of ChIP-seq reads map), (ii) data sets lacking enrichment for the canonical binding motif close to the peak summit and (iii) data sets having a combination of peaks with and without the TF canonical binding motif proximal to the peak-summit (25). Most ChIP-seq data sets were observed in category (iii). As direct TF–DNA interactions are expected to be enriched at ChIP-seq peak summits (25–30), Worsley Hunt *et al.* developed a heuristic approach specifically based on PWMs to automatically identify, in each ChIP-seq data set, this enrichment zone. The method determines the thresholds on the PWM scores and distances to the peak summits delimiting the enrichment zone that contains direct TF–DNA interactions. However, this method does not work with some more recent TFBS computational models (15,31,32).

In this study, we mapped direct TF–DNA interactions in the human genome in a refined manner by capitalizing on uniformly processed TF ChIP-seq data sets and computational tools modelling TFBSs. We provide (i) a new software to predict direct TF–DNA interactions within ChIP-seq peaks along with (ii) genome-wide predictions of such interactions in the human genome. Using an entropy-based algorithm, we have developed ChIP-eat, a tool that automatically identifies direct TF–DNA interactions using both ChIP-seq peaks and any computational model for TFBSs. We applied ChIP-eat to 1983 human ChIP-seq peak data sets from the ReMap database (16), accounting for 232 distinct TFs. The set of predicted direct TF–DNA interactions derived from PWMs covers >2% of the human genome. To make this resource available to the community, we have created UniBind (http://unibind.uio.no/), a web-interface providing public access to the predictions. We validated *a posteriori* these TFBS predictions using protein binding microarray (33) and ChIP-exo (34) data, and multiple ChIP-seq peak-callers. We used these TFBSs to (i) confirm that hotspots of ChIP-seq peaks (also known as high occupancy target regions (35)) are likely not derived from direct TF–DNA interactions, (ii) predict co-binding TFs and (iii) define *cis*-regulatory modules, which are enriched for disease- and trait-associated SNPs.

MATERIALS AND METHODS

ChIP-seq data

The ChIP-seq data sets considered were retrieved, processed, and classified as part of the last update (2018) of the ReMap database (16) (Supplementary Figure S1).

TF binding profiles

For 1983 ChIP-seq data sets used in the last ReMap update, we were able to manually assign TF binding profiles corresponding to the ChIP'ed TFs as position frequency matrices (PFMs) from the JASPAR (2018) database (24).

Training data sets

To train the TFBS computational models (see below), we considered 101 bp sequences centered around the peak summits as positive training sets. When required for training, negative training sets were obtained by shuffling the positive sequences using the *g* subcommand of the BiasAway (version 0.96) tool to match the %GC composition (25).

TFBS computational models

Position weight matrices. JASPAR PFMs were converted to PWMs as previously described in (36). For each ChIP-seq data set, PWMs were optimized using DiMO (version 1.6; default parameters with a maximum of 150 optimization steps) using the corresponding training sets (37). For TFBS predictions, we considered PWM *relative* scores, which were computed as $relative\ score = 100 \times (absolute\ score - min)/(max - min)$ where *absolute score* corresponds to the PWM absolute/raw score and *min* and *max* to the minimal and maximal absolute/raw PWM scores, respectively.

Binding energy models. JASPAR PFMs were converted to binding energy models (BEMs; (32)) using the implementation from the MARS Tools (https://github.com/kipkurui/MARSTools; Kibet and Machanick, bioRxiv, doi:10.1101/065615). We modified the implementation to return a BEM score corresponding to 1 – (original score) to consider the best site of the DNA sequence as the one with the highest BEM score (instead of the lowest one).

Transcription factor flexible models. First-order transcription factor flexible models (TFFMs) (version 2.0) were initialized with the DiMO-optimized PFMs and trained with default parameters (https://github.com/wassermanlab/TFFM; (31)) on the positive training sets.

DNAshapedTFBS models. The DNA shape-based models were trained on the training sets using the DNAshapedTFBS tool (version 1.0; https://github.com/amathelier/DNAshapedTFBS/; (15)). We trained three types of DNAshapedTFBS models with the following features: (i) DiMO-optimized PWM + DNA shape, (ii) first-order TFFM + DNA shape and (iii) 4-bits encoding + DNA shape following (15). We considered the first and second order DNA shape features helix twist, propeller twist, minor groove width, and roll with values extracted from GBShape (38).

Landscape plots

Each TFBS computational model was applied to each ChIP-seq data set independently. Following the strategy described in (25), we considered 1001 bp sequences centered around the peak summits, obtained using the bedtools (version 2.25) *slop* subcommand (39). The trained computational models were used to extract the best (maximal score) site per 1001 bp ChIP-seq peak region. For each ChIP-seq data set, landscape plots were constructed from the corresponding sites following the TFBS_Visualization tool (25). These scatter plots were also converted into heat maps using the *kde2d* function from the MASS R package (40).

Automated identification of the enrichment zone

To define the enrichment zone for each landscape plot, we automatically identified the thresholds for the TFBS computational model scores and distances to peak summits using the entropy-based algorithm from (41). The algorithm aims at identifying two classes of elements. Given a histogram, the algorithm selects the threshold that maximizes the within-class sum of the Shannon entropies for the elements in two classes (42). The two classes of elements identified are defined by the elements with values (i) above and (ii) below the threshold, respectively. This procedure optimally separates the input elements in two classes. Given a ChIP-seq data set, we applied the algorithm to the histograms of the TFBS computational model scores and distances to peak summits, independently. The maximum entropy implementation of the algorithm available in ImageJ (43) was used with default parameters.

The source code of the ChIP-eat software used to process ChIP-seq peak data sets to predict direct TF–DNA binding events is freely available at https://bitbucket.org/CBGR/chip-eat. Specifically, ChIP-eat trains a TFBS computational model and automatically defines the enrichment zone in the landscape plots to predict the underlying direct TF–DNA interactions. The identification of the enrichment zone has been applied to each TF ChIP-seq peak data set independently, allowing for the automatic detection of the thresholds that are specific to each data set with each TFBS computational model. Note that only the best hit per ChIP-seq peak has been considered to identify the enrichment zones and for all the downstream analyses.

Assessing the robustness of the enrichment zone identification

Random noise. For each ChIP-seq data set, we sampled the set of peaks using the seqtk (version 1.0) (https://github.com/lh3/seqtk) *sample* subcommand. The sequences of the sampled peaks were shuffled using the *fasta-shuffle-letters* subcommand of the MEME suite (version 4.11.4) (44) and added to the original set of ChIP-seq peaks. The automatic thresholding algorithm was applied to this new set. We tested the addition of shuffled peaks representing 10%, 25%, and 50% of the original set peaks.

Window size variability. For each ChIP-seq data set, we considered the region around the peak summit by extending with 300, 400, and 500 bp on each side using the bedtools *slop* subcommand. We considered ChIP-seq data sets where at least one TFBS was predicted within the enrichment zones obtained for all three window sizes.

Comparison with the heuristic approach to predict the enrichment zone. ChIP-eat was compared to the heuristic approach described in (25) and implemented in the TFBS_Visualization tool https://github.com/wassermanlab/TFBS_Visualization using the default parameters. The centrality of the TFBSs within the enrichment zones predicted by ChIP-eat and TFBS_Visualization was assessed using centrality *P*-value computations as described in the CentriMo tool (27). The statistical difference between the centrality *P*-values obtained with the heuristic method and ChIP-eat was assessed using a Mann-Whitney signed-rank test.

Genome coverage. The entire set of predicted TFBSs (within enrichment zones) was concatenated and then sorted using the *cat* and *sort* commands of the Unix operating system. The resulting set of locations was merged using the bedtools *merge* subcommand with default parameters. The genome coverage of the corresponding merged and non-overlapping positions was calculated as the percentage of the total number of nucleotides covered out of the total number of nucleotides in the hg38 version of the human genome.

TF–DNA binding affinity assessment with protein binding microarray data. Protein binding microarray (PBM) (45) data were retrieved from UniProbe (http://the_brain.bwh.harvard.edu/uniprobe/; (46)) for 40 TFs with available ChIP-seq data. For each ChIP-seq data set landscape plot, we extracted the DNA sequences at the sites within and outside of the predicted enrichment zone. The binding affinity of a TF to each site was computed as the median PBM intensity value of all the de Bruijn sequences containing the site sequence. The statistical difference between the distribution of PBM binding affinities from sites within and outside the enrichment zone was assessed using a two samples Mann-Whitney U test (47) implemented in the R package *stats*. A Bonferroni correction was applied to the computed *P*-values. The *P*-value density plot in Figure 3B was generated with the *density* R function with default parameters and the corresponding computed bandwidth was used to plot Supplementary Figure S10.

ChIP-exo data. ChIP-eat was applied with DiMO-optimized PFMs to the ChIP-exo data sets from (48), which were lifted over to hg38 using the liftOver tool (20). As for ChIP-seq peaks, we considered 1 001 bp regions centered around the peak summits.

ChIP-seq peaks from HOMER and BCP peak-callers. We successfully applied the HOMER (version 4.7.2) (49) and BCP (version 1.1) (50) peak-callers to 670 ENCODE ChIP-seq data sets (Supplementary Table S1). ChIP-eat was applied to the corresponding ChIP-seq peak regions with DiMO-optimized PFMs as described above. ChIP-seq peaks predicted to contain a direct TF–DNA interaction or not (using the enrichment zones) from the three peak-callers (MACS2 (51), HOMER, and BCP) were overlapped using the bedtools *intersect* subcommand. Hypergeometric tests were performed to assess the significance of the intersections using the R *phyper* function for every combination of two peak-callers with the following contingency matrix:

number of overlapping peaks **with** TFBSs from two peak-callers - 1	number of peaks **without** TFBSs from the two peak-callers
number of peaks **with** TFBSs from the two peak-callers	number of overlapping peaks from the two peak-callers

HOT/XOT regions. The high occupancy target (HOT) and extreme occupancy target (XOT) regions in all contexts were downloaded through the ENCODE data portal at http://encode-ftp.s3.amazonaws.com/modENCODE_VS_ENCODE/Regulation/Human/hotRegions/maphot_hs_selection_reg_cx_simP05_all.bed and http://encode-ftp.s3.amazonaws.com/modENCODE_VS_ENCODE/Regulation/Human/hotRegions/maphot_hs_selection_reg_cx_simP01_all.bed. ChIP-seq peaks were overlapped with the HOT/XOT regions using the bedtools *intersect* subcommand. The enrichment for overlap was assessed with a hypergeometric test using the R *phyper* function with the following contingency matrix:

number of peaks **without** TFBSs overlapping HOT/XOT regions - 1	number of peaks **with** TFBSs
number of peaks **without** TFBSs	total number of peaks

Identification of TFs with co-localized TFBSs. For each pair of distinct TFs (TF_A, TF_B), we extracted the closest TFBS associated with TF_B for each TFBS associated with TF_A and computed the geometric mean distance between midpoints of the paired TFBSs. With this approach, the geometric mean m_{AB} for the pair (TF_A, TF_B) is different from the geometric mean of the pair (TF_B, TF_A). With 232 TFs available in our analyses, we computed geometric means for 53 592 ordered pairs of TFs.

The colocalization of TFBSs for each TF pair was assessed using a Monte Carlo-based approach as follows. The number of TFBSs per TF ranged from 1 to 404 566, with 455 as the fifth percentile. We uniformly discretized the range [455, 414 172] to consider 50 TFBS set sizes (S_i for i in [1, 50]). We chose 414 172 as the maximum value to be able to compute a P-value for the set of 404 566 TFBSs. For each set size S_i, we created 500 sets of TFBSs by randomly selecting TFBSs from the total pool. Using these random sets, we computed null distributions for 500 Monte Carlo samples of geometric mean distances for each of the 2601 set size combinations. Specifically, this computation led to 2601 distributions of 500 geometric means. For the TF pair (TF_A, TF_B) with N_A and N_B TFBSs, respectively, we extracted the Monte Carlo sample of geometric mean distances M obtained from the random sets with S_A and S_B TFBSs, where $S_A = \min(S_i)$ with $S_i > N_A$ and $S_B = \min(S_i)$ with $S_i > N_B$. The empirical P-value associated with the pair (TF_A, TF_B) was computed as the number of times we observed a geometric mean smaller than m_{AB} from M over the 500 precomputed geometric means; if no smaller geometric mean was observed, the empirical P-value is defined as <0.002 (i.e. 1/500).

Since the expected geometric mean distance increases with a decreasing number of TFBSs, this P-value computation is conservative (under-estimated significance). The obtained P-values were corrected for multiple testing using the Benjamini–Hochberg method (52), only the TF pairs with a FDR $<5\%$ were considered significant.

The detailed null distribution values can be downloaded and reproduced at https://hyperbrowser.uio.no/geirksa_sandbox/u/gsandve/h/null-distributions-for-manuscript-a-map-of-direct-tf-dna-interactions-in-the-human-genome. These computations are based on running the static methods 'ConcatenateNullDistributionsTool.execute' and 'ComputeNullDistributionForEachCombinationFromSuiteVsSuiteTool.execute' (with argument values corresponding to parameter settings annotated in the Galaxy (53) history above) in the code provided at https://hyperbrowser.uio.no/geirksa_sandbox/static/hyperbrowser/files/div/hb.zip. The source code for the comparison with null distributions is available at https://bitbucket.org/CBGR/co-binding/.

GeneMANIA. We used the GeneMANIA software (54) to extract known protein–protein interactions from the list of TFs with significant co-localized TFBSs and plot the corresponding network.

Prediction of cis*-regulatory modules.* The TFBSs predicted by ChIP-eat were sorted and merged using the bedtools *sort* and *merge* subcommands. The CREAM tool (Madani Tonekaboni *et al.*, bioRxiv, doi:10.1101/222562) was applied to the merged TFBSs to define *cis*-regulatory modules (CRMs) as genomic regions enriched for clusters of TFBSs.

GWAS trait- and disease-associated single nucleotide polymorphism enrichment analysis. We assessed the enrichment for GWAS trait- and disease-associated single nucleotide polymorphisms (SNPs) at CRMs using the *traseR* R package (version 1.10.0 (55)). CRM genomic positions were lifted over to the hg19 version of the human genome to perform the analyses. The set of SNPs (as of 30 April 2018) considered by *traseR* combined data from dbGaP (56) and NHGRI (57) as described in the corresponding bioconductor package vignette (https://bioconductor.org/packages/release/bioc/vignettes/traseR/inst/doc/traseR.pdf).

Conservation analysis. The hg38 phastCons (58) scores for multiple alignments of 99 vertebrate genomes to the human genome were retrieved as a bigWig file at http://hgdownload.cse.ucsc.edu/goldenpath/hg38/phastCons100way/hg38.phastCons100way.bw. The TFBSs predicted by ChIP-eat were sorted and merged using the bedtools *sort* and *merge* subcommands. The locations overlapping CRMs were obtained using the bedtools *intersect* subcommand. The corresponding genomic locations (for all TFBSs and TFBSs in CRMs) in BED format were decomposed into 1 bp intervals using bedops v.2.4.14 (59) with the –chop 1 option. The phastCons scores at every bp were extracted with the *ex* subcommand of the bwtool (60) using the corresponding BED and phastCons bigWig files.

The UniBind web interface. All the TFBS predictions, corresponding ReMap ChIP-seq peaks, trained TFBS computational models, and CRMs are available through the UniBind database at http://unibind.uio.no/. The UniBind web interface was developed in Python using the model-view-controller framework Django. It uses MySQL to store TFBS metadata and Bootstrap as the frontend template engine. The source code is available at https://bitbucket.org/CBGR/unibind.

Statistical analyses. All statistical analyses were performed in the R environment (version 3.4.4).

RESULTS

Predicting direct TF–DNA interactions in the human genome from ChIP-seq data

Given a set of ChIP-seq peaks and a TFBS computational model such as a PWM, one can extract the best site per peak, which corresponds to the DNA subsequence of the peak with the highest score for the model. The higher the score, the stronger the computational evidence that the site is similar to TFBSs known to be bound by the TF (36). Moreover, it has been shown that the closer the site to the peak summit, the more likely it is to represent a direct TF–DNA interaction with experimental evidence from the ChIP-seq assay (25,27,30). Hence, direct TF–DNA interactions captured by ChIP-seq are enriched for high scores and small distances to the peak summits (Figure 1A,B). These characteristics have previously been used to automatically predict direct TF–DNA interactions by selecting score and distance thresholds defining these enrichment zones using a heuristic approach (25). This approach used pre-defined parameter values and was specifically designed for PWMs, but is not applicable to more recent TFBS computational models such as binding energy models (BEMs) (32), transcription factor flexible models (TFFMs) (31), and DNA shape-based models (DNAshapedTFBS) (15).

We aimed to predict direct TF–DNA interactions (TFBSs) within ChIP-seq peaks and developed the ChIP-eat software that automatically identifies the enrichment zone for any TFBS computational model. It uses a non-parametric, entropy-based algorithm originally designed to separate background/noise from foreground/signal in image processing (41) (Supplementary Figure S2). We applied this algorithm to the distributions of site scores and distance to peak summits independently to separate direct TF–DNA interaction events from other binding subtypes and ChIP-seq artifacts (Figure 1C,D; Materials and Methods). The two thresholds define the enrichment zone, which delimits the sites that are predicted as TFBSs with both experimental and computational evidence of direct TF–DNA interactions. With this approach, we automatically adjust the enrichment zone discovery specifically for each TF ChIP-seq peak data set and for each computational model. The identified enrichment zone defines the thresholds on the TFBS computational model scores and distances to the peak summits in a data set-specific manner.

We retrieved 1983 ChIP-seq peak data sets from ReMap (16), accounting for 232 TFs with a PFM available in the JASPAR database (24). Using DiMO-optimized PWMs, we compared the enrichment zones predicted by ChIP-eat with the ones obtained with the heuristic approach developed in (25). The enrichment zones predicted with ChIP-eat were more stringent than with the heuristic algorithm (Supplementary Figure S3A,B,D,E). The corresponding TFBSs predicted in the enrichment zones were more central to the peak summits with ChIP-eat than with the heuristic method as evaluated with CentriMo (27) (Supplementary Figure S3C, F). Moreover, ChIP-eat does not require any fixed values such as a predefined bin size (25) to predict the enrichment zones. Finally, ChIP-eat is not restricted to work with PWMs only and can be used with any TFBS computational model.

We applied ChIP-eat to the 1983 human ChIP-seq data sets with four types of computational TFBS models: DiMO-optimized PWMs, BEMs, TFFMs, and DNAshapedTFBS. These models were optimized for each ChIP-seq data set, independently (see Materials and Methods). In the following analyses, we focused on the predictions obtained with the DiMO-optimized PWMs (see Materials and Methods). This set of direct TF–DNA interactions (TFBSs) extracted from the enrichment zones covers >2% of the human genome, encompassing 8 304 135 distinct TFBS locations.

Predicted direct TF–DNA interactions are likely *bona fide* TFBSs

Robustness of the enrichment zone identification. The robustness of the method was first evaluated by applying ChIP-eat to genomic regions of ±300, 400, and 500 bp around the peak summits. The median distance threshold to the peak summit shifted from 72 bp using ±500 bp to 64 and 55 using ±400 and 300 bp, respectively. The median PWM scores thresholds were 85, 84.6 and 83.9 with ±500, 400, and 300 bp regions, respectively (see Supplementary Figure S8 for a visual representation using the 10 most frequent ChIP'ed TFs). The variability of the predicted enrichment zone when using different window sizes is similar to the variability between ChIP-seq data sets for the same TF (see below). Further, the number of predicted TFBSs within the enrichment zones were similar when using the different region sizes (Supplementary Figure S9). These analyses confirmed the robustness of the entropy-based thresholding algorithm to the window size considered. As previously used in (25), we considered the ±500 bp regions around the peak summits in the following analyses.

Considering the ChIP-seq data sets for the 10 most frequently ChIP'ed TFs, we observed that the thresholds on the PWM scores and distances to peak summits, defining the enrichment zones, were consistent between data sets for the same TF (Figure 2A,B). Namely, the median pairwise difference between PWM score thresholds for the same TF ranged from 1.7 to 3.7 and the median distance thresholds from 12 to 35 bp. As expected, the thresholds identified for distinct TFs are different (Figure 2C, D). Taken together, these results highlight that the entropy-based algorithm allows for the identification of enrichment zones specific to each TF and ChIP-seq data set, with consistent predictions between data sets for the same TF. Results were consistent with BEM, TFFM, and DNAshapedTFBS models (Supplementary Figures S4–S6).

We further evaluated the robustness of the method to noise by adding 10%, 25%, and 50% of shuffled sequences to the initial set of ChIP-seq peaks for all ChIP-seq peak data sets (see Materials and Methods). The median threshold on the distances to peak summits shifted from 73 bp in the initial set of ChIP-seq peaks to 70 bp with 10% noise, 67 bp with 25% noise, and to 63 bp when adding 50% noise. The median PWM score threshold was 85.2 for the initial set of ChIP-seq peaks and shifted to 85 when adding 10%

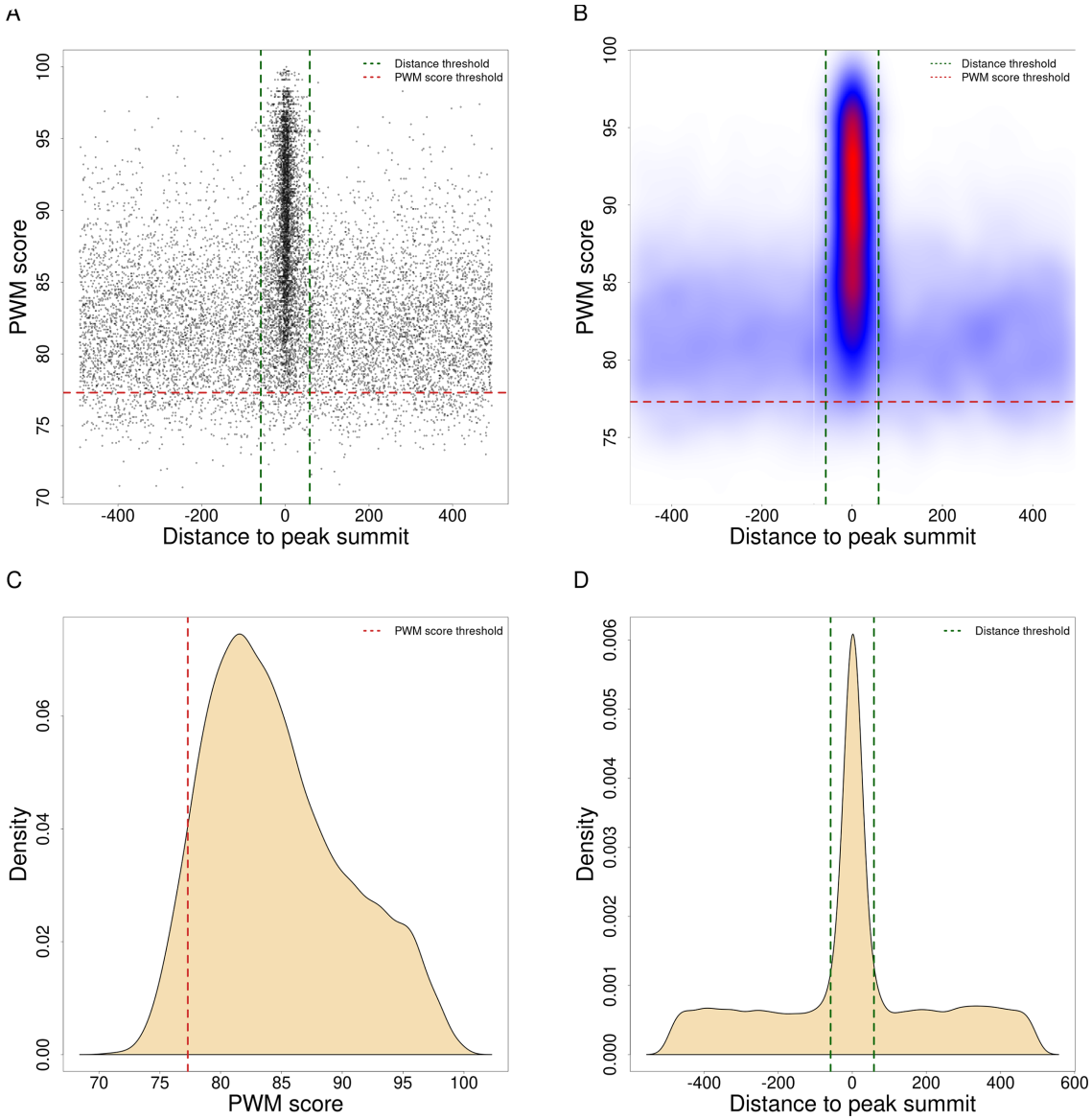


Figure 1. Automatic detection of the TFBS enrichment zone. Landscape plots (25) obtained with SRF ChIP-seq peaks using the DiMO-optimized PWM MA0083.3 from JASPAR are presented as scatter (**A**) and heatmap (**B**) plots. The enrichment zone (defined within the red and green dashed line boundaries, A-B) is automatically obtained by ChIP-eat with thresholds on PWM scores (red dashed lines; **C**) and distances to peak summits (green dashed lines; **D**). The enrichment zone provides TFBSs in ChIP-seq peaks (points in A) with supporting evidence for direct TF–DNA binding from the ChIP-seq assay (close distance to peak-summits, A-B, x-axis) and the computational model (PWM score, A-B, y-axis). Distances to peak summits in A, B and D are provided using a base pair unit.

of noise, to 84.8 when adding 25% of noise, and to 84.4 when adding 50% of noise. A visual representation for the 10 most frequently ChIP'ed TFs is available in Supplementary Figure S7. The variability of the thresholds defining the enrichment zones when adding noise is limited, within the range of variability between ChIP-seq peak data sets for the same TF (Figure 2). Taken together, these results show that the entropy-based thresholding algorithm delimiting the enrichment zones, as implemented in ChIP-eat, provides consistent results between data sets for the same ChIP'ed TF and is robust to the window sizes considered and random noise.

Validation using in vitro DNA binding affinities. To confirm *a posteriori* the high quality of our set of TFBS predictions, we assessed the TF binding affinity to DNA sequences derived experimentally from protein binding microarrays (PBM) (61). The PBM assay quantifies the binding affinity of a protein to all possible combinations of 8-mer DNA sequences. We retrieved PBM data from the UniPROBE database (46) for 40 different TFs present in our collection, corresponding to 249 ChIP-seq data sets (Supplementary Table S2). Note that the JASPAR PFMs for the ATF1, ATF3, and FOXJ2 TFs were originally derived from PBM data. For each ChIP-seq data set, we tested if the sites located in the enrichment zone presented higher binding affinity than sites outside (see Materials and Methods). The

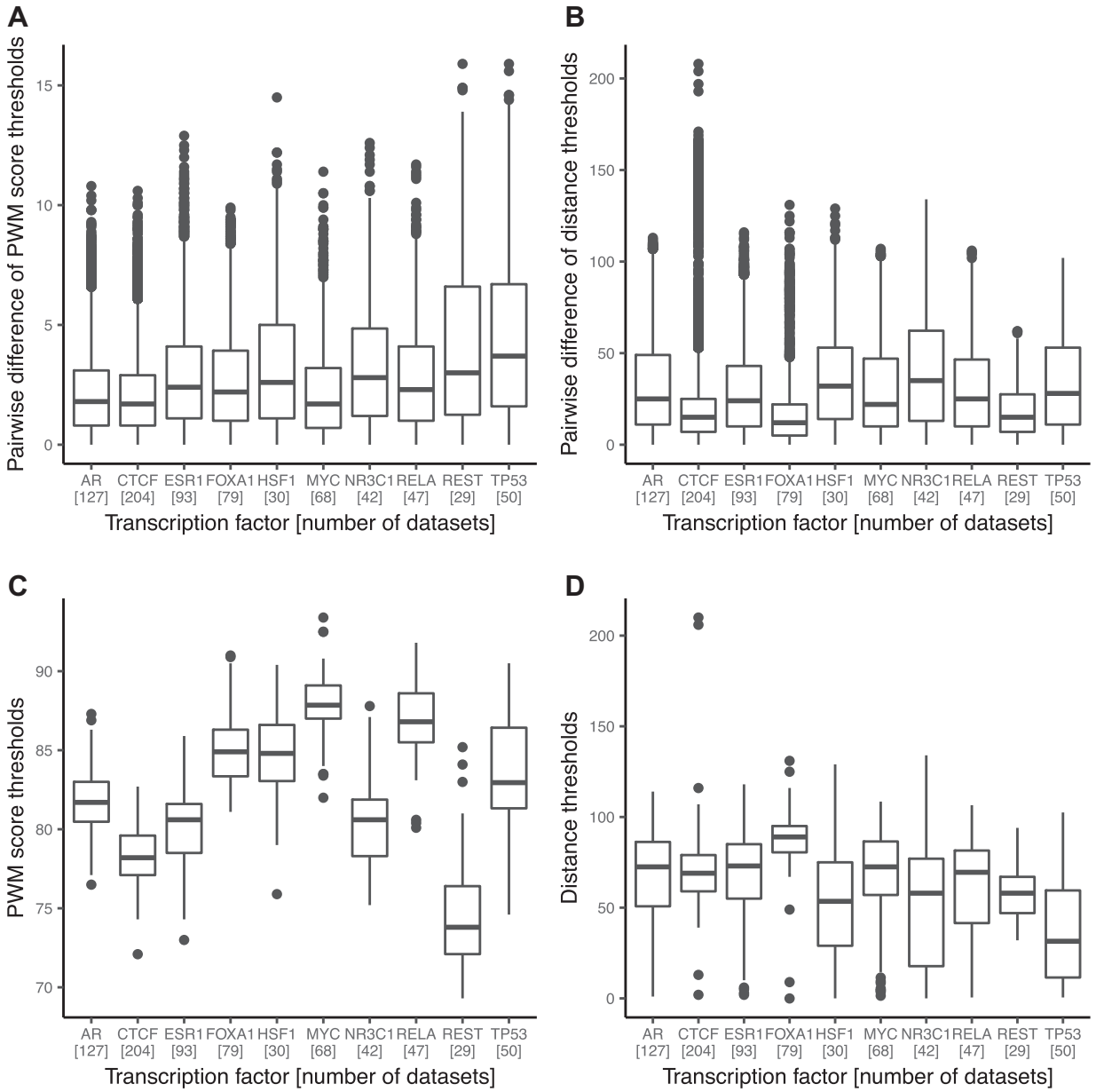


Figure 2. Assessment of the thresholds predicted by ChIP-eat across data sets. Boxplots of the pairwise differences for DiMO-optimized PWM score thresholds and distances to peak summits thresholds between ChIP-seq data sets for the same TF are provided in panels (**A**) and (**B**), respectively. Absolute variations of DiMO-optimized PWM score thresholds and distances to the peak summits within all data sets for the same TF are provided in panels (**C**) and (**D**), respectively. The ten TFs with the highest number of data sets were selected; the number of data sets for each TF is provided between brackets.

distributions of the binding affinity scores for sites within and outside the enrichment zones were compared using a Mann-Whitney U test (Figure 3A; Materials and Methods). Predicted direct TF–DNA interactions (sites within the enrichment zone) had significantly higher binding affinity than the other sites for 75% of the data sets with *P*-value <0.01 and 81% with *P*-value <0.05 (Figure 3B). Similar results were obtained when considering BEM, TFFM, and DNAshapedTFBSs computational models (Supplementary Figure S10). This analysis emphasizes that the sites predicted in the defined enrichment zones are likely to correspond to direct TF–DNA interactions.

Predicted direct TF–DNA interactions are found in high confidence ChIP-seq peaks. We hypothesized that the ChIP-seq signal at ChIP-seq peaks containing a predicted direct TF–DNA interaction were more likely to be higher than at the other peaks. To test this hypothesis, we looked at (i) the quality of the peaks based on *P*-values assigned to the peaks by the MACS2 peak-caller and (ii) the reproducibility of calling these peaks with multiple peak-callers (MACS2, HOMER, and BCP; see Materials and Methods).

We observed that the distribution of *P*-values assigned by MACS2 to the peaks containing a predicted TFBS were significantly (*P*-value < 0.01; Mann–Whitney *U* test) lower than for the rest of the peaks for 1862 (96%) data sets (Fig-

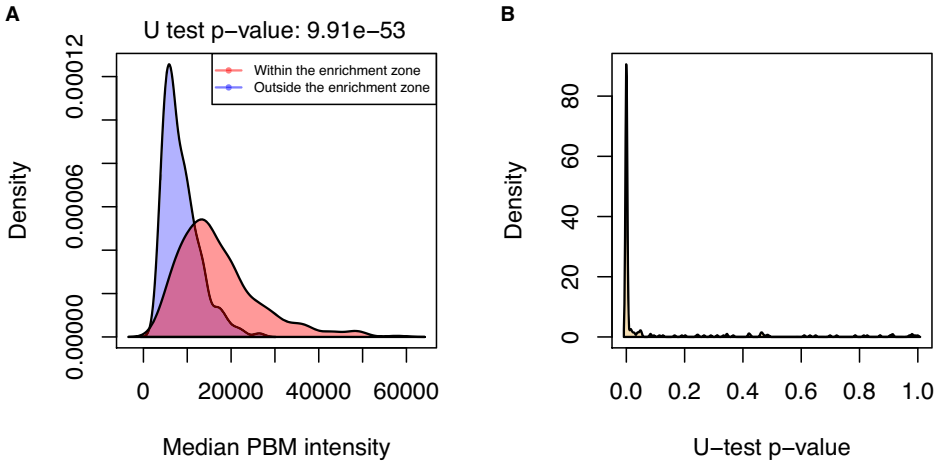


Figure 3. Binding affinity assessment for the predicted direct TF–DNA interactions. (**A**) Distribution of the median PBM intensity scores for the ENCSR000BMX GATA3 ChIP-seq data set between sequences at TFBSs (i.e. sites within the enrichment zone; in red) and sites outside the enrichment zone (in blue). (**B**) Distribution of Mann–Whitney *U* test *P*-values across the 249 data sets, showing distinct distributions of PBM intensity scores between sites within and outside the enrichment zones.

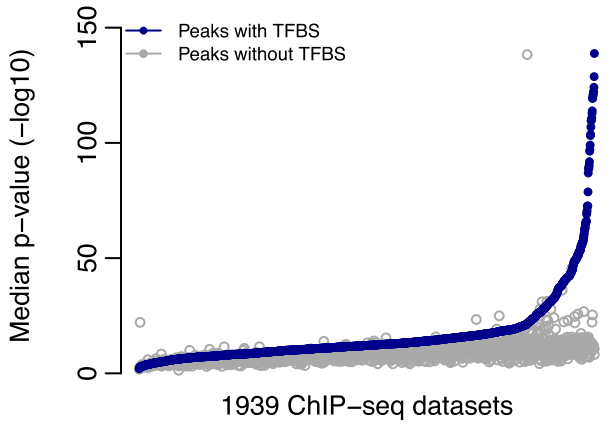


Figure 4. Quality assessment of the ChIP-seq peaks derived from direct TF–DNA interactions. Distribution of the median MACS2 *P*-values (y-axis) across all data sets. Values for peaks containing a predicted TFBS are provided in blue and values for the other peaks in grey. 1939 ChIP-seq data sets were predicted to contain direct TF–DNA interactions (x-axis).

ure 4). The other 77 data sets contained a reduced number of peaks (median of 837 compared to 18 968 for the complete set of ChIP-seq data sets), which can explain the lack of statistical significance. These results confirm that the predictions of direct TF–DNA interactions were found in ChIP-seq peaks of higher quality as assessed by MACS2.

To test ChIP-seq peak-calling reproducibility, we used two other peak-callers (HOMER and BCP) on 670 ChIP-seq data sets from ENCODE. Our choice of peak-callers was motivated by their distinct statistical approaches for peak prediction. While MACS2 and HOMER are based on an empirical model supported by a Poisson distribution, BCP uses a Bayesian approach implementing infinite-state hidden Markov models. We applied ChIP-eat to the ChIP-seq peaks to predict TFBSs. For each pair of peak-callers, we assessed whether the peaks predicted to contain a direct TF–DNA interaction were more prevalent (*P*-value < 0.01, hypergeometric test) in the set of peaks called by both peak-callers. This was observed for 63% of the data sets for MACS2 and BCP, 70% for MACS2 and HOMER, and 66% for HOMER and BCP. The data sets without significant enrichment had a median number of peaks predicted to be derived from direct TF–DNA interactions that was ∼7 fold smaller (e.g. 3358 compared to 22 499 between MACS2 and BCP) than for the data sets with significant enrichment, and a median number of peaks without TFBS ∼2 fold larger (e.g. 40 050 compared to 21 256 between MACS2 and BCP) (Supplementary Table S3). Moreover, the median quality scores assigned by the peak-callers to the peaks from the enriched data sets were significantly (*P*-value < 0.01, Mann–Whitney *U* test) higher than for the peaks in the other data sets (Supplementary Figure S11). It suggests that the data sets enriched for reproducible peaks containing predicted direct TF–DNA interactions are of better quality than the rest of the data sets.

Taken together, these results highlight that the ChIP-seq peaks in which ChIP-eat predicts direct TF–DNA interactions are of higher quality than the other peaks. Note that the ChIP-eat tool does not consider the peak quality when predicting direct TF–DNA interactions. These observations reinforce the confidence in the predicted TFBSs by ChIP-eat.

Predictions of direct TF–DNA interactions in ChIP-exo data

The ChIP-exo assay has been developed to provide a higher resolution than ChIP-seq to identify TFBSs *in vivo* (34). We aimed at assessing the performance of ChIP-eat on predicting direct TF–DNA interactions using ChIP-exo data. The ChExMix tool has recently been introduced to characterize protein-DNA binding event subtypes from ChIP-exo peak (48). ChExMix predicted different binding event subtypes for ChIP-exo data obtained for the TFs ESR1 and FOXA1, one of these subtypes corresponding to direct TF–DNA interactions (48). We applied ChIP-eat on the same ESR1 and FOXA1 ChIP-exo data sets. We compared the set of peaks identified to contain direct TF–DNA interactions

predicted by ChExMix and ChIP-eat in these two data sets. We found that 93.6% (for ESR1) and 91.3% (for FOXA1) of the peaks predicted to contain TFBSs by ChIP-eat were also predicted as direct binding events by ChExMix (Supplementary Table S4). The high overlaps between the predictions from ChExMix and ChIP-eat were confirmed by Jaccard similarity indexes of 63.7% and 68.7% for ESR1 and FOXA1, respectively. The similar results obtained with the two tools suggest that ChIP-eat, designed for the more noisy and less precise ChIP-seq data, is able to capture direct binding events from ChIP-exo data.

High-occupancy target regions are likely not derived from direct TF–DNA interactions

High-occupancy target (HOT) and extreme-occupancy target (XOT) regions are genomic regions where ChIP-seq peaks were observed for a large number of distinct ChIP'ed TFs (35,62,63). These regions are observed across species (63) and contain an unusually high frequency of ChIP-seq peaks (35,62,63). We used our set of high quality TFBS predictions to confirm that HOT/XOT regions were depleted of direct TF–DNA interactions. Indeed, we found that ChIP-seq peaks that do not contain a predicted TFBS were significantly enriched at HOT/XOT regions (odds ratio = 1.43 for HOT and 1.44 for XOT, *P*-value < 2.2e–16, hypergeometric test, Supplementary Table S5). Similar results were obtained when considering the three other computational models (BEM, TFFM, and DNAshapedTFBSs; Supplementary Table S5). This observation, combined with a previous study describing that HOT/XOT regions are likely to be derived from ChIP-seq artifacts (Wreczycka *et al.*, bioRxiv, 10.1101/107680), suggests that HOT/XOT regions are not derived from the direct binding of the ChIP'ed TFs.

Predicted direct TF–DNA interactions reveal co-binding TFs and cis-regulatory modules enriched for disease- and trait-associated SNPs

TFs are known to collaborate through specific co-binding at *cis*-regulatory modules (CRMs) to achieve their function (1,36). Hence, identifying co-binding TFs is critical to decipher transcriptional regulation of gene expression. We aimed at using our predicted direct TF–DNA interactions to reveal co-binding TFs and CRMs. We hypothesized that the distances between TFBSs of cooperating TFs are smaller than expected by chance. We tested this hypothesis for all pairs of TFs for which we predicted TFBSs (232 TFs, 53 592 pairs tested; see Materials and Methods). For each TF pair, we used a conservative Monte Carlo-based approach to compare the geometric mean of the distances between their TFBSs to the geometric mean distance expected by chance for a similar number of TFBSs randomly selected from the complete pool of TFBSs (see Materials and Methods). This approach predicted 150 pairs of TFs (accounting for 112 distinct TFs) with TFBSs closer in the genome than expected by chance (FDR < 5%; Supplementary Table S6). For 82% of the predicted TF pairs, we confirmed that the corresponding TFs physically interact using the protein-protein interaction networks from the GeneMANIA tool (54) (Supplementary Figure S12). This analysis further supports the biological relevance of the TFBSs predicted by ChIP-eat.

Next, we aimed to automatically identify CRMs, which correspond to clusters of direct TF–DNA interactions, using the clustering of genomic regions analysis method (CREAM; (Madani Tonekaboni *et al.*, bioRxiv, doi:10.1101/222562)). When considering our complete set of TFBSs, CREAM detected 61 934 CRMs in the human genome, encompassing 2 474 587 distinct TFBS locations. We found that the predicted CRMs were significantly enriched (FDR-corrected *P*-value = $2.9e^{-150}$) for disease- and trait-associated SNPs using traseR (55). Further, we observed that the TFBSs lying within the CRMs were more conserved than the TFBSs predicted outside (Supplementary Figure S13). Taken together, these results indicate a potentially functional role of the CRMs identified as clusters of direct TF–DNA interactions.

The UniBind web interface to access our collection of direct TF–DNA interactions

We catalogued the complete set of TFBS predictions from each prediction model, trained models, original ChIP-seq peaks from ReMap, and computed CRMs, and made them publicly available through UniBind at http://unibind.uio.no/. UniBind provides an interactive web interface with easy browsing, searching, and downloading for all our predictions (Figure 5). For instance, users can search for predictions for specific TFs, cell lines, and conditions.

The data can be searched by using the case insensitive search option available on the homepage. The database can be searched for each of the four TF binding models, cell/tissue type, and TF name using the 'Advanced Options', available on the homepage (Figure 5A). Search results are presented in a responsive and paginated table along with metadata information (Figure 5B), which can be clicked to view the detailed information and download TFBSs, summary plots, and ReMap ChIP-seq peaks (Figure 5C-D). All the metadata in the responsive tables can be downloaded as CSV files. UniBind displays by default the results obtained with the DiMO-optimized PWMs, but results obtained from all TFBS computational models along with the trained models are available for browsing and/or download.

DISCUSSION

To summarize, we have uniformly processed 1983 ChIP-seq peak data sets to predict high quality direct TF–DNA binding interactions in the human genome. The predictions were obtained using a non-parametric, entropy-based algorithm that automatically detects thresholds for TFBS computational model scores and distances to peak summits for each ChIP-seq data set. This new approach identified TFBSs supported by strong experimental and computational evidences for direct TF–DNA interactions. The accuracy of the predictions was *a posteriori* validated using the PBM *in vitro* assay, ChIP-exo data, and multiple ChIP-seq peak-calling algorithms. Our set of direct TF–DNA interactions confirmed that HOT genomic regions are likely not derived from direct binding of the TFs to the DNA. We used

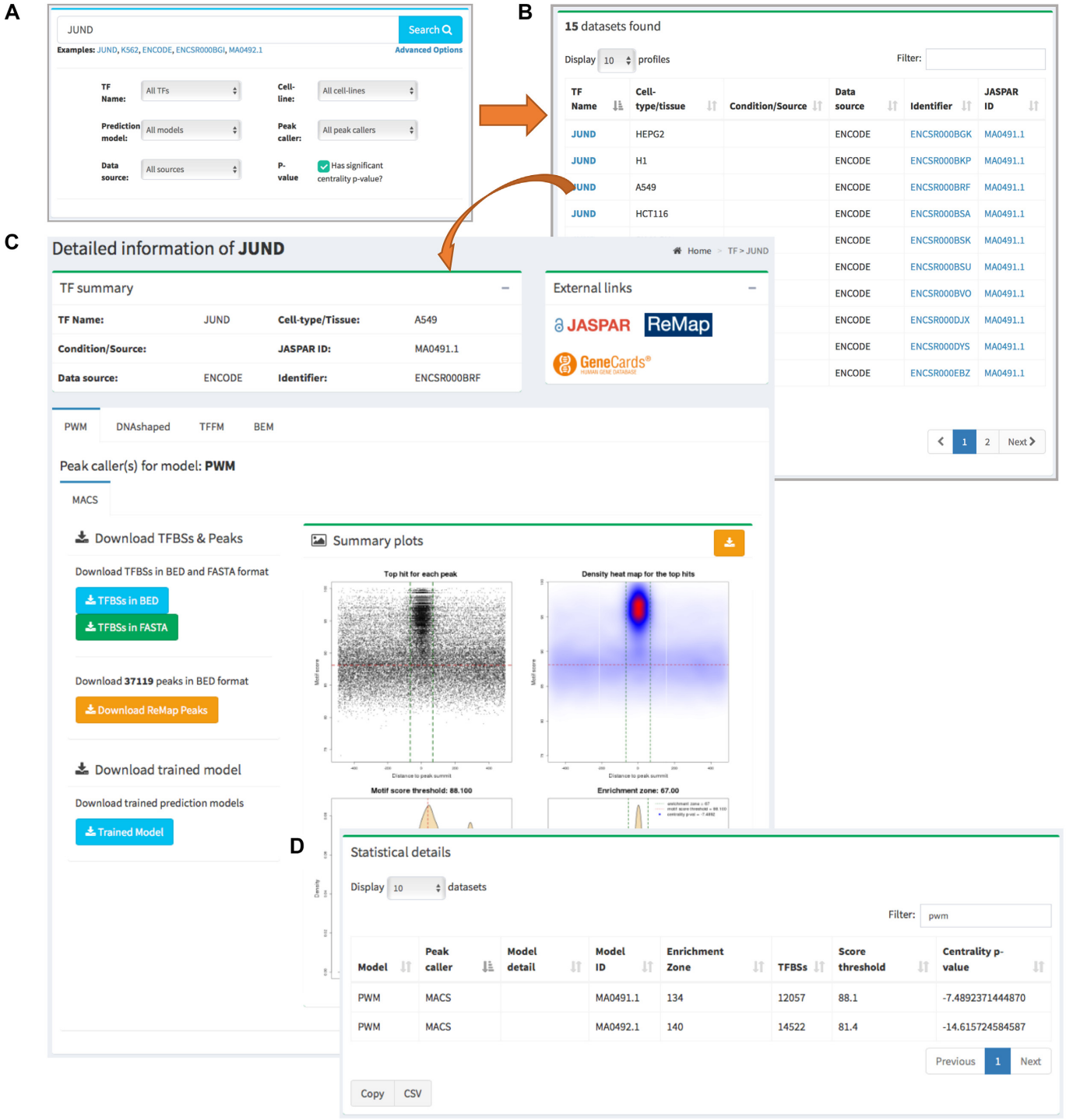


Figure 5. Overview of the UniBind user interface with interactive searching activity. (**A**) A quick and detailed search feature on the homepage. (**B**) A responsive table lists the searched data set(s), which can be clicked to view the details. (**C**) A detailed page shows the analysis for the JUND TF in cell-line A549, which is divided into sub-panels including the TF summary, external links, summary plots, and download options for each computational TFBS model. (**D**) Statistical details of the results.

our TFBSs to predict TFs with proximal binding events in the human genome, which could cooperate to achieve specific functions. Further, we defined *cis*-regulatory modules, which are clusters of TFBSs, that were enriched for disease- and trait-associated SNPs from GWAS. The complete set of predictions is publicly and freely available through the UniBind web-interface (http://unibind.uio.no/), in an effort to provide the community with an unprecedented collection of high quality direct TF–DNA interaction events in the human genome.

The output of ChIP-seq assays is generally composed of direct protein-DNA interactions, indirect binding of the protein to the DNA (through a co-binding partner), nonspecific protein binding to the DNA, and noise/bias/artifacts (4–6). Here, we specifically aimed at identifying direct TF–DNA interaction events by using an entropy-based algorithm (41). This algorithm was originally developed to discriminate between foreground and background in image processing. Hence, it assumes the presence of background (or noise) in the data. As a consequence, our approach is limited by the assumption that there is background/noise in the ChIP-seq data sets analyzed. We assume that this noise represents indirect binding of TFs, nonspecific binding, or ChIP-seq experimental artifacts. Moreover, our approach considered the best site per ChIP-seq peak (defined using TFBS computational models), which represents the best candidate. We recognize that other sites with lower scores could represent direct TF–DNA interactions. These limitations denote that our approach is stringent for the prediction of direct TF–DNA interactions, favoring specificity over sensitivity. The ChIP-seq peaks that our method did not predict to contain direct TF–DNA binding events could be further analyzed to discriminate other mechanisms for protein-DNA interactions from background noise, as proposed in the ChExMix tool established for ChIP-exo data (48).

The ChIP-eat pipeline developed for this study used four TFBS computational models to predict TF–DNA binding events. These models were specifically trained for each ChIP-seq data set to improve the quality of the predictions, as the best-performing computational model varies for different TFs or TF families (8,14,15). As a consequence, we advocate that a 'one-fits-all' TFBS prediction model is not optimal and that one should compare results from multiple models. With the predictions available through UniBind, users can assess which model would perform better for each data set. Of course, it requires to use a specific metric to compare performance. As our methods aimed at identifying enrichment zones centered around ChIP-seq peak summits, we suggest to rely on a centrality measure as implemented in the CentriMo method (27). In UniBind, we provide centrality *P*-values computed following (27) for the predictions from each model in each ChIP-seq data set. Moreover, the ChIP-eat pipeline is generalizable and users can incorporate other TFBS computational models to predict direct TF–DNA interactions and compare them to the ones already stored in UniBind.

While studies alike focus on determining where TFs directly interact with DNA, our understanding of how these TF–DNA interactions influence expression is limited. Surely, it is critical to decipher the relationship between TF–DNA interactions and transcriptional regulation (64). It is expected that a large portion of the TFBSs identified in our study are not functional, as suggested by the futility theorem (36). Nevertheless, functional TF binding events are likely to be clustered (65–68) and associated with stronger ChIP-seq peak signals (12,69). We expect that the direct TF–DNA interactions predicted in *cis*-regulatory modules and stored in UniBind are more likely to be enriched for functional events. Determining the specific set of functional TF–DNA interactions would require dedicated computational models and experiments.

DATA AVAILABILITY

Source code of the ChIP-eat software is available at https://bitbucket.org/CBGR/chip-eat and of UniBind at https://bitbucket.org/CBGR/unibind. The source code used for the identification of co-localized TFs is available at https://bitbucket.org/CBGR/co-binding. Users can browse and/or download the data through the UniBind web interface at http://unibind.uio.no/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

As research parasites (70), we would like to thank all the researchers who deposited their data. We thank Georgios Magklaras and his team for systems support, Manuela Zucknick and Andrea Cremaschi for statistical insights, Elisa Bjørgo and Ingrid Kjelsvik for management support, and Roza Berhanu Lemma, Jaime Castro-Mondragon, Oriol Fornes and Phillip Richmond for comments on the manuscript draft.

FUNDING

Norwegian Research Council (project #187615), Helse Sør-Øst, and the University of Oslo through the Centre for Molecular Medicine Norway (NCMM) (to A.M., A.K., M.G.); Ph.D. fellowship from the French Ministry of Higher Education and Research (to J.C.). Funding for open access charge: Norges Forskningsråd.
Conflict of interest statement. None declared.

REFERENCES

1. Lambert,S.A., Jolma,A., Campitelli,L.F., Das,P.K., Yin,Y., Albu,M., Chen,X., Taipale,J., Hughes,T.R. and Weirauch,M.T. (2018) The human transcription factors. *Cell*, **172**, 650–665.
2. Mathelier,A., Shi,W. and Wasserman,W.W. (2015) Identification of altered cis-regulatory elements in human disease. *Trends Genet.*, **31**, 67–76.
3. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
4. Teytelman,L., Thurtle,D.M., Rine,J. and van Oudenaarden,A. (2013) Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 18602–18607.
5. Jain,D., Baldi,S., Zabel,A., Straub,T. and Becker,P.B. (2015) Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments. *Nucleic Acids Res.*, **43**, 6959–6968.

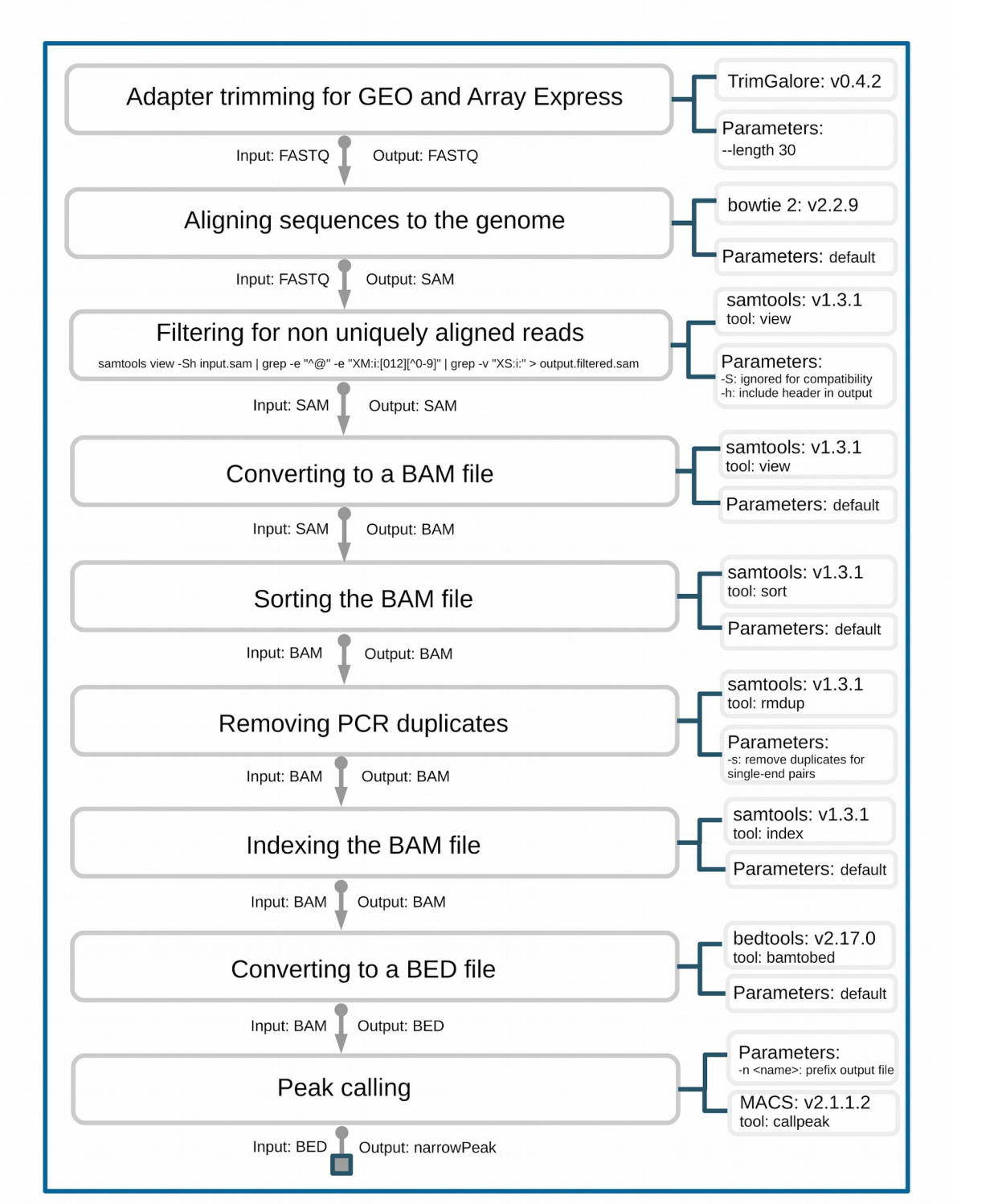
6. Worsley Hunt,R. and Wasserman,W.W. (2014) Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. *Genome Biol.*, **15**, 412.
7. Stormo,G.D. (2013) Modeling the specificity of protein-DNA interactions. *Quant Biol.*, **1**, 115–130.
8. Weirauch,M.T., Cote,A., Norel,R., Annala,M., Zhao,Y., Riley,T.R., Saez-Rodriguez,J., Cokelaer,T., Vedenko,A., Talukder,S. *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.
9. Kulakovskiy,I., Levitsky,V., Oshchepkov,D., Bryzgalov,L., Vorontsov,I. and Makeev,V. (2013) From binding motifs in ChIP-Seq data to improved models of transcription factor binding sites. *J. Bioinform. Comput. Biol.*, **11**, 1340004.
10. Eggeling,R., Roos,T., Myllymäki,P. and Grosse,I. (2015) Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data. *BMC Bioinformatics*, **16**, 375.
11. Siebert,M. and Söding,J. (2016) Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.*, **44**, 6055–6069.
12. Slattery,M., Zhou,T., Yang,L., Dantas Machado,A.C., Gordân,R. and Rohs,R. (2014) Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.*, **39**, 381–399.
13. Keilwagen,J. and Grau,J. (2015) Varying levels of complexity in transcription factor binding motifs. *Nucleic Acids Res.*, **43**, e119.
14. Yang,L., Orenstein,Y., Jolma,A., Yin,Y., Taipale,J., Shamir,R. and Rohs,R. (2017) Transcription factor family-specific DNA shape readout revealed by quantitative specificity models. *Mol. Syst. Biol.*, **13**, 910.
15. Mathelier,A., Xin,B., Chiu,T.-P., Yang,L., Rohs,R. and Wasserman,W.W. (2016) DNA shape features improve transcription factor binding site predictions in vivo. *Cell Syst.*, **3**, 278–286.
16. Chèneby,J., Gheorghe,M., Artufel,M., Mathelier,A. and Ballester,B. (2018) ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.*, **46**, D267–D275.
17. Yevshin,I., Sharipov,R., Valeev,T., Kel,A. and Kolpakov,F. (2017) GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res.*, **45**, D61–D67.
18. Zhou,K.-R., Liu,S., Sun,W.-J., Zheng,L.-L., Zhou,H., Yang,J.-H. and Qu,L.-H. (2017) ChIPBase v2.0: decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data. *Nucleic Acids Res.*, **45**, D43–D50.
19. Mei,S., Qin,Q., Wu,Q., Sun,H., Zheng,R., Zang,C., Zhu,M., Wu,J., Shi,X., Taing,L. *et al.* (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.*, **45**, D658–D662.
20. Hinrichs,A.S., Karolchik,D., Baertsch,R., Barber,G.P., Bejerano,G., Clawson,H., Diekhans,M., Furey,T.S., Harte,R.A., Hsu,F. *et al.* (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.*, **34**, D590–D598.
21. Montgomery,S.B., Griffith,O.L., Sleumer,M.C., Bergman,C.M., Bilenky,M., Pleasance,E.D., Prychyna,Y., Zhang,X. and Jones,S.J.M. (2006) ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics*, **22**, 637–640.
22. Kent,W.J. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
23. Fornes,O., Gheorghe,M., Richmond,P.A., Arenillas,D.J., Wasserman,W.W. and Mathelier,A. (2018) MANTA2, update of the Mongo database for the analysis of transcription factor binding site alterations. *Sci Data*, **5**, 180141.
24. Khan,A., Fornes,O., Stigliani,A., Gheorghe,M., Castro-Mondragon,J.A., van der Lee,R., Bessy,A., Chèneby,J., Kulkarni,S.R., Tan,G. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D1284.
25. Worsley Hunt,R., Mathelier,A., Del Peso,L. and Wasserman,W.W. (2014) Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. *BMC Genomics*, **15**, 472.
26. Guo,Y., Mahony,S. and Gifford,D.K. (2012) High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.*, **8**, e1002638.
27. Bailey,T.L. and Machanick,P. (2012) Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.*, **40**, e128.
28. Kulakovskiy,I.V., Boeva,V.A., Favorov,A.V. and Makeev,V.J. (2010) Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, **26**, 2622–2623.
29. Jothi,R., Cuddapah,S., Barski,A., Cui,K. and Zhao,K. (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, **36**, 5221–5231.
30. Wilbanks,E.G. and Facciotti,M.T. (2010) Evaluation of algorithm performance in ChIP-Seq peak detection. *PLoS One*, **5**, e11471.
31. Mathelier,A. and Wasserman,W.W. (2013) The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, **9**, e1003214.
32. Zhao,Y., Ruan,S., Pandey,M. and Stormo,G.D. (2012) Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics*, **191**, 781–790.
33. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. 3rd and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
34. Rhee,H.S. and Pugh,B.F. (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, **147**, 1408–1419.
35. Yip,K.Y., Cheng,C., Bhardwaj,N., Brown,J.B., Leng,J., Kundaje,A., Rozowsky,J., Birney,E., Bickel,P., Snyder,M. *et al.* (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.*, **13**, R48.
36. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
37. Patel,R.Y. and Stormo,G.D. (2014) Discriminative motif optimization based on perceptron training. *Bioinformatics*, **30**, 941–948.
38. Chiu,T.-P., Yang,L., Zhou,T., Main,B.J., Parker,S.C.J., Nuzhdin,S.V., Tullius,T.D. and Rohs,R. (2015) GBshape: a genome browser database for DNA shape annotations. *Nucleic Acids Res.*, **43**, D103–D109.
39. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
40. Venables,W.N. and Ripley,B.D. (2002) *Modern Applied Statistics with S Springer*, NY.
41. Kapur,J.N., Sahoo,P.K. and Wong,A.K.C. (1985) A new method for gray-level picture thresholding using the entropy of the histogram. *Comput. Vis. Graph. Image Process.*, **29**, 140.
42. Shannon,C.E. (1948) A Mathematical Theory of Communication. *Bell Syst. Tech. J.*, **27**, 623–656.
43. Schneider,C.A., Rasband,W.S. and Eliceiri,K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods*, **9**, 671–675.
44. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
45. Bulyk,M.L., Gentalen,E., Lockhart,D.J. and Church,G.M. (1999) Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nat. Biotechnol.*, **17**, 573–577.
46. Hume,M.A., Barrera,L.A., Gisselbrecht,S.S. and Bulyk,M.L. (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **43**, D117–D122.
47. Mann,H.B. and Whitney,D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, **18**, 50–60.
48. Yamada,N., Lai,W.K.M., Farrell,N., Pugh,B.F. and Mahony,S. (2018) Characterizing protein-DNA binding event subtypes in ChIP-exo data. *Bioinformatics*, doi:10.1093/bioinformatics/bty703.
49. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell*, **38**, 576–589.
50. Xing,H., Mo,Y., Liao,W. and Zhang,M.Q. (2012) Genome-wide localization of protein-DNA binding and histone modification by a

Bayesian change-point method with ChIP-seq data. *PLoS Comput. Biol.*, **8**, e1002613.
51. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
52. Hochberg,Y. and Benjamini,Y. (1990) More powerful procedures for multiple significance testing. *Stat. Med.*, **9**, 811–818.
53. Afgan,E., Baker,D., Batut,B., van den Beek,M., Bouvier,D., Cech,M., Chilton,J., Clements,D., Coraor,N., Grüning,B.A. *et al.* (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.*, **46**, W537–W544.
54. Warde-Farley,D., Donaldson,S.L., Comes,O., Zuberi,K., Badrawi,R., Chao,P., Franz,M., Grouios,C., Kazi,F., Lopes,C.T. *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**, W214–W20.
55. Chen,L. and Qin,Z.S. (2015) traseR: an R package for performing trait-associated SNP enrichment analysis in genomic intervals. *Bioinformatics*, **32**, 1214–1216.
56. Mailman,M.D., Feolo,M., Jin,Y., Kimura,M., Tryka,K., Bagoutdinov,R., Hao,L., Kiang,A., Paschall,J., Phan,L. *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.*, **39**, 1181–1186.
57. Welter,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Flicek,P., Manolio,T., Hindorff,L. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.
58. Siepel,A. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
59. Neph,S., Kuehn,M.S., Reynolds,A.P., Haugen,E., Thurman,R.E., Johnson,A.K., Rynes,E., Maurano,M.T., Vierstra,J., Thomas,S. *et al.* (2012) BEDOPS: high-performance genomic feature operations. *Bioinformatics*, **28**, 1919–1920.
60. Pohl,A. and Beato,M. (2014) bwtool: a tool for bigWig files. *Bioinformatics*, **30**, 1618–1619.
61. Berger,M.F. and Bulyk,M.L. (2006) Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods Mol. Biol.*, **338**, 245–260.
62. Xie,D., Boyle,A.P., Wu,L., Zhai,J., Kawli,T. and Snyder,M. (2013) Dynamic trans-Acting factor colocalization in human cells. *Cell*, **155**, 713–724.
63. Boyle,A.P., Araya,C.L., Brdlik,C., Cayting,P., Cheng,C., Cheng,Y., Gardner,K., Hillier,L.W., Janette,J., Jiang,L. *et al.* (2014) Comparative analysis of regulatory information and circuits across distant species. *Nature*, **512**, 453–456.
64. Whitfield,T.W., Wang,J., Collins,P.J., Christopher Partridge,E., Aldred,S., Trinklein,N.D., Myers,R.M. and Weng,Z. (2012) Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.*, **13**, R50.
65. Hnisz,D., Abraham,B.J., Lee,T.I., Lau,A., Saint-André,V., Sigova,A.A., Hoke,H.A. and Young,R.A. (2013) Super-enhancers in the control of cell identity and disease. *Cell*, **155**, 934–947.
66. Wilczyński,B. and Furlong,E.E.M. (2010) Dynamic CRM occupancy reflects a temporal map of developmental progression. *Mol. Syst. Biol.*, **6**, 383.
67. Whyte,W.A., Orlando,D.A., Hnisz,D., Abraham,B.J., Lin,C.Y., Kagey,M.H., Rahl,P.B., Lee,T.I. and Young,R.A. (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**, 307–319.
68. He,Q., Bardet,A.F., Patton,B., Purvis,J., Johnston,J., Paulson,A., Gogol,M., Stark,A. and Zeitlinger,J. (2011) High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. *Nat. Genet.*, **43**, 414–420.
69. Fisher,W.W., Li,J.J., Hammonds,A.S., Brown,J.B., Pfeiffer,B.D., Weiszmann,R., MacArthur,S., Thomas,S., Stamatoyannopoulos,J.A., Eisen,M.B. *et al.* (2012) DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 21330–21335.
70. Longo,D.L. and Drazen,J.M. (2016) Data sharing. *N. Engl. J. Med.*, **374**, 276–277.

A map of direct TF-DNA interactions in the human genome

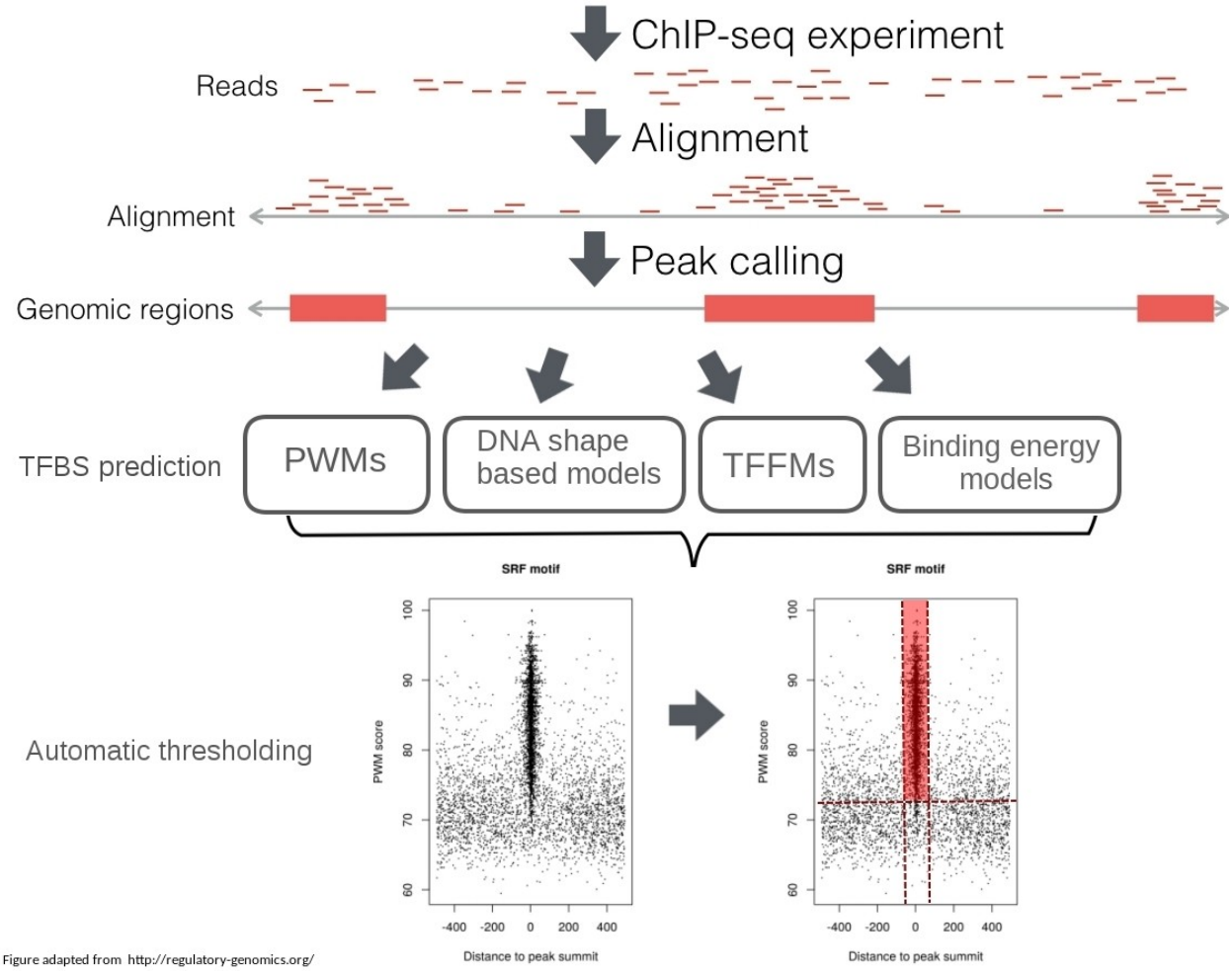
Marius Gheorghe[1], Geir Kjetil Sandve[2], Aziz Khan[1], Jeanne Chèneby[3], Benoit Ballester[3], and Anthony Mathelier[1,4,*]

[1] Centre for Molecular Medicine Norway (NCMM), University of Oslo, Oslo, Norway
[2] Department of Informatics, University of Oslo, Oslo, Norway
[3] Aix Marseille Université, INSERM, TAGC, Marseille, France
[4] Department of Cancer Genetics, Institute for Cancer Research, Radiumhospitalet, Oslo, Norway
* To whom correspondence should be addressed. Email: anthony.mathelier@ncmm.uio.no

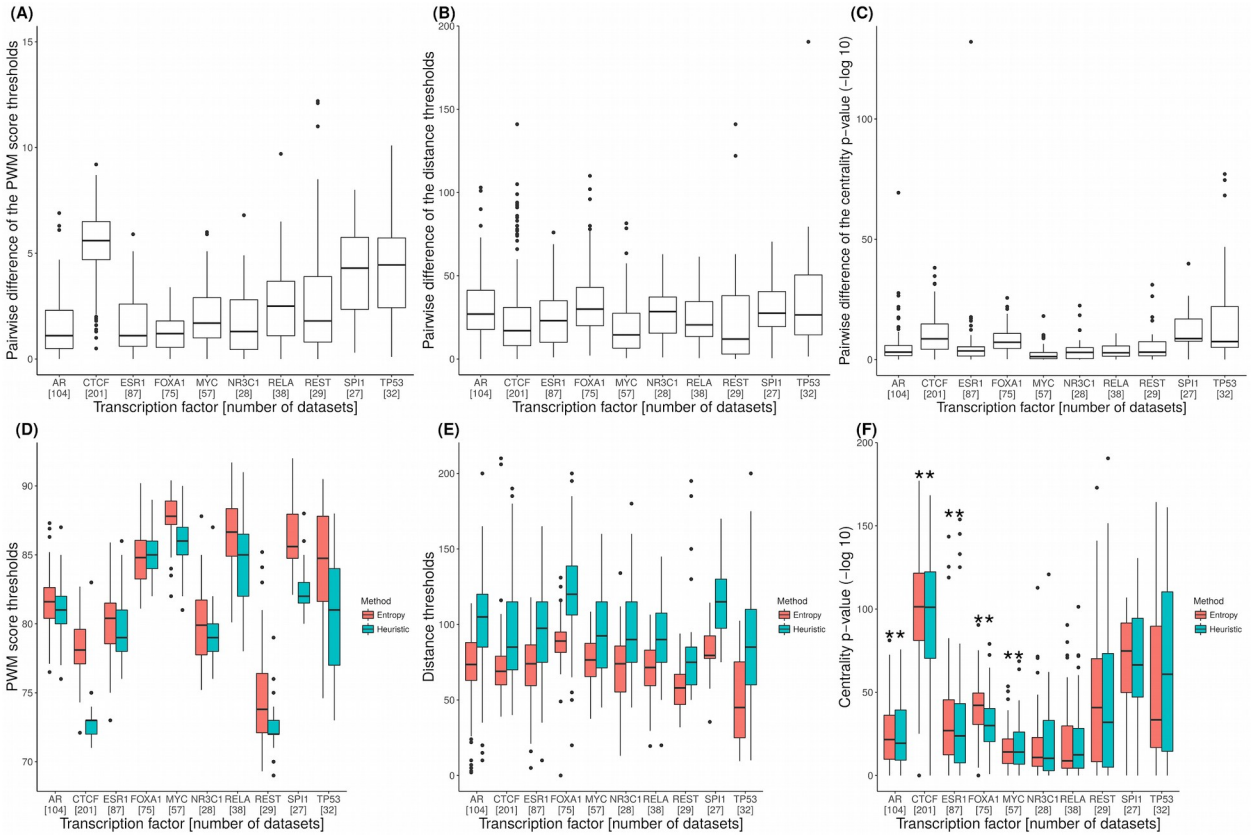


Workflow adapted from: https://github.com/mahmoudibrahim/JAMM/wiki/ChIP-Seq-Alignment-and-Processing-Pipeline

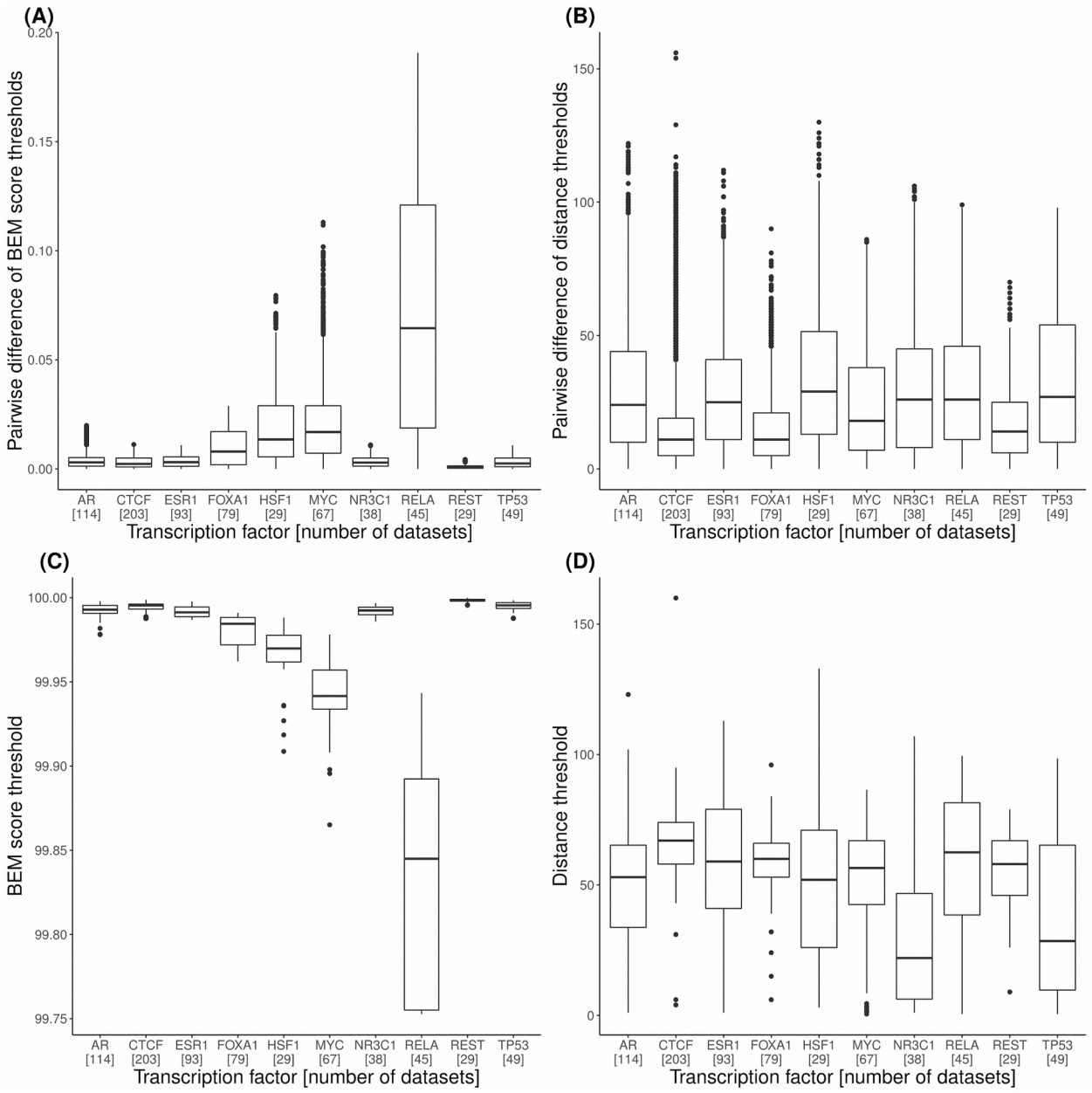
Supplementary Figure S1: The raw read alignment and filtering workflow. This diagram shows the detailed steps followed in the processing of the ChIP-seq data sets, from raw data to ChIP-seq peak calling used for ReMap 2018 (1).



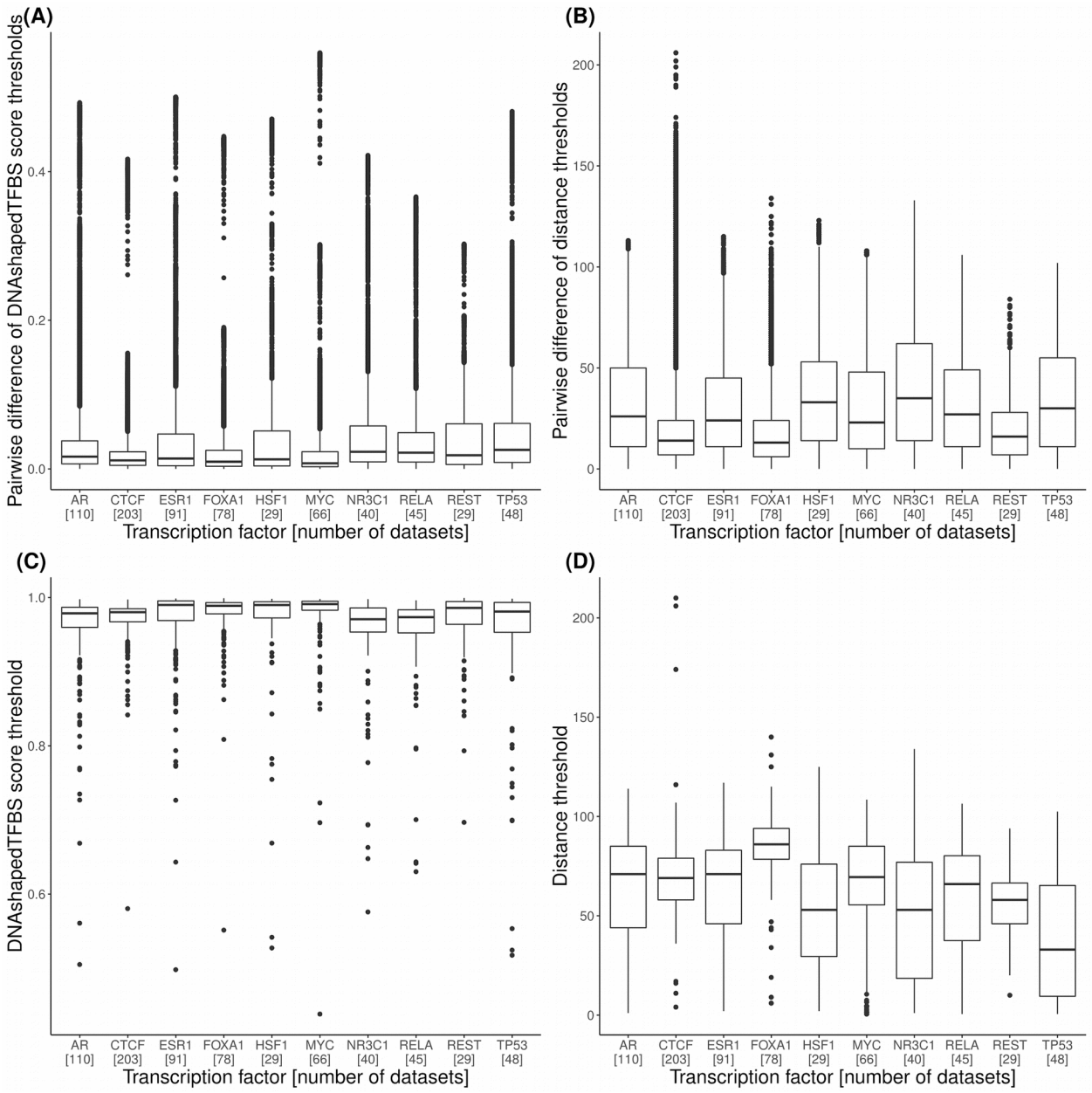
Supplementary Figure S2: A high level diagram of the ChIP-eat processing workflow. Raw reads were aligned to the hg38 version of the human genome and ChIP-seq peaks were called using MACS2 as part of the ReMap 2018 project (1). Subsequently, four TFBS computational models were trained and applied independently on each set of ChIP-seq peaks. Finally, a non-parametric, entropy-based algorithm is used to delineate the enrichment zone to predict direct TF-DNA interactions. The figure was adapted from http://regulatory-genomics.org/.



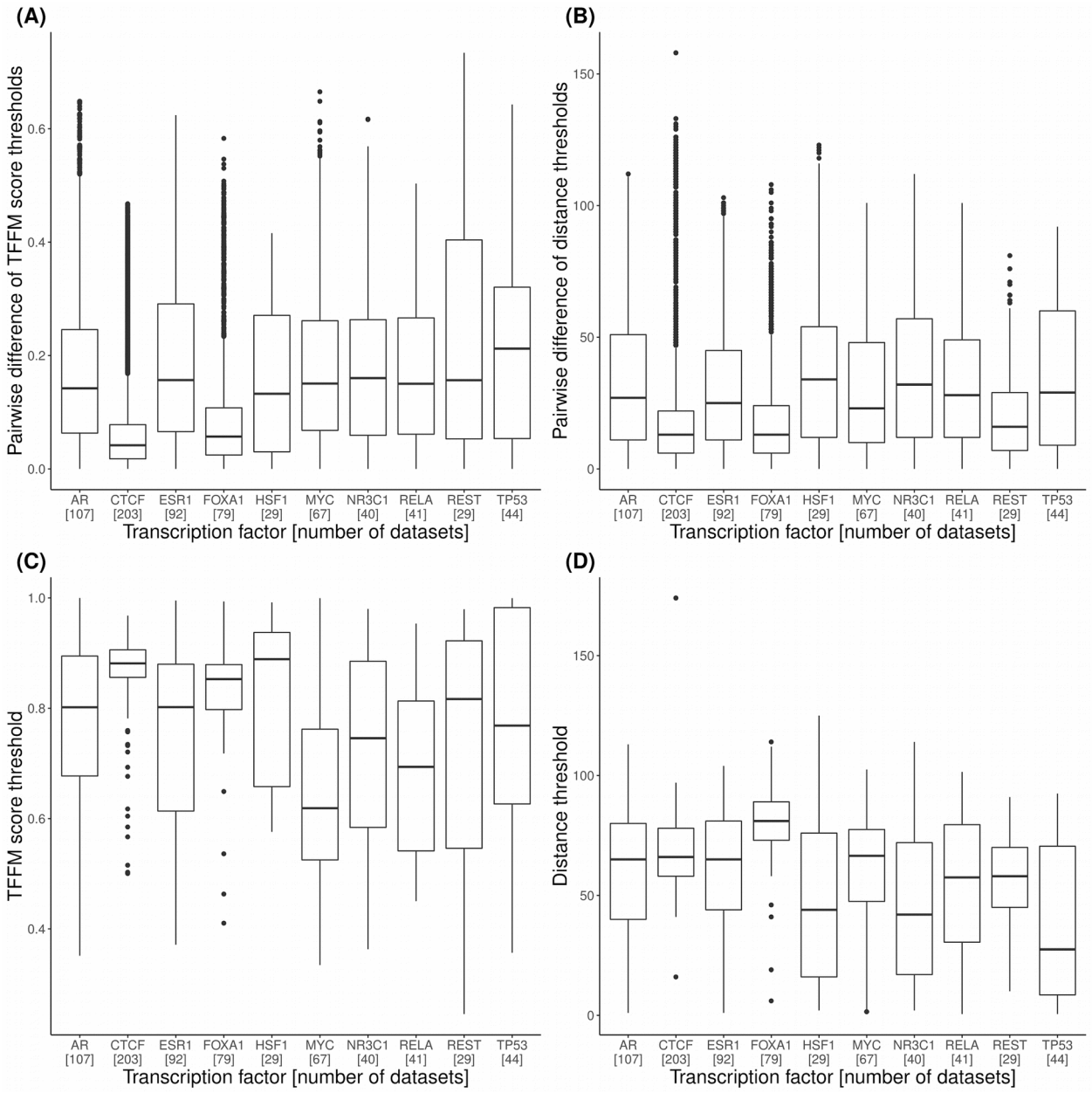
Supplementary Figure S3: Comparison between ChIP-eat and the heuristic approach from (2). For each ChIP-seq peak data sets for the 10 most frequently ChIP'ed TFs, ChIP-eat and the heuristic approach were used to define enrichment zones for TFBSs. Boxplots of the pairwise difference in PWM scores (**A**), distances to peak summits (**B**), and centrality p-values (**C**) are displayed in the upper panels. Boxplots of the absolute values for the predicted thresholds on the PWM scores (**D**), distances to peak summits (**E**), and centrality p-values (-log10 p-values are provided, i.e. the higher the more central the TFBSs are) (**F**) are displayed in the lower panels. Statistically significant difference ($p < 0.05$, Mann-Whitney signed rank test) between the distributions of the centrality p-values are highlighted by '**'. The number of ChIP-seq data sets for each TF is provided between brackets in the x-axis labels. Centrality p-values were computed following the method used in the CentriMo tool (3).



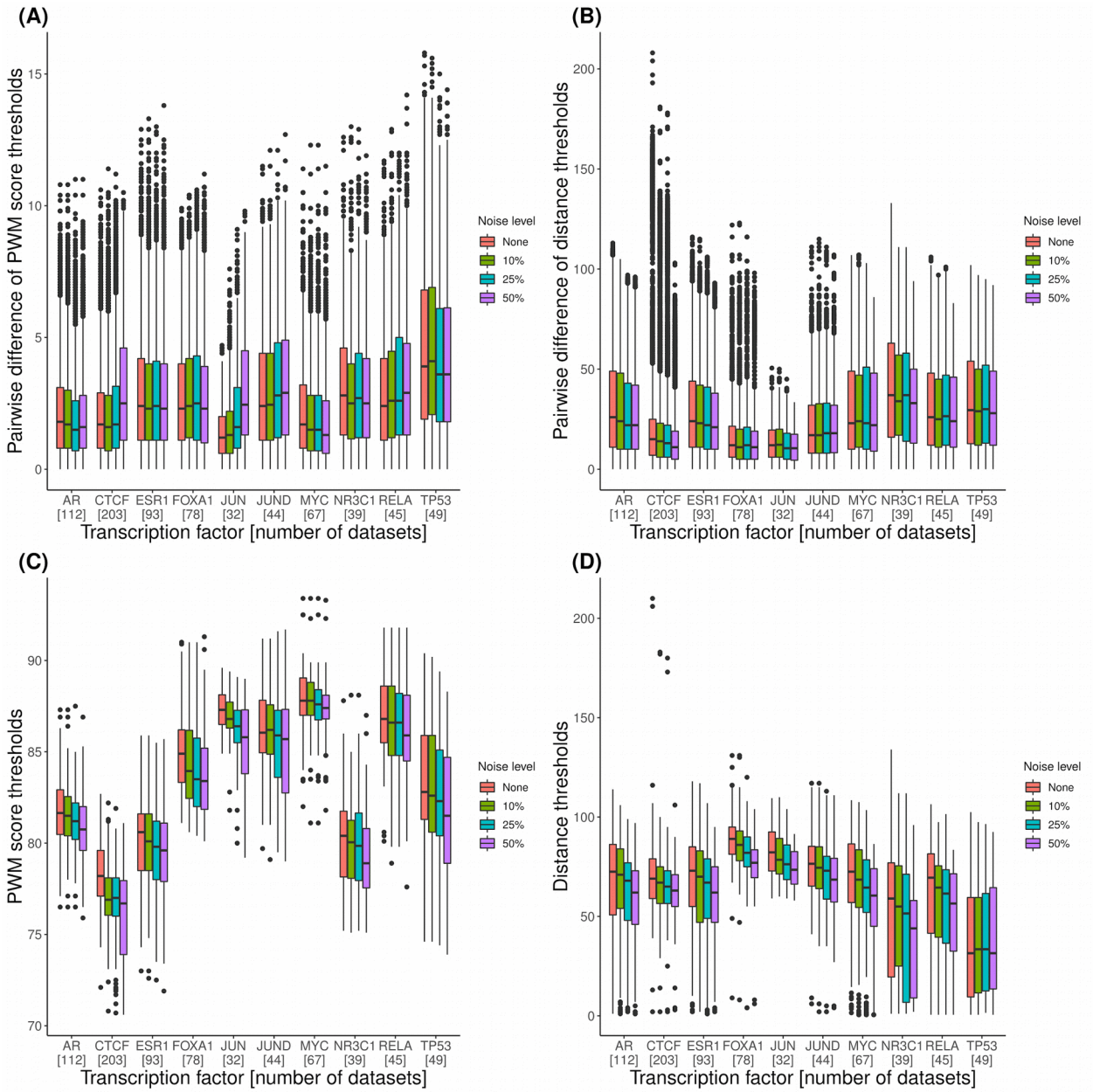
Supplementary Figure S4: Assessment of ChIP-eat enrichment zones using BEM computational models. Boxplots of the pairwise differences for the thresholds on BEM scores (**A**) and distances to peak summits (**B**) between ChIP-seq data sets for the same TF are provided in the upper panels. Boxplot of the absolute threshold values for the BEM scores (**C**) and distances to the peak summits (**D**) are provided in the lower panels. The ten TFs with the highest number of data sets were selected; the number of data sets for each TF is provided between brackets.



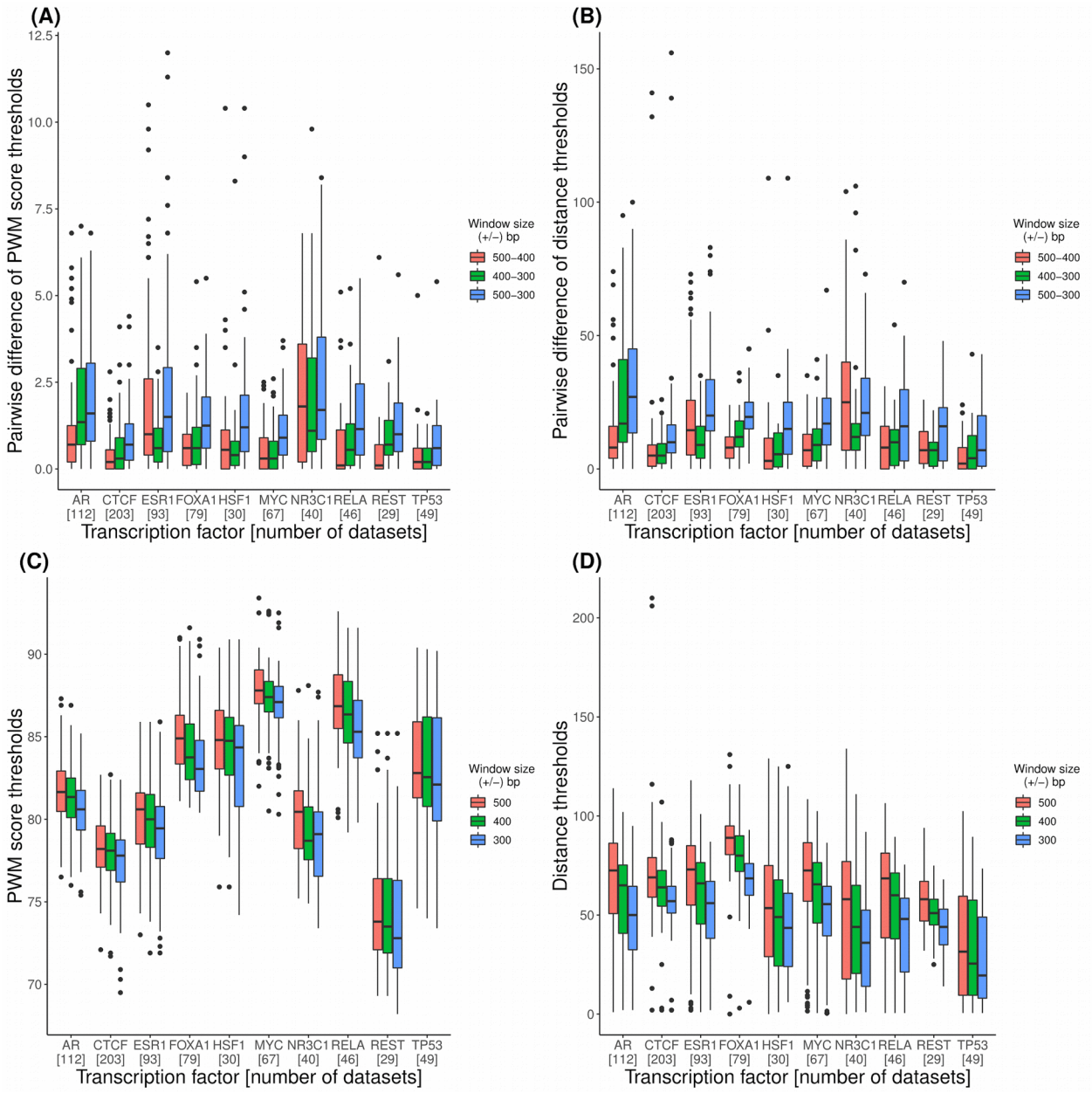
Supplementary Figure S5: Assessment of ChIP-eat enrichment zones using DNAshapedTFBS computational models. Boxplots of the pairwise differences for the thresholds on DNAshapedTFBS scores (**A**) and distances to peak summits (**B**) between ChIP-seq data sets for the same TF are provided in the upper panels. Boxplot of the absolute threshold values for the DNAshapedTFBS scores (**C**) and distances to the peak summits (**D**) are provided in the lower panels. The ten TFs with the highest number of data sets were selected; the number of data sets for each TF is provided between brackets.



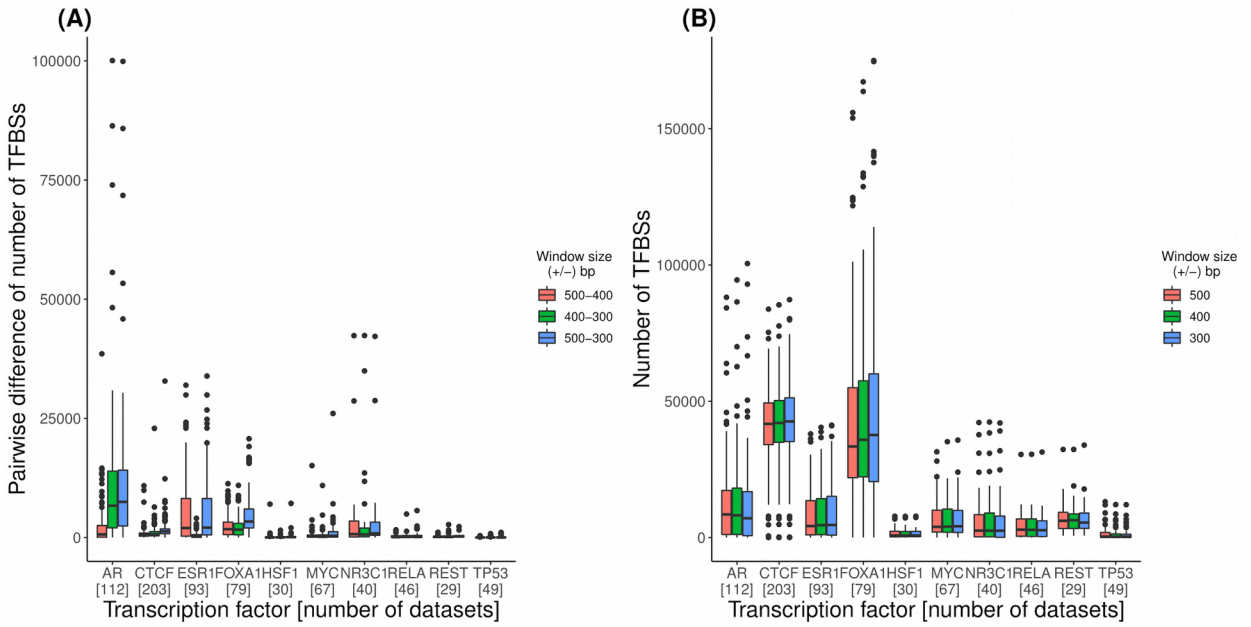
Supplementary Figure S6: Assessment of ChIP-eat enrichment zones using TFFM computational models. Boxplots of the pairwise differences for the thresholds on TFFM scores (**A**) and distances to peak summits (**B**) between ChIP-seq data sets for the same TF are provided in the upper panels. Boxplot of the absolute threshold values for the TFFM scores (**C**) and distances to the peak summits (**D**) are provided in the lower panels. The ten TFs with the highest number of data sets were selected; the number of data sets for each TF is provided between brackets.



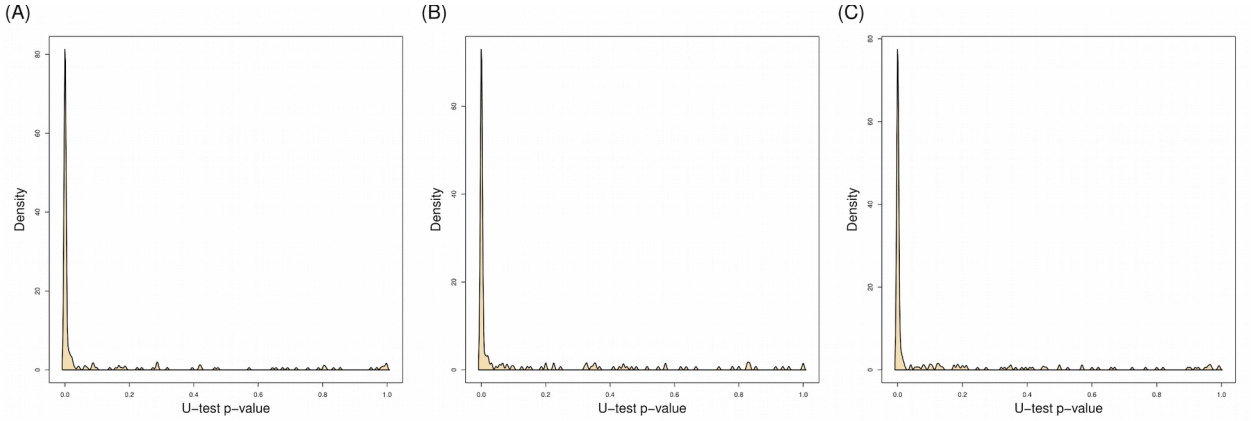
Supplementary Figure S7: Assessment of the robustness to noise for ChIP-eat to define enrichment zones. Boxplots of the pairwise differences for the thresholds on PWM scores (**A**) and distances to peak summits (**B**) between ChIP-seq data sets for the same TF are provided in the upper panels. Boxplot of the absolute threshold values for the PWM scores (**C**) and distances to the peak summits (**D**) are provided in the lower panels. In all panels, results from the original sets of peaks are provided in red. We added noise (random sequences) to the initial set of ChIP-seq peaks for each dataset (see Materials and Methods) to test the variability of the thresholds. Three levels of noise were assessed: 10% (green), 25% (blue), and 50% (violet).



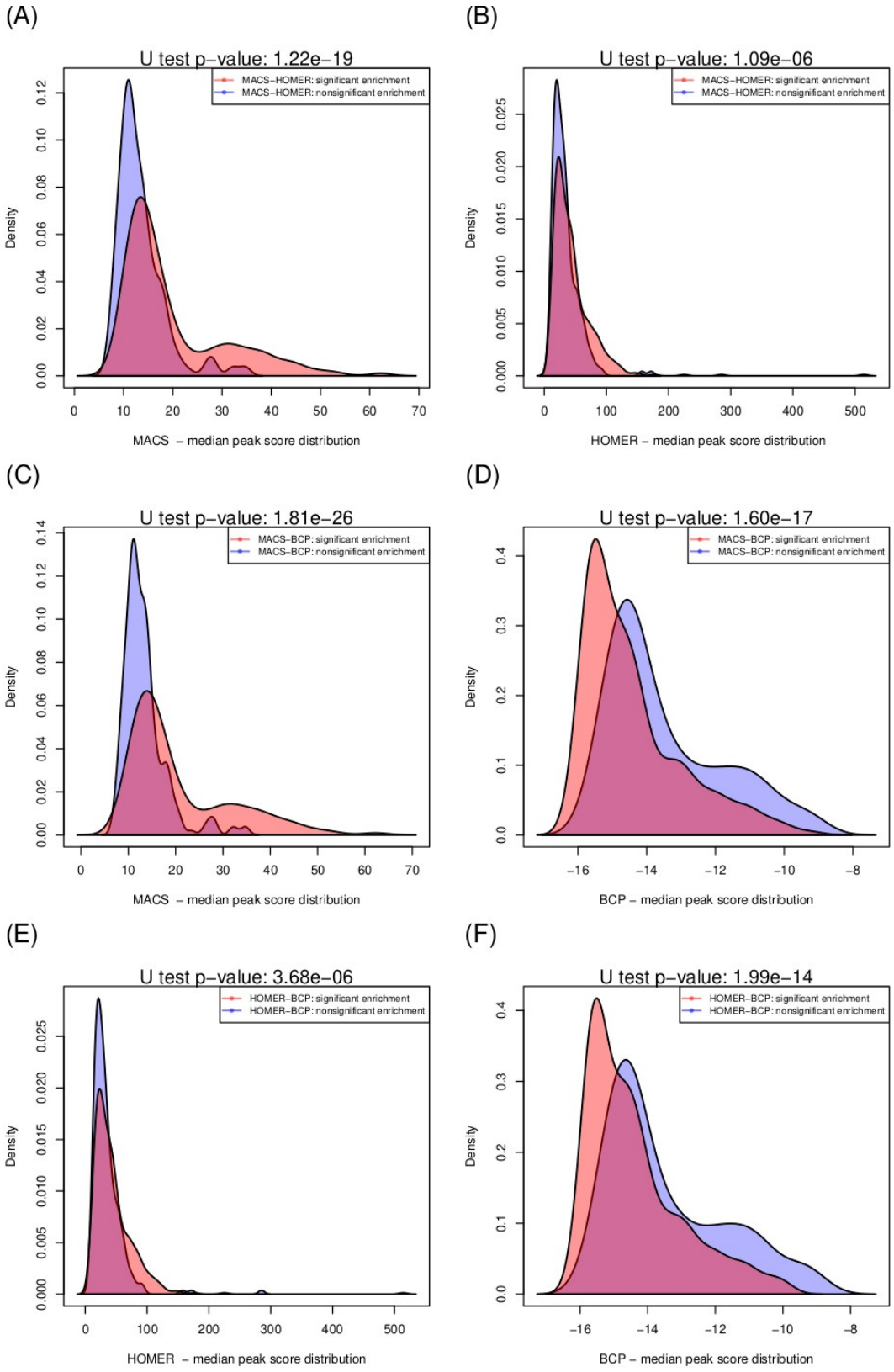
Supplementary Figure S8: Assessment of the robustness to different window sizes around the ChIP-seq peak summit for ChIP-eat to define enrichment zones. Boxplots of the pairwise differences between each two different window sizes for the thresholds on PWM scores (**A**) and distances to peak summits (**B**) between ChIP-seq data sets for the same TF are provided in the upper panels. Boxplot of the absolute threshold values for the PWM scores (**C**) and distances to the peak summits (**D**) are provided in the lower panels. In all panels, results from the original window size (+/- 500 bp) are denoted in red. We have used three different window sizes and scored the sequences for each data set (see Materials and Methods) to test the variability of the thresholds. The different window sizes were: +/-500 bp, +/- 400 bp, and +/- 300 bp around the ChIP-seq peak summit.



Supplementary Figure S9: Assessment of the robustness in the number of predicted TFBSs of the ChIP-eat pipeline when defining the enrichment zone using different window sizes. Boxplots of the pairwise differences in the number of predicted TFBSs between each two different window sizes (**A**) and the absolute number of predicted TFBSs (**B**) for the 10 most frequent TFs. In both panels, the results from the original window size (+/- 500 bp) are denoted in red. We have used three different window sizes, scored the sequences for each data set (see Materials and Methods) and assessed the variability in the number of predicted TFBSs. The different window sizes were: +/-500 bp, +/- 400 bp, and +/- 300 bp around the ChIP-seq peak summit.

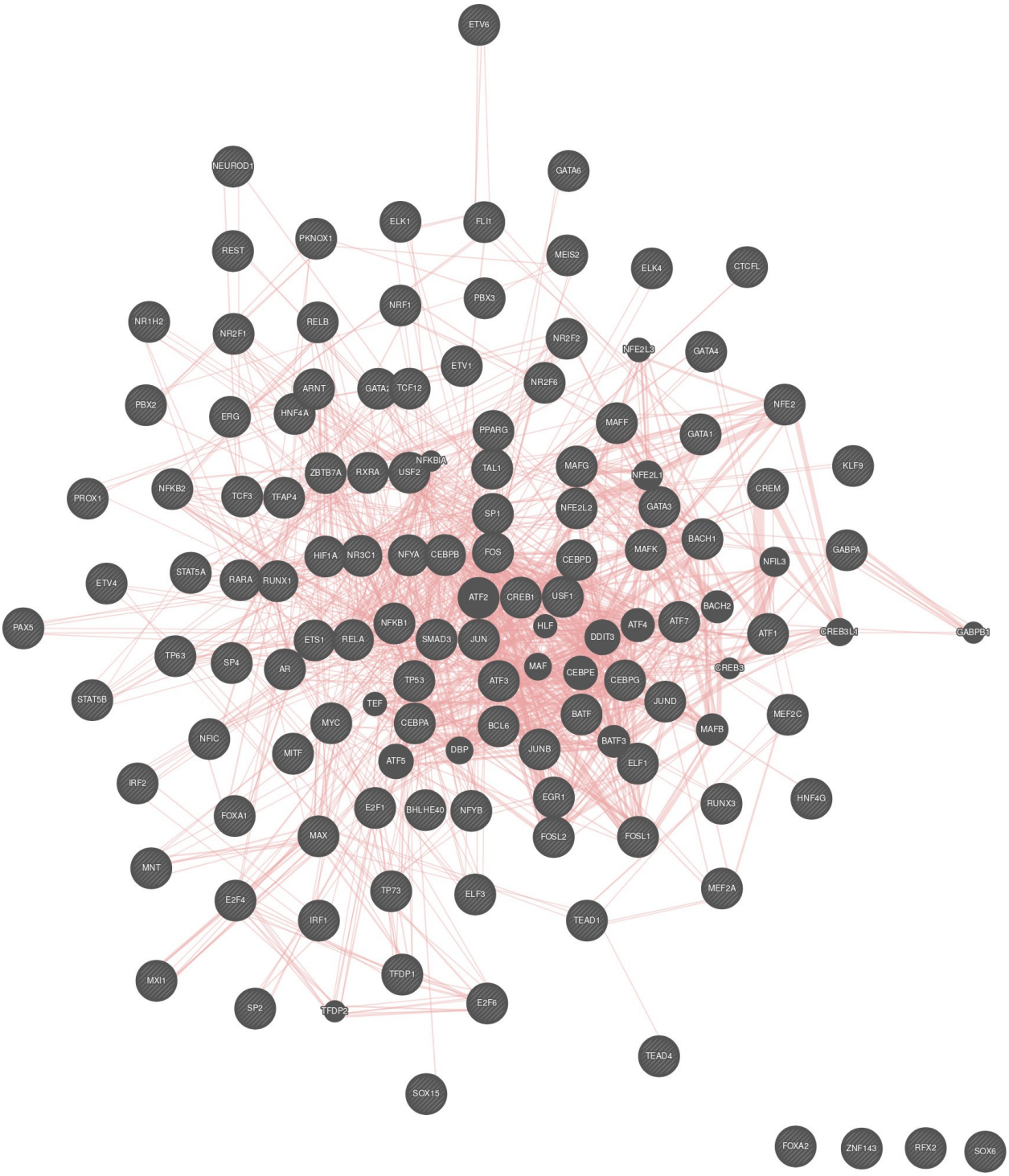


Supplementary Figure S10: Assessment of the binding affinity of the predited direct TF-DNA interactions. The PBM validation analysis (see Materials and methods) was performed for the other three computational models: BEM (**A**), TFFM (**B**), and DNAshapedTFBSs (**C**). The plots show the Mann-Whitney U-test p-values across the distinct distributions of PBM intensity scores between sites within the enrichment zone and outside the enrichment zones.

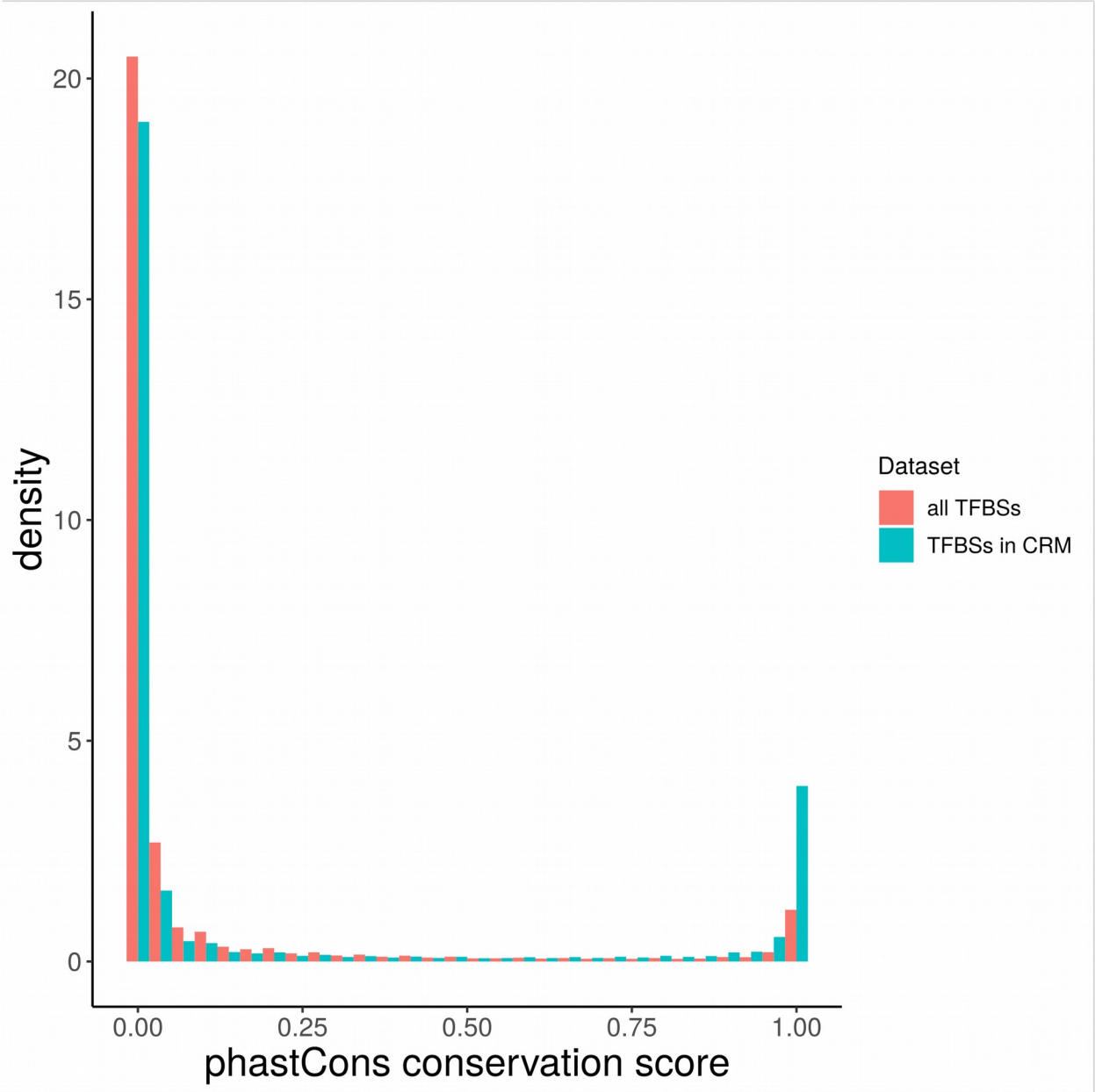


Supplementary Figure S11: Distribution of the ChIP-seq peak caller (MACS2, HOMER, and BCP)

median scores. For each pair of ChIP-seq peak caller, we provide the distribution of the median peak scores assigned by the peak callers to the ChIP-seq peak data sets. For each comparison, we partitioned the data sets in two sets: (i) the ChIP-seq data sets for which there is an enrichment for peaks, from the two peak callers, predicted to contain direct TF-DNA interactions (red), and (ii) the other data sets not presenting a significant enrichment (blue) (see Materials and Methods). Difference between each pair of distributions was assessed using a Mann-Whitney U test. Information about the peak caller used are provided in the legends and x-axis label for each panel.



Supplementary Figure S12: Protein-protein interaction network with the TFs predicted to have co-localized TFBSs across the human genome (see Materials and Methods). TFs are presented as nodes and known protein-protein interactions as edges. This plot was created by providing the list of unique TFs (112) as input to the GeneMANIA tool (4).



Supplementary Figure S13: PhastCons conservation scores at TFBSs. Normalized histograms depicting the densities (y-axis) of the phastCons conservation (5) scores (x-axis) at each bp for all TFBSs predicted by ChIP-eat (red) and the TFBSs within predicted *cis*-regulatory modules (blue).

REFERENCES

1. Chèneby,J., Gheorghe,M., Artufel,M., Mathelier,A. and Ballester,B. (2018) ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.*, **46**, D267–D275.

2. Worsley Hunt,R., Mathelier,A., Del Peso,L. and Wasserman,W.W. (2014) Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. *BMC Genomics*, **15**, 472.

3. Bailey,T.L. and Machanick,P. (2012) Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.*, **40**, e128.

4. Warde-Farley,D., Donaldson,S.L., Comes,O., Zuberi,K., Badrawi,R., Chao,P., Franz,M., Grouios,C., Kazi,F., Lopes,C.T., *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**, W214–20.

5. Siepel,A. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

D. Conclusion

ChIP-eat a été appliqué à 1983 ChIP-seq humains pour 232 TFs différents. Au final, 8 304 135 TFBS ont été identifiés. Ces résultats sont disponibles sur le site Unibind. Sur ces 8 millions, 61 934 ont été identifiés par des clusters de TFBS prédits. Il est intéressant de noter que ces CREs sont considérablement enrichis (valeur P corrigée par le FDR = 2,9e-150) pour les SNP associés à des maladies et à des traits de caractère par traseR.

La fiabilité de la méthode a été vérifiée par des expériences microarray pour 40 TFs. Les séquences d'ADN prédites comme fixant directement les TFs sont similaires à celles des microarrays. Les régions d'enrichissement de fixation directe définies par les seuils des scores des PPM et les distances aux sommets des pics ont ensuite été examinées. Les régions obtenues pour le même TF sont cohérentes entre elles, et ce pour les 10 TFs les plus fréquemment utilisés. En considérant les ensembles de données ChIP-seq les plus fréquemment utilisés, pour les 10 TFs, nous avons observé que les seuils des scores PPM et des distances aux sommets des pics, définissant les zones d'enrichissement, étaient cohérents avec les ensembles de données pour la même TF. Nous observons également que les événements de liaison TF fonctionnels sont susceptibles d'être regroupés et associés à des signaux de ChIP-seq plus forts

La méthode présente certaines limites. A l'origine, cet algorithme a été développé pour différencier le premier plan et l'arrière-plan dans le traitement des images. Il suppose donc la présence de bruit de fond (ou de bruit) dans les données. Cette approche est limitée si l'hypothèse, selon laquelle il y aurait du bruit de fond dans les ensembles de données ChIP-seq analysés, n'est pas confirmée. D'autre part, il s'avère possible que d'autres sites ayant des scores plus faibles présentent des interactions directes entre la TF et l'ADN. Afin de pallier cette limite, il serait intéressant d'étudier les pics des séquences ChIP des événements directs de liaison TF/ADN, que la méthode ne prévoyait pas d'englober. Cela permettrait d'identifier d'autres mécanismes d'interactions protéine/ADN.

On s'attend à ce qu'une partie des TFBS identifiés dans notre étude ne soient pas fonctionnels, comme le suggère le théorème de la futilité[187]. Néanmoins, les événements de liaison TF fonctionnels sont susceptibles d'être regroupés et associés à des signaux de crête ChIP-seq plus forts

Conclusion

I Caractérisation des éléments régulateurs de la transcription à haut débit

Durant ma thèse, mes différents projets ont porté sur l'identification des éléments cis-régulateurs de la transcription (CRE) et, en particulier, sur l'identification des régions de fixation des protéines régulatrices de la transcription. Mes travaux m'ont conduit à développer une amélioration majeure de la base de données ReMap, et dans une certaine mesure de celle de JASPAR. L'objectif du projet ReMap est de cataloguer les éléments régulateurs en se basant sur les événements de fixation des protéines régulatrices de la transcription. Il met à la disposition de la communauté scientifique un catalogue inédit de régions de fixation annoté et retraité uniformément pour l'Homme et chez *A. thaliana.* Les données et outils du projet ReMap sont utilisés par la communauté scientifique (79 citations pour ReMap 2015, 48 citations pour ReMap 2018) dans des projets variés comme l'identification de la fixation de facteurs dans des régions d'intérêt (ex : A repetitive acidic region contributes to the extremely rapid degradation of the cell-context essential protein TRIM52[188], INKA2, a novel p53 target that interacts with the serine/threonine kinase PAK4[189], etc.), ou comme des projets intégratifs nécessitant des données de haute qualité (Inferring Interaction Networks From Multi-Omics Data[190], C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution[191], …). Les données ReMap sont utilisées dans le but de générer les profils de fixation des protéines à l'ADN, dans la base de données JASPAR. Les données de ReMap et de JASPAR ont également été utilisées pour développer ChIP-eat, méthode capable d'identifier les événements de fixation directs des protéines à l'ADN, à partir de données ChIP. Ces données sont disponibles sur la base de données UniBind.

A. Challenge de l'annotation des CREs

ReMap a permis l'identification de 1,7 millions de régions régulatrices (clusters de pics) chez l'Homme et *A. Thaliana* grâce à l'analyse intégrative des données ChIP. Ces éléments régulateurs annotés sont très hétérogènes puisqu'ils comprennent : les promoteurs, les enhancers et les insulateurs pour différentes lignées cellulaires et différents tissus. La méthode ReMap actuelle pour construire les CREs consiste en une intersection des pics présents dans le catalogue. Cette méthode ne permet pas de détecter individuellement chaque élément régulateur, en particulier, dans les régions très denses en pics (Figure I.1).

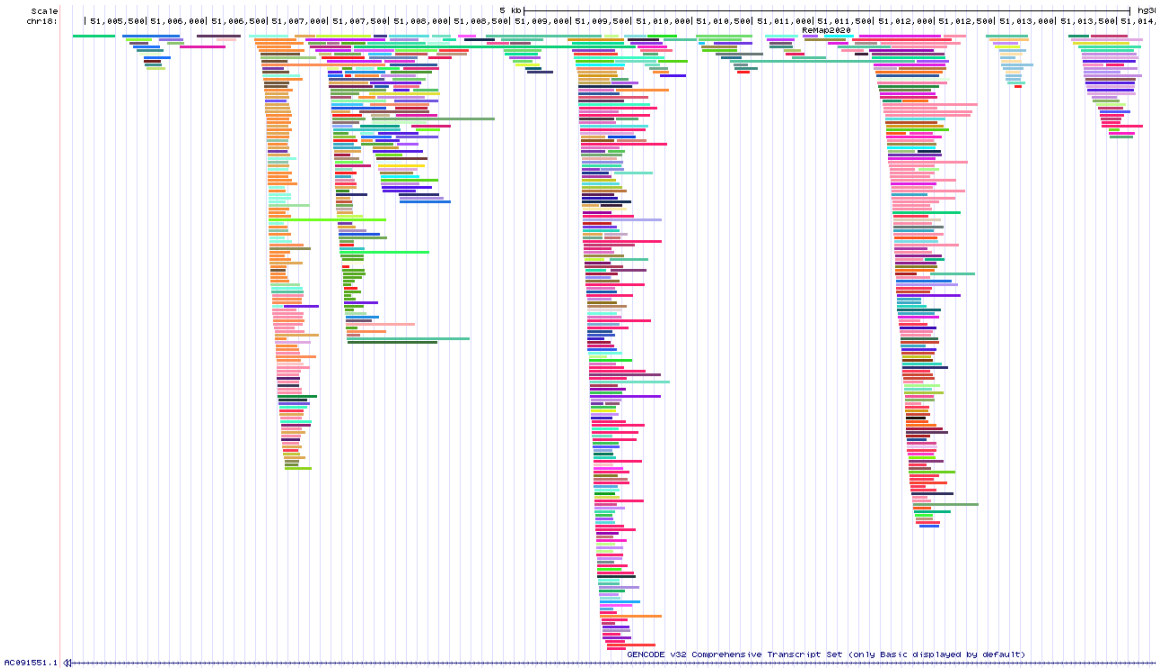


Figure I.1 : Visualisation des pics de fixation des protéines chez l'Homme par le navigateur de génome UCSC.
Au niveau du promoteur de ce gène, une seule région régulatrice représente plusieurs éléments régulateurs.

Les méthodes issues de l'analyse d'expériences biologiques sont soumises aux limites de résolution de ces analyses. Dans le cas du ChIP-seq, l'information de fixation des protéines peut recouvrir des régions de plusieurs centaines de paires de bases (pb) (Résultats I.C. Conclusion ReMap2018).

Une approche consiste à identifier les sites de fixation des TFs (TFBS) situés sous les pics de ChIP-seq (cistrome) afin de définir plus précisément les éléments régulateurs[192,193]. La base de données GTRD propose une ébauche de cistrome pour plusieurs espèces[194]. Une des limites de cette méthode, est due à la qualité des bases de données de TFBS. En effet, le motif de fixation de certaines protéines n'est pas connu ou n'est pas confirmé par des expériences biologiques (ex : liste de profils de fixation putatifs chez JASPAR, voir chapitre concerné). Par conséquent, les CREs formés par de telles protéines ne peuvent pas être détectés lors de la recherche de TFBS.

D'autres approches, proposent l'identification des CREs par l'analyse d'expériences biologiques comme le CAGE, développé par le consortium FANTOM et doté d'une résolution très fine. Cette méthode permet de différencier des promoteurs et des promoteurs alternatifs proches de quelques dizaines de pb. En contrepartie, la méthode CAGE ne permet d'identifier que les régions et les éléments régulateurs transcrits.

Le consortium ENCODE propose une liste de CRE chez l'Homme construite à partir de plusieurs données génomiques[195]. Les CREs ont été identifiés grâce à plusieurs analyses telles que, l'intégration de données d'ouverture de la chromatine (DNase-seq), l'expression des gènes (pucs à ADN), l'analyse des motifs dans les clusters de DNase-seq

et la corrélation entre l'ouverture de la chromatine et la transcription. Comme pour les données issues de FANTOM, ces analyses requièrent de nombreuses expériences biologiques à grande échelle afin d'affiner la résolution de la détection des CRE. Il est donc particulièrement intéressant d'utiliser les nombreuses données biologiques disponibles publiquement via ces consortia et entrepôts de données (GEO, ENA) et de les intégrer dans les analyses bioinformatiques chargées d'identifier les CREs. En revanche, le développement des méthodes bioinformatiques se trouve limité du fait de l'hétérogénéité de composition des CREs, comme, par exemple, l'absence de séquences consensus dans les enhancers, ou la variabilité des éléments dans le promoteur (Introduction I.B. Les éléments régulateurs de la transcription).

L'identification des éléments régulateurs très proches ou se chevauchant peut s'avérer complexe. Deux types de facteurs peuvent limiter cette identification. Il s'agit, d'une part, des difficultés inhérentes aux expérimentations biologiques visant à capturer les interactions protéine/ADN et, d'autre part, de celles liées à l'analyse intégrative de nombreuses données.

1. Importance des analyses intégratives des données ChIP-seq

Actuellement, la technique biologique ChIP-seq est la plus utilisée pour identifier les événements de fixation des protéines à l'ADN. Cette méthode est peu coûteuse, rapide, mais la qualité varie grandement en fonction de la préparation expérimentale des librairies. Les contrôles qualité de ReMap s'avèrent indispensables pour filtrer les expériences. De plus, la précision de la capture des protéines et des analyses dépend fortement du protocole utilisé[196,197]. Le séquençage à faible profondeur ou la taille trop faible des fragments de lecture, le manque de réplicas biologiques ou techniques, ainsi que le single-end limitent considérablement la précision de l'alignement des fragments de lecture sur le génome de référence et, ce, indépendamment de la qualité de l'expérience. Cette hétérogénéité dans les expériences se reflète dans les résultats du *peak-calling*. Une normalisation des données devient nécessaire afin de comparer directement l'intensité du signal entre deux expériences de ChIP dans le cas où les protocoles de ces deux expériences seraient très différents. Ces différences de protocole influent également sur la taille des pics trouvés (Résultats I.C. Conclusion ReMap2018).

La qualité de la capture des interactions protéine-ADN pour les expériences de ChIP-seq est dépendante de la spécificité de l'anticorps utilisé. Il a été démontré que les anticorps capturent, eux aussi, des interactions transitoires avec l'ADN ou avec d'autres protéines[173,198] ce qui conduit à l'identification de régions faussement positives par le ChIP-seq. Au niveau du protocole expérimental, les réplicas biologiques et les contrôles IgG permettent d'identifier ces interactions non significatives. Il est important de contrôler la qualité de cette fixation via des métriques tel que le FRiP (Fraction of Reads in Peaks).

Malgré ces facteurs, les analyses intégratives d'expériences de ChIP-seq sont extrêmement intéressantes. Elles permettent de différencier les informations pertinentes

(régions à forte densité de fixation de protéines régulatrices) du bruit (pics faux positifs dûs à l'utilisation d'un protocole non adapté, à la capture non spécifique des anticorps, etc.). La collecte de données issues de nombreuses expériences ChIP-seq différentes permet d'utiliser la convergence des données pour identifier les CREs Ce n'est pas le cas si on se base sur une seule expérience. Le projet ReMap offre cette opportunité. De plus, des modifications peuvent être proposées afin d'améliorer la qualité du catalogue ReMap.

La liste de CREs proposée par ReMap peut être affinée par les analyses suivantes. Ces analyses se basent sur l'application de méthodes bioinformatiques identifiant les TFBS, sur l'utilisation de données biologiques complémentaires (données sur la conformation de la chromatine) ou plus précises (DAP-seq ChIP-exo), et sur le développement de méthodes d'identification de cluster de pics de fixation de protéines (Résultats I.C.2. L'annotation des clusters de fixation des pics ReMap 2018 et II.F. Projet AtyPeak ReMap 2020).

La présence de TFBS peut être identifiée par la recherche de profils de fixation des TFs. Des outils en ligne tels que matrix-scan de RSAT[125] ou PWMScan[199] permettent l'identification de TFBS à partir de motifs de fixation de TFs de bases de données externes. En revanche, cette méthode ne permet pas de vérifier les protéines régulatrices, comme les cofacteurs qui ne se fixent pas directement à l'ADN. Une autre problématique, propre à ReMap, est le choix de la base de données de profils de fixation des protéines. Une des plus populaires, JASPAR produit ses pics de fixations à partir de données issues de ReMap ce qui engendre un problème circulaire où la qualité de l'un fausse le résultat de l'autre.

L'analyse de données biologiques complémentaires permettrait de confirmer la présence de CREs identifiés par ReMap. L'utilisation de données issues d'expériences biologiques n'utilisant pas les anticorps pour capturer les interactions protéines/ADN permet l'identification de captures non spécifiques. Nous avons fait ce choix et nous utilisons pour ReMap2020 des données issues de DAP-seq. En revanche, cette méthode ne prend pas en compte la conformation de la chromatine. L'utilisation de données DAP-seq doit donc se faire en complémentarité avec l'utilisation de données *in vivo*, telles ChIP-seq ou avec celles de données d'accessibilité de la chromatine telles que l'ATAC-seq.

Les données d'accessibilité de la chromatine peuvent aussi être utilisées pour confirmer les CREs identifiés par lignée cellulaire par ReMap. Des expériences biologiques telles que l'ATAC-seq ou le DNase -seq permettent d'identifier toutes les régions ouvertes le long de l'ADN pour un type cellulaire et à des conditions données. De nombreuses expériences d'ATAC-seq sont disponibles dans les entrepôts de données publiques (le papier décrivant la méthode a été cité plus de 1000 fois[119]. L'utilisation des données brutes d'ATAC-seq, pour les traiter uniformément permettrait d'analyser le recoupement avec les CREs identifiés pour chaque lignée cellulaire. Si plusieurs expériences d'accessibilité du génome existent pour une même lignée cellulaire, seules les régions confirmées par deux expériences ou plus sont conservées. La méthode ATAC-seq étant plus simple à réaliser que le DNase-seq, le nombre de données disponibles issues

d'expériences d'ATAC-seq est largement supérieur à celui des données issues de DNase-seq (366 ATAC-seq contre 29 DNase-seq pour l'Homme[200]).

Une évolution intéressante de ReMap consisterait à utiliser des données issues d'expériences de single-cell. Cette démarche permettrait d'identifier les éléments régulateurs en fonction des populations de cellules et, par conséquent, d'analyser les variations des éléments régulateurs en fonction de la population étudiée. La précision de capture des anticorps complique la réalisation des expériences de ChIP-seq en single-cell[201]. Malgré cela, certaines études commencent à émerger[183,202]. Afin de pallier le manque d'expériences de ChIP-seq single-cell disponibles, à ce jour, des données issues d'autres expériences de fixations protéines/ADN peuvent être intégrées. Je citerais, par exemple, le DamID (DNA adenine methyltransferase identification). Cette expérience permet d'identifier les interactions protéines/ADN grâce à la fusion des protéines d'intérêt et de l'ADN méthyltransférase. Les régions d'interaction sont ensuite identifiées par la méthylation de l'adénine résultant de la fixation de la protéine chimérique à l'ADN. Cette méthode est intéressante car elle a été couplée avec succès au single-cell. L'analyse de ChIP-seq single-cell couplée avec DamID single-cell permettrait d'enrichir la base de données ReMap avec des éléments régulateurs plus spécifiques que ceux issus du séquençage de masse (bulk).

L'identification de CREs peut être améliorée. On observe une très grande variation dans la taille des pics pour une même protéine (Résultats I.C. Conclusions ReMap2018). Certaines protéines de grande taille couvrant une large région de l'ADN ont tendance à produire des pics plus longs (ex : BRD4). Les pics de grande taille sont la cause majeure de la difficulté à délimiter les éléments régulateurs proches. La figure I.1 l'illustre très bien. Un des solutions serait de ne pas utiliser les pics issus de protéines recouvrant de larges régions, au détriment des informations qu'ils apportent. Dans le cas de ReMap, les protéines ayant les pics les plus larges sont : JARID2[203], CHD1[204], DDX5[205], EZH2[206] et SMARCC1[207]. Quatre de ces protéines jouent un rôle dans la conformation de la chromatine et dans le recrutement d'autres protéines régulatrices. Leurs présences, à elles seules, permettent d'identifier des complexes protéiques jouant un rôle dans la régulation de la transcription. Afin de conserver ces larges pics, il est indispensable de développer des méthodes limitant l'impact de la variation de taille des pics si on souhaite détecter les éléments régulateurs plutôt que les régions régulatrices. J'ai fait un premier pas dans cette voie (Résultats I.C. L'annotation des clusters de fixations de pics ReMap2018), mais n'ai, malheureusement, pas eu le temps de tester cette méthode à grande échelle.

2. Analyse Big Data et données biologiques

Les limites évoquées, inhérentes à l'expérimentation mettent en évidence la nécessité d'un contrôle qualité pour chaque expérience. C'est une des raisons pour lesquelles, il est extrêmement intéressant d'intégrer de nombreuses données issues d'expériences similaires afin de se baser, non pas sur une seule expérience, mais sur la redondance de nombreuses expériences indépendantes. Dans l'exemple de ReMap, l'annotation des

éléments régulateurs est basée sur la convergence et le clustering des pics de fixation de protéines à l'ADN ce qui élimine les événements de fixation isolés.

En revanche, le traitement de données Big Data exige certaines précautions. En effet multiplier, les expériences de fixation des protéines à l'ADN multiplie également le nombre d'erreurs, dans les résultats. Il faut tirer parti de la grande redondance des données pour éliminer les observations aberrantes. C'est le cas pour les pics non-redondants de protéines décrits par ReMap. Ces pics sont soutenus par plusieurs dizaines d'événements de fixation issus d'expériences réalisées par des laboratoires différents. Il est, de plus, nécessaire d'effectuer des filtres non plus à l'échelle de l'expérience unique, mais à l'échelle de la base de données dans son ensemble. Prenons l'exemple des analyses d'enrichissement fonctionnel pour lesquelles un test statistique est réalisé pour chaque région d'intérêt. Ce procédé conduit à réaliser plusieurs centaines voire plusieurs milliers de tests, d'où la nécessité de rectifier les résultats à l'aide d'une correction mutlitesting telle que Bonferroni (ex : l'outil d'enrichissement fonctionnel GREAT utilise plusieurs types de corrections multitest[208]). Dans le cas de ReMap, la solution s'avère moins facile au regard de l'hétérogénéité des données, provenant, pour certaines, d'expériences totalement différentes (ex : ChIP-exo, DAP-seq). C'est pourquoi, nous nous sommes tournés vers des analyses a posteriori portant sur la corrélation entre les pics de différents TFs (Résultats II.F. Projet AtyPeak ReMap 2020).

ReMap risque également de faire face, dans le futur, à une limite logicielle. ReMap identifie 1,7M de CREs putatifs. Des logiciels comme Snakemake possèdent des "soft limits" ou limites non strictes au niveau de la construction du schéma interne du pipeline. Plus le nombre d'expériences et d'étapes augmente, plus la construction du schéma interne est longue et coûteuse en ressources. Dans le cas de l'analyse des CREs, après plusieurs tests j'ai estimé que, pour construire son schéma interne sans saturer, Snakemake ne pouvait pas traiter plus de 200 000 analyses avec 4 règles distinctes. La complexité du traitement des données doit donc être évaluée en amont de l'analyse. Une méthode simple à mettre en place pour limiter la complexité du schéma interne, consisterait, par exemple, à isoler les analyses en sub-workflows. Une autre solution serait de transférer les workflows d'analyses de CREs sur un autre gestionnaire de workflow. La construction du schéma interne des pipelines Nextflow est dynamique et supporte des workflows plus complexes.

B. Caractérisation des données produites par ReMap

L'identification des CREs et des régions de fixation des protéines régulatrices est une étape indispensable à une meilleure compréhension des mécanismes de régulation de la transcription. Les analyses à l'échelle du génome d'un organisme exigent un catalogue le plus complet et le plus précis possible. Avec l'avènement de nouvelles méthodes de séquençage et d'analyses à haut débit, de nombreux catalogues voient le jour[195,209,210]. Les catalogues se basant sur la fixation des protéines régulatrices de la transcription permettent de recenser, sans a priori, les régions régulatrices. Il est donc nécessaire

d'annoter et de classer ces régions régulatrices afin d'en identifier les rôles, les mécanismes. Les résultats que nous avons obtenus avec ReMap permettent de nombreuses analyses.

Il serait intéressant de classer et de caractériser les régions identifiées en fonction du type d'éléments régulateurs. Par exemple, la figure I.2 met en évidence la présence d'un cluster important au niveau du promoteur du gène, mais elle permet de discerner également un second cluster représentant sûrement un promoteur alternatif. Il serait aussi pertinent d'identifier, le long du génome humain, une liste de promoteur alternatifs en identifiant les clusters proches des promoteurs dans les régions introniques. L'identification des promoteurs alternatifs est extrêmement importante suite à la découverte de leurs liens avec des maladies[211] telles que la maladie d'Alzheimer[212] ou avec le traitement dans la dystrophie musculaire de Duchenne[213]. Le consortium FANTOM s'est, certes, déjà intéressé à l'identification de ces promoteurs alternatifs, mais la base de données ReMap fournit des informations complémentaires. ReMap fournit des informations concernant plus de 600 lignées cellulaires et tissus dans des conditions différentes alors que FANTOM5 n'en propose que 400. Il serait intéressant de comparer l'enrichissement en protéines de régulation de la transcription entre le promoteur primaire et le promoteur alternatif. Cela permettrait d'identifier différents mécanismes dans l'activation du gène. Cette méthode peut être transposée à d'autre régions géniques spécifiques (ex : pseudogène, régions répétées, etc.) afin d'identifier des profils de fixation propres à ces régions.

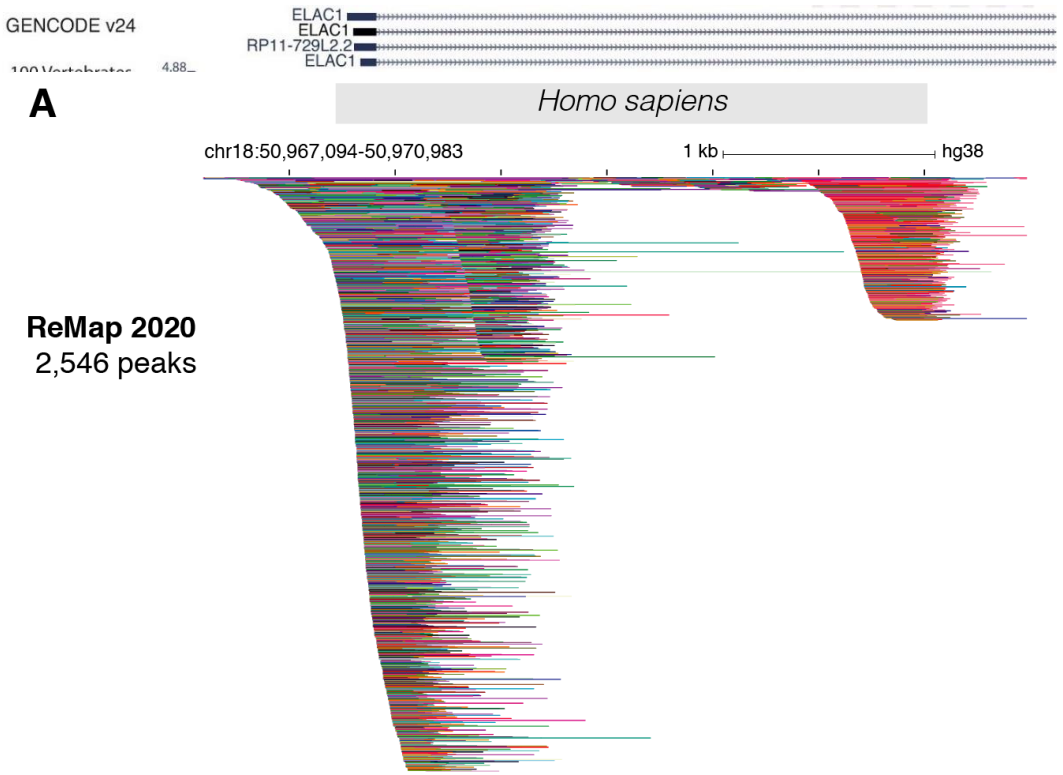


Figure I.2 : Visualisation des données de fixation des protéines régulatrices de ReMap au niveau d'une région promotrice chez l'Homme.

Une autre exploitation possible des données ReMap, est l'identification et la caractérisation des régions insulatrices. Certaines régions du génome présentent une surreprésentation de la fixation du facteur CTCF et de la cohésine RAD21 (Figure I.3). Ces protéines jouent un rôle important dans la conformation de l'ADN (Introduction I.A.4 Coopération entre les protéines régulatrices). Elles délimitent les TADs et permettent d'identifier les insulateurs. Dans le cas des données ReMap, ces pics proviennent d'expériences réalisées sur différentes lignées cellulaires ou tissus. Les données de fixations de CTCF et RAD21 issues de ReMap permettraient donc d'établir un catalogue d'insulateurs putatifs pour les différentes lignées et pour les différents tissus cellulaires. Caractériser ces insulateurs contribuera à définir les limites hypothétiques des TADs sans qu'il soit nécessaire d'utiliser des expériences de capture de la conformation comme le Hi-C. Identifier les insulateurs spécifiques aux lignées cellulaires aiderait à définir la probabilité d'interaction entre les éléments régulateurs comme les enhancers et les promoteurs[214,215].

L'étude des associations entre protéines régulatrices permettrait de mieux comprendre la hiérarchie entre les différentes protéines. L'observation a montré que certaines protéines jouent le rôle de "Master Regulator"[136]. Ces protéines ont des fonctions permettant de faciliter le recrutement de nombreuses protéines au niveau des éléments régulateurs ou d'initier la transcription de gènes produisant des protéines régulatrices[216,217]. Ces master regulator sont des cibles privilégiées lors de l'étude des pathologies[218,219].

Au cours de ma thèse, je me suis également intéressée aux enhancers transcrits, un des sous-groupe d'enhancers produisant, de façon bidirectionnelle, des ARN courts non polyadénylés (eRNA). Différents travaux ont montré que ces eRNA possédaient des fonctions capables de stimuler la transcription au niveau de certains gènes[71–73] .Il peut s'avérer complexe d'identifier des enhancers transcrits dans la mesure où les eRNA sont rapidement dégradés et sont transcrits à des niveaux faibles. Grâce à la méthode CAGE, le consortium FANTOM5 a identifié de nombreuses régions transcrites de façon bidirectionnelles, pour de nombreux types cellulaires chez l'Homme et chez la souris[220]. Cette approche nécessite de gros moyens financiers et une technique particulière. Nous avons fait le choix, d'employer des analyses intégratives pour identifier ces enhancers transcrits. Nous savons que les enhancers sont transcrits par l'ARN polymérase II (Pol II). Le traitement uniforme des données de ChIP-seq de Pol II issues de GEO et d'ENCODE permet l'identification des régions de fixations de Pol II. L'annotation des CREs de ReMap intersectant les régions de fixation de Pol II pourrait permettre d'identifier les régions régulatrices potentiellement transcrites. Il est envisageable d'aller encore plus loin en analysant la forme du signal de fixation de la Pol II au niveau des enhancers transcrits. Plusieurs méthodes permettent de vérifier cette hypothèse. Dans un premier temps, un contrôle positif consisterait à comparer la liste d'enhancers putatifs à celle de FANTOM5. Dans un deuxième temps, il conviendrait de vérifier que la fixation de l'a Pol II se fait bien au niveau de l'enhancer et non pas du gène et, pour ce faire, étudier la fixation des facteurs généraux tels que TFIIB dont ReMap nous fournit les données.

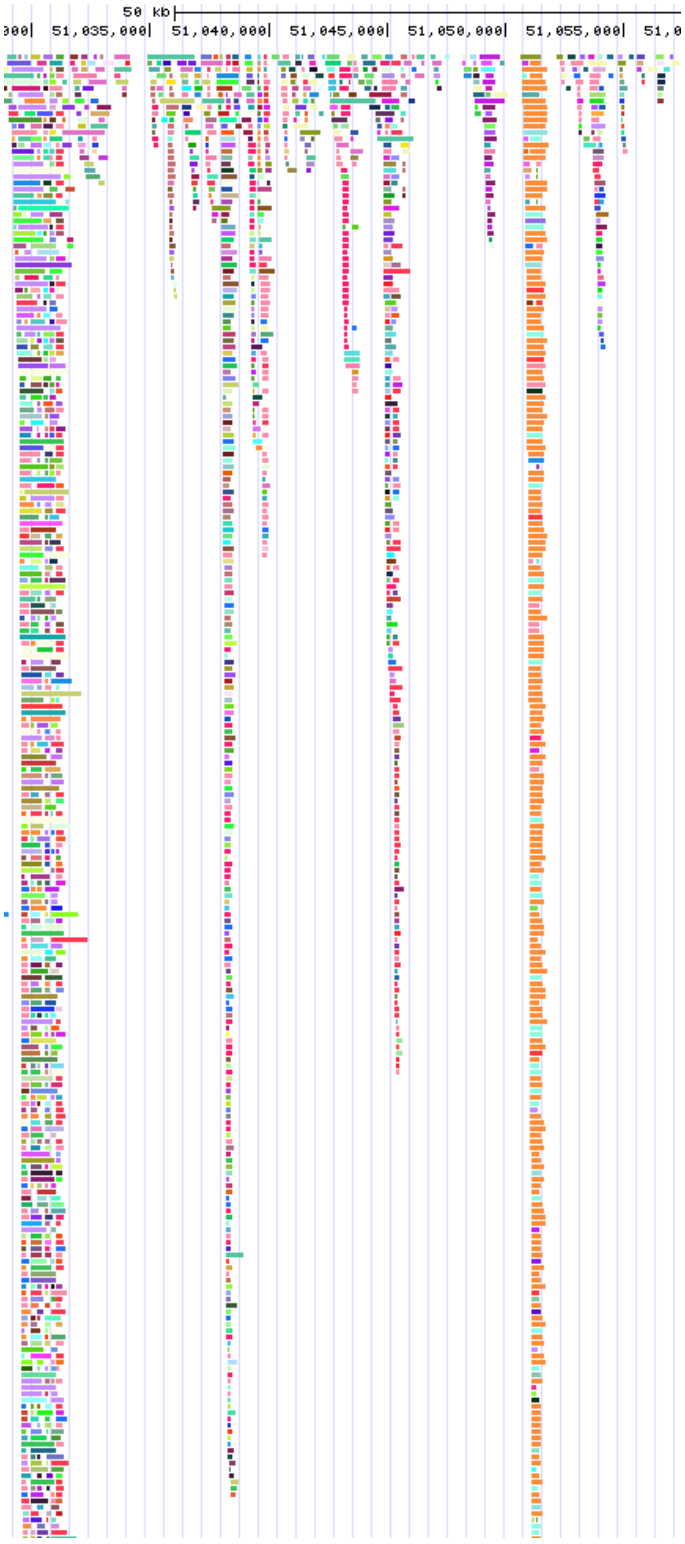


Figure I.3 : Visualisation d'un cluster de CTCF et RAD21 dans les données ReMap.
Orange : Redondance de 380 pics de CTCF provenant de plus de 100 lignées cellulaires différentes. Bleu clair : Redondance de 78 pics de Rad21 provenant de 47 lignées cellulaires. Ces deux protéines représentent plus de 80% de ce cluster.

II Reproductibilité dans la recherche

Le partage des données est un des principes fondamentaux de la recherche. Pour répondre à cet impératif, les analyses et données produites doivent être compréhensibles et reproductibles. C'est d'autant plus vrai dans le cadre des études intégratives à grande échelle pour lesquelles l'annotation et le traitement uniforme des données sont indispensables. Lors de de ma thèse, j'ai veillé à appliquer ces concepts au projet d'intégration de données ReMap.

A. Reproductibilité du traitement et des analyses

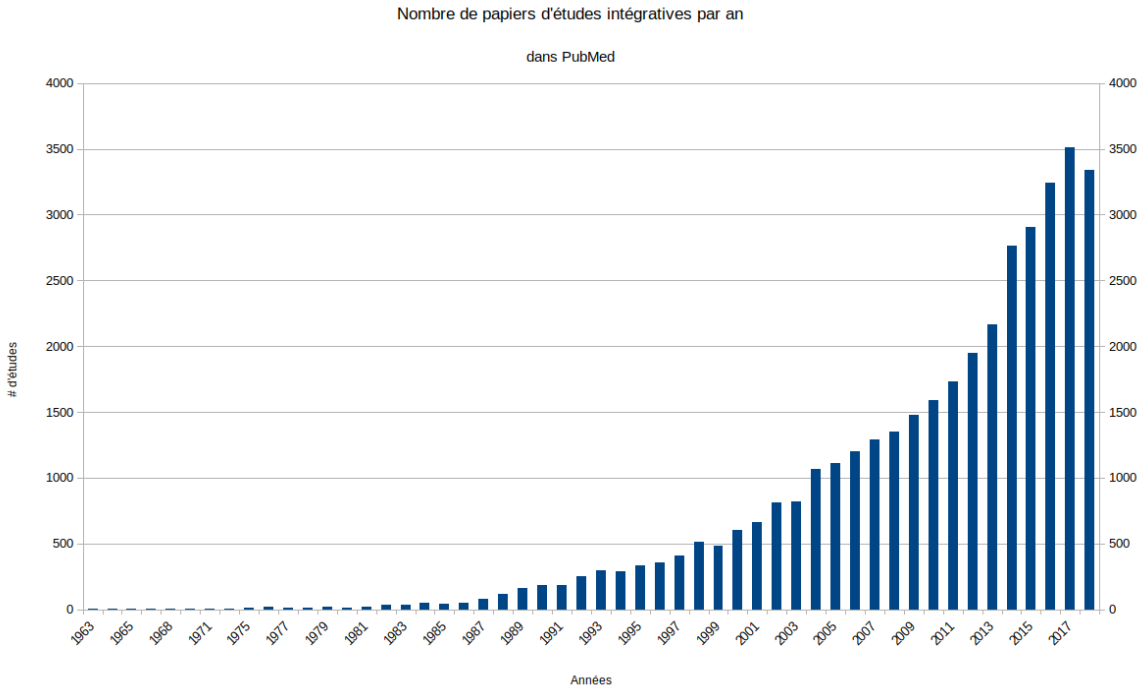


Figure II.1 : Nombre d'articles scientifiques, recensés dans GEO, contenant les mots "integrative study" dans le titre, l'abstract ou les mots clés.

Les études intégratives sont de plus en plus courantes en biologie (Figure II.1). Elles nécessitent des moyen informatiques et statistiques de plus en plus complexes. Depuis quelques années, la communauté scientifique s'est concentrée sur l'amélioration des outils et des méthodes afin de faciliter les études intégratives. De nombreux outils sont utilisés pour aider au développement et au déploiement de workflows tels que Snakemake (cité 449 fois depuis sa parution en 2012), Nextflow (cité 141 fois depuis 2017) et Bioconda (cité 49 fois depuis sa parution en 2018). L'avantage de ces outils est d'être ouverts et libres de droit et, aussi, de pouvoir être déployés au sein de structure locales.

a : https://elixir-europe.org
b : www.france-bioinformatique.fr

Le projet Elixir[a] a pour but de développer et de rendre accessibles les ressources bioinformatiques à la communauté scientifique Européenne. L'institut Français de Bioinformatique[b] (IFB) fait partie d'Elixir et propose un accès libre et gratuit à différentes plateformes de stockage et de calculs de données scientifiques. Ces solutions sont à mettre en opposition aux services commerciaux proposés par Amazon, Google ou Microsoft. Ces services (souvent des clouds) posent question quant à la traçabilité de la localisation des données et quant à la sécurité. Quelle que soit la méthode utilisée, il devient nécessaire d'appliquer les mêmes standards de partage aux protocoles biologiques et aux analyses informatiques.

B. FAIRification

La reproductibilité des données passe aussi par une meilleure annotation des données et métadonnées produites. Les principes FAIR ont mis à disposition depuis 2016 des lignes de conduite à suivre et à discuter pour aider à pérenniser les données. L'Institut Français de Bioinformatique essaye d'améliorer l'interopérabilité des bases de données et des outils produits par les chercheurs français. Ce travail est indispensable pour profiter pleinement des ressources produites par les différentes équipes de chercheurs. Cet effort appliqué au recensement des outils et des bases de données rend les travaux divulgués, accessibles et réutilisables par le plus grand nombre.

La FAIRification des données constitue la prochaine étape des études intégratives. Les avancées technologiques facilitent considérablement la production de données à grande échelle. La reconnaissance de l'intérêt de la bioinformatique rend la production d'outils de plus en plus aisée. Nous arrivons donc à un moment clé, où la limite dans la recherche en bioinformatique n'est plus seulement technique, mais devient aussi structurelle. Pour exemple, le traitement informatique (production) des données de ReMap 2020 a demandé environ un mois de travail et, ce, sans compter le développement des scripts. La réflexion, en amont, sur la structure des données et des métadonnées, quant à elle, a nécessité plusieurs mois. Le travail que j'ai effectué sur ReMap et en collaboration avec l'équipe de JASPAR m'a fait prendre conscience de l'importance de la mise en place d'un cahier des charges simple et applicable à tous les types d'analyses bioinformatiques. Les groupes de réflexions, comme FAIR, sont indispensables pour guider les chercheurs vers un meilleur partage d'information.

C. Projet UniMap

La FAIRification des données est un processus long et difficile. Pour les entrepôts de données biologiques tels que GEO, le manque de directives concernant le dépôt de données complique l'utilisation des données à grande échelle. Pour ce qui concerne le projet ReMap, nous annotons manuellement les données de fixation des protéines à l'ADN afin de pouvoir les traiter uniformément. Nous avons pris la décision de créer un outil web pour faciliter ce processus d'annotation manuelle des données.

L'objectif du projet UniMap est de regrouper les outils et les connaissances nécessaires à l'annotation des données issues d'une plateforme web. Ce projet concerne principalement les données issues de GEO. En effet, des difficultés émergent lors de l'annotation de ces données, comme l'absence de métadatas uniformément annotées ou l'absence d'association entre fichiers d'une même expérience. Les données figurant sur GEO sont regroupées par papier. Par conséquent, les réplicas de différentes expériences se trouvent mélangés. Il faut donc, pour chaque expérience, répertorier l'ensemble des fichiers la concernant. Il faut, ensuite, distinguer les fichiers de contrôle des fichiers de l'expérience et, aussi, identifier les réplicas biologiques et les réplicas techniques.

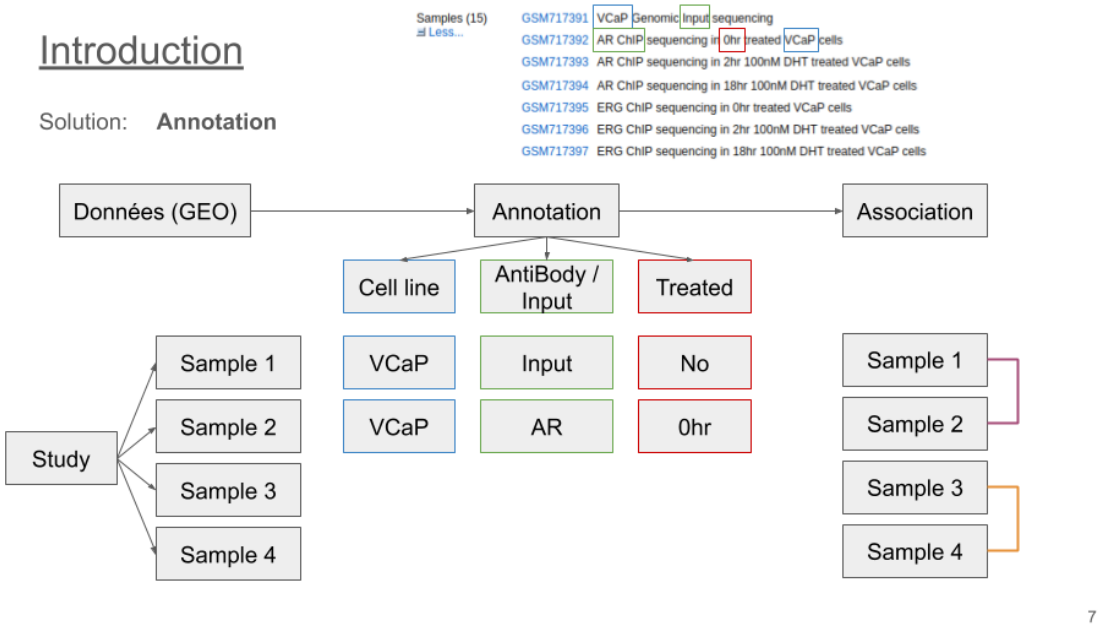


Figure II.2 : Pipeline d'annotation d'UniMap.
Les métadonnées de ChIP-seq concernant les fichiers stockés sur GEO sont directement récupérés. Les métadonnées sont regroupées par étude (study). Ces métadonnées sont ensuite annotées par un utilisateur humain. Pour chaque fichier il faut identifier si c'est : un contrôle (input), si non, quelle est la protéine cible, la lignée cellulaire concernée et son traitement. La troisième étape consiste à associer les fichiers. Il faut identifier les réplicas et les contrôles pour chaque expérience.

Figure II.2 : Thomas Rosnet

L'objectif est de regrouper en une seule page :

- Les métadatas pertinentes de GEO : listes des fichiers, types d'expériences (ex : ChIP-seq paired-end), protéine cible ou anticorps utilisés pour les ChIPs, type cellulaire ou tissu utilisé, lien vers protocole/article, organisme.
- Les ontologies officielles : HGNC, BRENDA et Cellosaurus.
- Les noms utilisés par la base de données ReMap.

Une fois l'annotation effectuée, UniMap doit être en mesure de fournir toutes les informations nécessaires au traitement automatique des données par le pipeline ReMap (Figure II.2).

Bibliographie

1. Cramer, P. *et al.* Structure of Eukaryotic RNA Polymerases. *Annu. Rev. Biophys.* **37**, 337–352 (2008).
2. Dieci, G., Fiorino, G., Castelnuovo, M., Teichmann, M. & Pagano, A. The expanding RNA polymerase III transcriptome. *Trends Genet.* **23**, 614–622 (2007).
3. Kang, J. *et al.* LRF acts as an activator and repressor of the human β-like globin gene transcription in a developmental stage dependent manner. *Biochem. Cell Biol.* **97**, 380–386 (2019).
4. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10**, 252–263 (2009).
5. Nguyen, N. T. T. *et al.* RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Res.* **46**, W209–W214 (2018).
6. Gonze, D., Pinloche, S., Gascuel, O. & van Helden, J. Discrimination of yeast genes involved in methionine and phosphate metabolism on the basis of upstream motifs. *Bioinformatics* **21**, 3490–3500 (2005).
7. Kamanu, F. K. *et al.* Mutations and Binding Sites of Human Transcription Factors. *Front. Genet.* **3**, 100 (2012).
8. Fridley, B. L. & Biernacka, J. M. Gene set analysis of SNP data: benefits, challenges, and future directions. *Eur. J. Hum. Genet.* **19**, 837–843 (2011).
9. *Campbell, Reece, Mitchell & Taylor, Biology: Concepts and Connections | Pearson*.
10. *McGraw-Hill encyclopedia of science & technology : an international reference work in twenty volumes including an index.* (McGraw-Hill, 2002).
11. Olins, D. E. & Olins, A. L. Chromatin history: our view from the bridge. *Nat. Rev. Mol. Cell Biol.* **4**, 809–814 (2003).
12. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362–366 (2009).
13. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).
14. Turner, B. M. Cellular memory and the histone code. *Cell* **111**, 285–91 (2002).
15. Rosenfeld, J. A. *et al.* Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics* **10**, 143 (2009).
16. King, H. W. & Klose, R. J. The pioneer factor OCT4 requires the chromatin remodeller BRG1 to support gene regulatory element function in mouse embryonic stem cells. *Elife* **6**, (2017).
17. Iwafuchi-Doi, M. *et al.* The Pioneer Transcription Factor FoxA Maintains an

Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. *Mol. Cell* **62**, 79–91 (2016).

18. Grebbin, B. M. & Schulte, D. PBX1 as Pioneer Factor: A Case Still Open. *Front. Cell Dev. Biol.* **5**, 9 (2017).
19. Wang, H. *et al.* Purification and functional characterization of a histone H3-lysine 4-specific methyltransferase. *Mol. Cell* **8**, 1207–17 (2001).
20. Nakamura, T. *et al.* ALL-1 is a histone methyltransferase that assembles a supercomplex of proteins involved in transcriptional regulation. *Mol. Cell* **10**, 1119–28 (2002).
21. Rea, S. *et al.* Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* **406**, 593–599 (2000).
22. Mizzen, C. A. *et al.* The TAF(II)250 subunit of TFIID has histone acetyltransferase activity. *Cell* **87**, 1261–70 (1996).
23. Schiltz, R. L. *et al.* Overlapping but distinct patterns of histone acetylation by the human coactivators p300 and PCAF within nucleosomal substrates. *J. Biol. Chem.* **274**, 1189–92 (1999).
24. Zhang, B. *et al.* A dynamic H3K27ac signature identifies VEGFA-stimulated endothelial enhancers and requires EP300 activity. *Genome Res.* **23**, 917–27 (2013).
25. Kuzmichev, A., Jenuwein, T., Tempst, P. & Reinberg, D. Different EZH2-containing complexes target methylation of histone H1 or nucleosomal histone H3. *Mol. Cell* **14**, 183–93 (2004).
26. Smale, S. T. & Kadonaga, J. T. The RNA Polymerase II Core Promoter. *Annu. Rev. Biochem.* **72**, 449–479 (2003).
27. Yang, C., Bolotin, E., Jiang, T., Sladek, F. M. & Martinez, E. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene* **389**, 52–65 (2007).
28. Deng, W. & Roberts, S. G. E. A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes Dev.* **19**, 2418–23 (2005).
29. Jabbari, K. & Bernardi, G. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene* **333**, 143–149 (2004).
30. Gardiner-Garden, M. & Frommer, M. CpG Islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).
31. Illingworth, R. S. *et al.* Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome. *PLoS Genet.* **6**, e1001134 (2010).
32. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–22 (2011).
33. Hu, S. *et al.* DNA methylation presents distinct binding sites for human transcription factors. *Elife* **2**, (2013).
34. Domcke, S. *et al.* Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**, 575–579 (2015).
35. Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of

promoter DNA methylation in the human genome. *Nat. Genet.* **39**, 457–466 (2007).

36. Schübeler, D. Function and information content of DNA methylation. *Nature* **517**, 321–326 (2015).

37. Liu, X., Kraus, W. L. & Bai, X. Ready, pause, go: regulation of RNA polymerase II pausing and release by cellular signaling pathways. *Trends Biochem. Sci.* **40**, 516–525 (2015).

38. Brookes, E. & Pombo, A. Modifications of RNA polymerase II are pivotal in regulating gene expression states. *EMBO Rep.* **10**, 1213–1219 (2009).

39. Meinhart, A. & Cramer, P. Recognition of RNA polymerase II carboxy-terminal domain by 3′-RNA-processing factors. *Nature* **430**, 223–226 (2004).

40. Spilianakis, C. G., Lalioti, M. D., Town, T., Lee, G. R. & Flavell, R. A. Interchromosomal associations between alternatively expressed loci. *Nature* **435**, 637–645 (2005).

41. Biggar, S. R. & Crabtree, G. R. Cell signaling can direct either binary or graded transcriptional responses. *EMBO J.* **20**, 3167–3176 (2001).

42. Giorgetti, L. *et al.* Noncooperative interactions between transcription factors and clustered DNA binding sites enable graded transcriptional responses to environmental inputs. *Mol. Cell* **37**, 418–28 (2010).

43. Zinzen, R. P., Senger, K., Levine, M. & Papatsenko, D. Computational models for neurogenic gene expression in the Drosophila embryo. *Curr. Biol.* **16**, 1358–65 (2006).

44. Chopra, V. S. & Levine, M. Combinatorial patterning mechanisms in the Drosophila embryo. *Briefings Funct. Genomics Proteomics* **8**, 243–249 (2009).

45. Biddie, S. C. *et al.* Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Mol. Cell* **43**, 145–55 (2011).

46. Lupien, M. *et al.* FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132**, 958–70 (2008).

47. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–93 (2009).

48. Liu, F. & Wu, D. Roles of CTCF in conformation and functions of chromosome. *Semin. Cell Dev. Biol.* **90**, 168–173 (2019).

49. Lee, B.-K. & Iyer, V. R. Genome-wide studies of CCCTC-binding factor (CTCF) and cohesin provide insight into chromatin structure and regulation. *J. Biol. Chem.* **287**, 30906–13 (2012).

50. Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.* **16**, 245–257 (2015).

51. Danino, Y. M., Even, D., Ideses, D. & Juven-Gershon, T. The core promoter: At the heart of gene expression. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1849**, 1116–1131 (2015).

52. Juven-Gershon, T. & Kadonaga, J. T. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev. Biol.* **339**, 225–229 (2010).

53. Zabidi, M. A. & Stark, A. Regulatory Enhancer-Core-Promoter Communication via

Transcription Factors and Cofactors. *Trends Genet.* **32**, 801–814 (2016).

54. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.* **16**, 144–154 (2015).

55. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**, 1170–1187 (2016).

56. Zabidi, M. a *et al.* Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2015).

57. Ko, J. Y. S. K. H. Functional Enhancers As Master Regulators of Tissue-Specific Gene Regulation and Cancer Development. *Mol. Cells* **40**, 169–177 (2017).

58. Ong, C.-T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.* **12**, 283–293 (2011).

59. Pandini, G. *et al.* Androgens Up-regulate the Insulin-like Growth Factor-I Receptor in Prostate Cancer Cells. *Cancer Res.* **65**, 1849 LP – 1857 (2005).

60. Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional Regulatory Elements in the Human Genome. *Annu. Rev. Genomics Hum. Genet.* **7**, 29–59 (2006).

61. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21931–6 (2010).

62. Bae, M. G., Kim, J. Y. & Choi, J. K. Frequent hypermethylation of orphan CpG islands with enhancer activity in cancer. *BMC Med. Genomics* **9**, 38 (2016).

63. Raisner, R. *et al.* Enhancer Activity Requires CBP/P300 Bromodomain-Dependent Histone H3K27 Acetylation. *Cell Rep.* **24**, 1722–1729 (2018).

64. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371–375 (2014).

65. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, (2013).

66. Edelman, L. B. & Fraser, P. Transcription factories: genetic programming in three dimensions. *Curr. Opin. Genet. Dev.* **22**, 110–4 (2012).

67. Rieder, D., Trajanoski, Z. & McNally, J. G. Transcription factories. *Front. Genet.* **3**, 221 (2012).

68. de Santa, F. *et al.* A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. *PLoS Biol.* **8**, (2010).

69. Hah, N., Murakami, S., Nagari, A., Danko, C. G. & Kraus, W. L. Enhancer transcripts mark active estrogen receptor binding sites Enhancer transcripts mark active estrogen receptor binding sites. 1210–1223 (2013). doi:10.1101/gr.152306.112

70. Arner, E. *et al.* Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science (80-.).* **347**, 1010–1014 (2015).

71. Schaukowitch, K. *et al.* Enhancer RNA Facilitates NELF Release from Immediate Early Genes. *Mol. Cell* **56**, 29–42 (2014).

72. Kearns, N. A. *et al.* Cas9 effector-mediated regulation of transcription and differentiation in human pluripotent stem cells. *Development* **141**, 219–23 (2014).

73. Bose, D. A. & Berger, S. L. eRNA binding produces tailored CBP activity profiles to regulate gene expression. *RNA Biol.* **14**, 1655–1659 (2017).

74. Gaszner, M. & Felsenfeld, G. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat. Rev. Genet.* **7**, 703–713 (2006).

75. Capuano, F., Mülleder, M., Kok, R., Blom, H. J. & Ralser, M. Cytosine DNA Methylation Is Found in *Drosophila melanogaster* but Absent in *Saccharomyces cerevisiae, Schizosaccharomyces pombe* , and Other Yeast Species. *Anal. Chem.* **86**, 3697–3702 (2014).

76. Tost, J. DNA Methylation: An Introduction to the Biology and the Disease-Associated Changes of a Promising Biomarker. *Mol. Biotechnol.* **44**, 71–81 (2010).

77. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796–815 (2000).

78. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

79. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–51 (2001).

80. Britten, R. J. & Kohne, D. E. Repeated Sequences in DNA. *Science (80-.).* **161**, 529 LP – 540 (1968).

81. Pennisi, E. ENCODE Project Writes Eulogy for Junk DNA. *Science (80-.).* **337**, 1159 LP – 1161 (2012).

82. Liu, A. M. F., New, D. C., Lo, R. K. H. & Wong, Y. H. Reporter Gene Assays. in 109–123 (Humana Press, Totowa, NJ, 2009). doi:10.1007/978-1-60327-545-3_8

83. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).

84. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).

85. Eberwine, J., Sul, J.-Y., Bartfai, T. & Kim, J. The promise of single-cell sequencing. *Nat. Methods* **11**, 25–27 (2014).

86. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 13770–3 (1996).

87. Langmead, B. Aligning Short Sequencing Reads with Bowtie. in *Current Protocols in Bioinformatics* **32**, 11.7.1-11.7.14 (John Wiley & Sons, Inc., 2010).

88. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

89. Schbath, S. *et al.* Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis. *J. Comput. Biol.* **19**, 796–813 (2012).

90. Menzel, P., Frellsen, J., Plass, M., Rasmussen, S. H. & Krogh, A. On the Accuracy of Short Read Mapping. in 39–59 (Humana Press, Totowa, NJ, 2013). doi:10.1007/978-1-62703-514-9_3

91. Hatem, A., Bozdağ, D., Toland, A. E. & Çatalyürek, Ü. V. Benchmarking short sequence mapping tools. *BMC Bioinformatics* **14**, 184 (2013).

92. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

93. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

94. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

95. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).

96. Aparicio, O., Geisberg, J. V. & Struhl, K. Chromatin Immunoprecipitation for Determining the Association of Proteins with Specific Genomic Sequences In Vivo. in *Current Protocols in Cell Biology* **23**, 17.7.1-17.7.23 (John Wiley & Sons, Inc., 2004).

97. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408–19 (2011).

98. Kidder, B. L., Hu, G. & Zhao, K. ChIP-Seq: technical considerations for obtaining high-quality data. *Nat. Immunol.* **12**, 918–922 (2011).

99. Xiong, L. *et al.* Genome-wide Identification and Characterization of Enhancers Across 10 Human Tissues. *Int. J. Biol. Sci.* **14**, 1321–1332 (2018).

100. Huda, A., Mariño-Ramírez, L., Landsman, D. & Jordan, I. K. Repetitive DNA elements, nucleosome binding and human gene expression. *Gene* **436**, 12–22 (2009).

101. Bartlett, A. *et al.* Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nat. Protoc.* **12**, 1659–1672 (2017).

102. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–31 (2012).

103. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

104. Ibrahim, M. M., Lacadie, S. A. & Ohler, U. JAMM: a peak finder for joint analysis of NGS replicates. *Bioinformatics* **31**, 48–55 (2015).

105. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).

106. Stalker, J. *et al.* The Ensembl Web site: mechanics of a genome browser. *Genome Res.* **14**, 951–5 (2004).

107. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

108. Boyle, A. P. *et al.* High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.* **21**, 456–64 (2011).

109. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* **6**, 283–289 (2009).

110. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).

111. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–37 (2007).

112. Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887–98 (2008).

113. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

114. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499–509 (2004).

115. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).

116. Kumar, V. *et al.* Uniform, optimal signal processing of mapped deep-sequencing data. *Nat. Biotechnol.* **31**, 615–622 (2013).

117. He, H. H. *et al.* Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Methods* **11**, 73–78 (2014).

118. Tsompana, M. & Buck, M. J. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin* **7**, 33 (2014).

119. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

120. Fickett, J. W. & Hatzigeorgiou, A. G. Eukaryotic promoter recognition. *Genome Res.* **7**, 861–78 (1997).

121. Scherf, M., Klingenhoff, A. & Werner, T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.* **297**, 599–606 (2000).

122. Bajic, V. B., Tan, S. L., Suzuki, Y. & Sugano, S. Promoter prediction analysis on the whole human genome. *Nat. Biotechnol.* **22**, 1467–1473 (2004).

123. Wingender, E., Dietze, P., Karas, H. & Knüppel, R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* **24**, 238–241 (1996).

124. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* (2017).

125. Turatsinze, J.-V., Thomas-Chollier, M., Defrance, M. & van Helden, J. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat. Protoc.* **3**, 1578–1588 (2008).

126. Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**,

1647–1649 (2012).

127. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–50 (2005).

128. Ganley, A. R. D. & Kobayashi, T. Phylogenetic Footprinting to Find Functional DNA Elements. in 367–379 (Humana Press, 2007). doi:10.1007/978-1-59745-514-5_23

129. Odom, D. T. *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.* **39**, 730–732 (2007).

130. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).

131. Shen, S. Q. *et al.* Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Res.* **26**, 238–55 (2016).

132. Valensisi, C., Liao, J. L., Andrus, C., Battle, S. L. & Hawkins, R. D. cChIP-seq: a robust small-scale method for investigation of histone modifications. *BMC Genomics* **16**, 1083 (2015).

133. Teng, M. *et al.* A benchmark for RNA-seq quantification pipelines. *Genome Biol.* **17**, 74 (2016).

134. Sundararaman, B. *et al.* Resources for the Comprehensive Discovery of Functional RNA Elements. *Mol. Cell* **61**, 903–13 (2016).

135. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752–1779 (2011).

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

137. Gerstein, M. B. *et al.* Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project. *Science (80-.).* **330**, 1775 LP – 1787 (2010).

138. Roy, S. *et al.* Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. *Science (80-.).* **330**, 1787 LP – 1797 (2010).

139. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

140. Functional annotation of a full-length mouse cDNA collection. *Nature* **409**, 685–690 (2001).

141. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).

142. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–63 (2005).

143. Ravasi, T. *et al.* An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**, 744–52 (2010).

144. Forrest, A. R. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–70 (2014).

145. de Hoon, M., Shin, J. W. & Carninci, P. Paradigm shifts in genomics through the

FANTOM projects. *Mamm. Genome* **26**, 391–402 (2015).

146. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2012).

147. Athar, A. *et al.* ArrayExpress update – from bulk to single-cell expression data. *Nucleic Acids Res.* **47**, D711–D715 (2019).

148. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).

149. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

150. Koster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

151. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).

152. Grüning, B. *et al.* Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* **15**, 475–476 (2018).

153. Yates, B. *et al.* Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.* **45**, D619–D625 (2017).

154. Gremse, M. *et al.* The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.* **39**, D507–D513 (2011).

155. Bao, S. *et al.* Evaluation of next-generation sequencing software in mapping and assembly. *J. Hum. Genet.* **56**, 406–414 (2011).

156. Ruffalo, M., LaFramboise, T. & Koyuturk, M. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* **27**, 2790–2796 (2011).

157. Thomas, R., Thomas, S., Holloway, A. K. & Pollard, K. S. Features that define the best ChIP-seq peak calling algorithms. *Brief. Bioinform.* **18**, bbw035 (2016).

158. Kharchenko, P. V, Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* **26**, 1351–1359 (2008).

159. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576–589 (2010).

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

161. Dozmorov, M. G. *et al.* Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data. *BMC Bioinformatics* **16**, S10 (2015).

162. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

163. Vilella, A. J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19**, 327–35 (2009).

164. Griffon, A. *et al.* Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Res.* **43**, e27 (2015).

165. E pluribus unum. *Nat. Methods* **7**, 331–331 (2010).

166. Oki, S. *et al.* Ch IP -Atlas: a data-mining suite powered by full integration of public Ch IP -seq data. *EMBO Rep.* **19**, (2018).

167. Mei, S. *et al.* Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.* **45**, D658–D662 (2017).

168. Yevshin, I., Sharipov, R., Valeev, T., Kel, A. & Kolpakov, F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res.* **45**, D61–D67 (2017).

169. Worsley Hunt, R., Mathelier, A., del Peso, L. & Wasserman, W. W. Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. *BMC Genomics* **15**, 472 (2014).

170. Bailey, T. L. & Machanick, P. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.* **40**, e128–e128 (2012).

171. Wilbanks, E. G. & Facciotti, M. T. Evaluation of Algorithm Performance in ChIP-Seq Peak Detection. *PLoS One* **5**, e11471 (2010).

172. Kendall, M. G. (Maurice G. & Ord, J. K. *Time series*. (E. Arnold, 1990).

173. Jain, D., Baldi, S., Zabel, A., Straub, T. & Becker, P. B. Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments. *Nucleic Acids Res.* **43**, 6959–6968 (2015).

174. Pickrell, J. K., Gaffney, D. J., Gilad, Y. & Pritchard, J. K. False positive peaks in ChIP-seq and other sequencing-based functional assays caused by unannotated high copy number regions. *Bioinformatics* **27**, 2144–2146 (2011).

175. Teytelman, L., Thurtle, D. M., Rine, J. & van Oudenaarden, A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 18602–7 (2013).

176. Sharma, N. L. *et al.* The ETS family member GABPα modulates androgen receptor signalling and mediates an aggressive phenotype in prostate cancer. *Nucleic Acids Res.* **42**, 6256–6269 (2014).

177. Vlieghe, D. *et al.* A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.* **34**, D95–D97 (2006).

178. Stormo, G. D. Modeling the specificity of protein-DNA interactions. *Quant. Biol.* **1**, 115–130 (2013).

179. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822 (1990).

180. Thomas-Chollier, M. *et al.* RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res.* **40**, e31–e31 (2012).

181. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).

182. Sandelin, A. & Wasserman, W. W. Constrained Binding Site Diversity within Families of Transcription Factors Enhances Pattern Discovery Bioinformatics. *J. Mol. Biol.* **338**, 207–215 (2004).

183. Grosselin, K. *et al.* High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat. Genet.* **51**, 1060–1066 (2019).

184. Worsley Hunt, R. & Wasserman, W. W. Non-targeted transcription factors motifs are a systemic component of ChIP-seq datasets. *Genome Biol.* **15**, 412 (2014).

185. Kapur, J. N., Sahoo, P. K. & Wong, A. K. C. A new method for gray-level picture thresholding using the entropy of the histogram. *Comput. Vision, Graph. Image Process.* **29**, 140 (1985).

186. Hume, M. A., Barrera, L. A., Gisselbrecht, S. S. & Bulyk, M. L. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein–DNA interactions. *Nucleic Acids Res.* **43**, D117–D122 (2015).

187. Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5**, 276–287 (2004).

188. Hacker, K. *et al.* A repetitive acidic region contributes to the extremely rapid degradation of the cell-context essential protein TRIM52. *Sci. Rep.* **9**, 7901 (2019).

189. Liu, Y., Tanikawa, C., Ueda, K. & Matsuda, K. INKA2, a novel p53 target that interacts with the serine/threonine kinase PAK4. *Int. J. Oncol.* **54**, 1907–1920 (2019).

190. Hawe, J. S., Theis, F. J. & Heinig, M. Inferring Interaction Networks From Multi-Omics Data. *Front. Genet.* **10**, 535 (2019).

191. Kouno, T. *et al.* C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. *Nat. Commun.* **10**, 360 (2019).

192. González, S. *et al.* ReLA, a local alignment search tool for the identification of distal and proximal gene regulatory regions and their conserved transcription factor binding sites. *Bioinformatics* **28**, 763–770 (2012).

193. Abdulghani, M., Jain, A. & Tuteja, G. Genome-wide identification of enhancer elements in the placenta. *Placenta* **79**, 72–77 (2019).

194. Yevshin, I., Sharipov, R., Kolmykov, S., Kondrakhin, Y. & Kolpakov, F. GTRD: a database on gene transcription regulation—2019 update. *Nucleic Acids Res.* **47**, D100–D105 (2019).

195. Sheffield, N. C. *et al.* Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Res.* **23**, 777–88 (2013).

196. Nakato, R. & Shirahige, K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Brief. Bioinform.* **18**, bbw023 (2016).

197. Zhang, Q. *et al.* Systematic evaluation of the impact of ChIP-seq read designs on genome coverage, peak identification, and allele-specific binding detection. *BMC Bioinformatics* **17**, 96 (2016).

198. Krebs, W. *et al.* Optimization of transcription factor binding map accuracy utilizing knockout-mouse models. *Nucleic Acids Res.* **42**, 13051–13060 (2014).

199. Ambrosini, G., Groux, R. & Bucher, P. PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. *Bioinformatics* **34**, 2483–2484 (2018).

200. Li, Z. *et al.* Identification of transcription factor binding sites using ATAC-seq.

Genome Biol. **20**, 45 (2019).

201. Clark, S. J., Lee, H. J., Smallwood, S. A., Kelsey, G. & Reik, W. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biol.* **17**, 72 (2016).
202. Rotem, A. *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33**, 1165–1172 (2015).
203. Pasini, D. *et al.* JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. *Nature* **464**, 306–310 (2010).
204. Sims, R. J. *et al.* Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. *Mol. Cell* **28**, 665–76 (2007).
205. Rossow, K. L. & Janknecht, R. Synergism between p68 RNA helicase and the transcriptional coactivators CBP and p300. *Oncogene* **22**, 151–156 (2003).
206. Pasini, D., Bracken, A. P., Jensen, M. R., Denchi, E. L. & Helin, K. Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. *EMBO J.* **23**, 4061–4071 (2004).
207. Euskirchen, G., Auerbach, R. K. & Snyder, M. SWI/SNF chromatin-remodeling factors: multiscale analyses and diverse functions. *J. Biol. Chem.* **287**, 30897–905 (2012).
208. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
209. Barakat, T. S. *et al.* Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell Stem Cell* **23**, 276-288.e8 (2018).
210. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309-1324.e18 (2018).
211. Vacik, T. & Raska, I. Alternative intronic promoters in development and disease. *Protoplasma* **254**, 1201–1206 (2017).
212. Huin, V. *et al.* Alternative promoter usage generates novel shorter MAPT mRNA transcripts in Alzheimer's disease and progressive supranuclear palsy brains. *Sci. Rep.* **7**, 12589 (2017).
213. Burton, E. A., Tinsley, J. M., Holzfeind, P. J., Rodrigues, N. R. & Davies, K. E. A second promoter provides an alternative target for therapeutic up-regulation of utrophin in Duchenne muscular dystrophy. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 14025–30 (1999).
214. Hanssen, L. L. P. *et al.* Tissue-specific CTCF–cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nat. Cell Biol.* **19**, 952–961 (2017).
215. Smith, E. M., Lajoie, B. R., Jain, G. & Dekker, J. Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus. *Am. J. Hum. Genet.* **98**, 185–201 (2016).
216. Rahman, N. I. A., Abdul Murad, N. A., Mollah, M. M., Jamal, R. & Harun, R. NFIX as a Master Regulator for Lung Cancer Progression. *Front. Pharmacol.* **8**, 540 (2017).
217. Saini, S., Sharma, I., Kaur, N. & Pati, P. K. Auxin: a master regulator in plant root

development. *Plant Cell Rep.* **32**, 741–757 (2013).

218. Tiwari, N. *et al.* Sox4 Is a Master Regulator of Epithelial-Mesenchymal Transition by Controlling Ezh2 Expression and Epigenetic Reprogramming. *Cancer Cell* **23**, 768–783 (2013).

219. Yang, J. *et al.* Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. *Cell* **117**, 927–39 (2004).

220. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–61 (2014).

Résumé

Le processus de régulation de la transcription des gènes repose très largement sur l'existence de séquences d'ADN non codantes dans le génome. Ces séquences d'ADN, appelées "éléments cis-régulateurs", ont la particularité de recruter de nombreuses protéines capables de réguler le niveau de transcription des gènes. Parmi ces protéines, les facteurs de transcription sont capables de se fixer directement sur l'ADN. Les facteurs de transcription coopèrent avec d'autres protéines régulatrices, les cofacteurs, afin de réguler la transcription. Les protéines régulatrices de la transcription permettent la fixation et la régulation de l'enzyme d'ARN polymérase II qui transcrit les gènes en ARN messager. Leurs fixations sur les éléments cis-régulateurs permettent une régulation des gènes dans l'espace et dans le temps. Pour mieux comprendre la régulation de l'expression des gènes, il est nécessaire d'identifier les éléments cis-régulateurs dans le génome afin de caractériser et d'identifier les mécanismes d'action des éléments régulateurs et des protéines qui leur sont liés. Le développement rapide des méthodes de séquençage à haut débit a permis l'identification des interactions ADN/protéines à grande échelle. L'accumulation massive des données de séquençage dans les banques de données publiques permet l'intégration de nombreuses expériences capturant les interactions entre les facteurs de transcription et l'ADN par des moyens bioinformatiques. Le but de mon doctorat a été d'annoter et traiter de façon uniforme les données brutes issues d'expériences de séquençage ayant pour objectif d'identifier les régions de fixation des protéines régulatrices pour l'Homme puis chez *Arabidopsis Thaliana*. Nous avons traité des données de ChIP-seq, ChIP-exo et DAP-seq afin d'élaborer plusieurs catalogues de régions régulatrices chez l'homme et chez Arabidopsis *Thaliana.* Toutes ces données sont disponibles au sein du projet ReMap. Pour *Arabidopsis Thaliana, n*ous avons complété ces données par une analyse de toutes les marques d'histones. Pour effectuer ces analyses, nous avons développé des workflows reproductibles, scalables et portables sur des architectures différentes. Cette analyse intégrative à haut débit nous a permis d'identifier de nombreux nouveaux éléments cis-régulateurs. Ces données ont également été utilisées pour identifier les sites de fixations reconnus par les facteurs de transcription et pour consolider la base de données JASPAR pour l'Homme et pour *Arabidopsis Thaliana*. Enfin, ce catalogue a été utilisé dans le développement d'une nouvelle méthode appliquant un algorithme basé sur l'entropie. Cet algorithme permet de différencier les événements de fixations directes et indirectes par les protéines dans les résultats de ChIP-seq.

Abstract

The regulation of gene transcription is largely based on the existence of non-coding DNA sequences in the genome. These DNA sequences, called "cis-regulatory elements", have the particularity of recruiting many proteins capable of regulating the level of gene transcription. Among these proteins, transcription factors are capable of directly binding to DNA. Transcription factors cooperate with other regulatory proteins, called cofactors, to regulate transcription. Transcription regulatory proteins allow the binding and regulation of the RNA polymerase II enzyme that transcribes genes into messenger RNA. The fixation of transcription factors on the cis-regulatory elements allows the regulation of genes in space and time. To better understand the regulation of gene expression, it is necessary to identify cis-regulatory elements in the genome in order to characterize and identify the mechanisms of action of regulatory elements and the proteins that are linked to them. The rapid development of high throughput sequencing methods has made it possible to identify DNA/protein interactions on a large scale. The massive accumulation of sequencing data in public databases allows the integration of many experiments that capture the interactions between transcription factors and DNA through bioinformatics. The purpose of my PhD was to annotate and process in a uniform way the raw data from sequencing experiments whose objective is to identify the binding regions of regulatory proteins for humans and then for Arabidopsis Thaliana. We processed data from ChIP-seq, ChIP-exo and DAP-seq to develop several catalogues of regulatory regions in humans and Arabidopsis Thaliana. All this data is available within the ReMap project. We completed them with an analysis of all histone brands for Arabidopsis Thaliana. To carry out these analyses we have developed reproducible, scalable and portable workflows working on different architectures. This high throughput integrative analysis has allowed us to identify many new cis-regulatory elements. These data were also used to identify the attachment sites recognized by the transcription factors and to consolidate the JASPAR database for humans and Arabidopsis Thaliana. Finally, this catalogue was used in the development of a new method applying an entropy-based algorithm to differentiate between direct and indirect protein binding events in ChIP-seq results.