Étude des éléments cis-régulateurs à différentes échelles

Fayrouz Hammal

Aix*Marseille université
Socialement engagée

NNT : 2023AIXM0001

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 10 mai 2023 par

Fayrouz HAMMAL

Étude des éléments cis-régulateurs à différentes échelles

Approches fondamentales à appliquées

Discipline
Biologie santé

Spécialité
Génomique et Bioinformatique

École doctorale
ED 62 - Sciences de la Vie et de la Santé

Laboratoire/Partenaires de recherche

TAGC - Theories and Approaches of Genomic Complexity

Région Sud - Conseil Régional de la région Provence-Alpes-Côte d'Azur

ABD - Advanced BioDesign

Composition du jury

Camille BERTHELOT CR - INSTITUT PASTEUR	Rapportrice
Gaël CRISTOFARI DR - IRCAN	Rapporteur
Anaïs BARDET CR - IGBMC	Examinatrice
Pascal RIHET PR - TAGC	Examinateur
Denis PUTHIER PR - TAGC	Examinateur
Benoît BALLESTER CR - TAGC	Directeur de thèse
Mileidys PEREZ ALEA CSO - ABD	Membre invitée

RÉGION SUD PROVENCE ALPES CÔTE D'AZUR

advanced biodesign

umr U 1090 TAGC theories and approaches of genomic complexity

Affidavit

Je soussignée, Fayrouz Hammal, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Benoit Ballester, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille le 25 février 2023

Liste de publications et participation aux conférences

Liste des publications réalisées dans le cadre du projet de thèse :

1. **Hammal, F.**, de Langen, P., Bergon, A., Lopez, F., & Ballester, B. (2022). ReMap 2022 : a database of human, mouse, drosophila and arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. Nucleic Acids Research, 50(D1), D316-D325.
2. Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., Fornes, O., Y Leung, T., Aguirre, A., **Hammal, F.**, Schmelter, D., Baranasic, D., Ballester, B., ... , & Mathelier, A. (2022). JASPAR 2022 : the 9th release of the open-access database of transcription factor binding profiles. Nucleic acids research, 50(D1), D165-D173.

Participation aux conférences au cours de la période de thèse :

1. 2022. Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM), Rennes - Poster
2. 2022. ECCB2022 : the 21st European Conference on Computational Biology, Sitges (Espagne) - Poster

Liste de participation aux missions de culture scientifique :

1. 2021. Fête de la Science : atelier « Les bêtes de laboratoire au service de la génétique », Marseille (16h)
2. 2021. Formation à la médiation scientifique Chercheur.se en classe, Marseille (30h)
3. 2020. La Nuit Européenne des Chercheur.e.s, Marseille (3h)

Résumé

L'ADN non-codant a longtemps été considéré comme inutile. Cependant, grâce aux avancées technologiques dans le domaine de la génomique, les chercheurs ont réalisé l'importance de ces régions dans la régulation des gènes. Les acteurs de la régulation sont les éléments *cis*-régulateurs (CRE) tels que les promoteurs et enhancers. Ces éléments sont fixés par des protéines appelées facteurs de la transcription (TF), qui se fixent à des sites de fixation spécifique. Les TF agissent seuls ou sous forme de complexes pour recruter l'ARN polymérase II et initier la transcription. Avec l'avénement des techniques de séquençage à haut débit, une technique permettant de détecter la fixation des protéines à l'ADN a vu le jour, le ChIP-seq. Les données résultantes sont stockées dans des entrepôts de données tels que GEO. Cependant, il existe une grande diversité dans la manière dont les expériences ChIP-seq sont conçues. Le projet ReMap, lancé en 2012, vise à identifier les régions régulatrices en annotant et intégrant ces données uniformormément. En 2022, j'ai ajouté deux nouvelles espèces au catalogue : *Mus musculus* et *Drosophila melanogaster* et la mise à jour pour l'Homme et *Arabidopsis thaliana*. Le catalogue chez l'Homme comprend 1210 TF et 182 millions de pics ChIP-seq. L'augmentation continue des données intégrées a nécessité de nouveaux filtres qualités. Ces données sont visualisables sur le navigateur de génomes de UCSC et peuvent être filtrés en fonction des TF et biotypes. Enfin, nous avons tenté de déterminer une spécificité tissulaire des modules de régulations ReMap à l'aide de la méthode ChromHMM. Les résultats préliminaires permettent de distinguer 11 tissus avec un modèle à 25 états.

La deuxième partie de ma thèse se concentre sur l'impact des éléments transposables (TE) sur l'insertion de site de fixation au TF dans les génomes au cours de l'évolution. Il y a plus de 50 ans, Barbara McClintock a découvert pour la première fois les éléments transposables dans le maïs (*Zea mays*) et les a appelés "éléments de contrôle". Depuis, les chercheurs s'éfforcent de classer et caractériser ces TE. Des travaux en génomiques décrivent les TE comme une source abondante de matériaux pour l'assemblage et la modification des systèmes régulateurs des gènes eucaryotes au cours de l'évolution. Nous avons donc réalisé une analyse à grande échelle afin de déterminer l'étendue de ce phénomène. Pour ce faire, nous avons réalisé une analyse d'enrichissement des 1210 pics TF de ReMap sur les TE avec l'outil LOLA. On identifie donc 15,441 paires de TE/TF significativement associés. A l'aide de l'outil FIMO nous avons détecté la présence des motifs de fixation des TF dans la séquence des TE associés pour 7,757 paires. Nous avons observé une spécificité de l'association TE/TF aux groupes de TF et aux familles de TE. Après avoir visualisé l'alignement des séquences de TE nous observons que les motifs sont alignés, témoignant de leur conservation au cours de l'évolution. Nous avons également observé que l'âge d'insertion des TE associés est différent pour chaque groupe de TF.

Enfin, nous avons réalisé un projet en collaboration avec Advanced BioDesign (ABD) qui porte sur la régulation des ALDH. En effet, la surexpression des ALDH est associée à un mauvais pronostic chez les patients atteints d'AML. ABD a donc développé un traitement inhibiteur des ALDH, le DIMMATE. Dans ce contexte, nous cherchons à mieux comprendre les mécanismes de la régulation de l'ALDH1A1. J'ai donc cartographié les éléments régulateurs autour de ce gène à l'aide de données multi-omiques. Au cours de notre analyse, nous avons identifié 14 régions régulatrices autour du gène ALDH1A1. Nos résultats ont ensuite été communiqués à nos collaborateurs afin de procéder aux validation expérimentales qui sont encore en cours.

Ces travaux se sont portés à trois niveaux de recherche 1) à très grande échelle, 2) à l'échelle génomique "évolutive" et 3) focalisés sur un locus précis.

Mots clés : ADN non codant, régulation, éléments *cis*-régulateurs, facteurs de transcription, ChIP-seq, ReMap, éléments transposables, TFBS, évolution, ALDH, AML

Abstract

Non-coding DNA has long been considered as junk. However, thanks to technological advances in genomics, researchers have realized the importance of these regions in gene regulation. The regulation is mediated by *cis*-regulatory elements (CRE) such as promoters and enhancers. These elements are bound by proteins called transcription factors (TF), which bind to specific binding sites. TF act alone or as complexes to recruit RNA polymerase II and initiate transcription. With the advent of high-throughput sequencing techniques, a technique for detecting protein binding to DNA has emerged, ChIP-seq. The resulting data is stored in data repositories such as GEO. However, there is a great diversity in metadata and processing of ChIP-seq experiments. It is in this context that the ReMap project was launched in 2012. This project aims to identify CRE by annotating and integrating these data in a uniform manner. In the first part of my thesis, I added two new species to the catalog, mouse (*Mus musculus*) and fly (*Drosophila melanogaster*), and also updated it for humans and *Arabidopsis thaliana*. The human catalog includes 1210 TF and 182 million ChIP-seq peaks. The continuous increase in integrated ChIP-seq data required new filters to ensure data quality. These data are visualizable on the UCSC genome browser and can be filtered by TF and biotypes. Finally, we attempted to determine tissue specificity of the regulatory modules of ReMap using the ChromHMM method. Preliminary results distinguish 11 tissues with a 25-state model.

The second part of my thesis focuses on the impact of transposable elements (TE) on the insertion of TF binding site in genomes during evolution. More than 50 years ago, Barbara McClintock first discovered transposable elements in maize (*Zea mays*) and called them "controlling elements". Since then, researchers have been working to classify and characterize these TE. Many genomic studies describes TE as an abundant source of materials for the assembly and modification of eukaryotic gene regulatory systems during evolution. We therefore decided to carry out a large-scale analysis to determine the extent of this phenomenon on the human genome. To do so, we performed an enrichment analysis of the 1210 ReMap TF peaks on TE using the LOLA tool. We identified 15,441 significantly associated TE/TF pairs. Using the FIMO tool, we detected the presence of TF binding motifs in the TE sequence associated with 5,691 pairs. We observed a specificity of TE/TF association to TF groups and TE families. After visualizing the TE sequence alignment, we observed that the motifs were aligned, indicating their conservation during evolution. We also observed that the insertion age of associated TE is different for each TF group.

Finally, we conducted a project in collaboration with Advanced BioDesign (ABD) on the regulation of ALDH. Indeed, ALDH overexpression is associated with poor prognosis in patients with AML. ABD has developed an ALDH inhibitor treatment, DIMMATE. In this context, we aim to better understand the mechanisms of ALDH regulation. I mapped the regulatory elements around this gene using multi-omics data. During our analysis, we identified 14 regulatory regions around the ALDH1A1 gene. Our results were then communicated to our collaborators for experimental validation that are still ongoing.

These works were conducted at three levels of research : i) at a very large scale, ii) evolutionary/fundamental genomics, and iii) focused on a specific locus.

Keywords : Non-coding DNA, regulation, *cis*-regulatory elements, transcription factors, ChIP-seq, ReMap, transposable elements, TFBS, evolution, ALDH, AML.

Remerciements

Tout d'abord, je tiens à remercier mon directeur de thèse, Benoit Ballester, sans qui rien n'aurait été possible. Merci Benoit pour la formidable opportunité que tu m'as offerte. Je sais que je n'ai pas été l'étudiante la plus parfaite et ces années ont été particulièrement difficiles, mais tu as été un mentor attentif, impliqué et compréhensif tout au long des trois années que j'ai passé au TAGC et j'ai beaucoup appris sous ta tutelle.

Je tiens à remercier Camille Berthelot, Gaël Cristofari, Anaïs Bardet, Pascal Rihet, Denis Puthier et Mileidys Perez Alea d'avoir accepté de faire partie de mon jury de soutenance. Je remercie également Pascal Rihet de m'avoir acceuillie dans son laboratoire et d'avoir été co-directeur les deux premières années de thèse.

Je remercie l'INSERM et la Région SUD pour le programme "Emplois Jeunes Doctorants", ainsi que le Service Recherche, Enseignement Supérieur, Santé, Innovation pour avoir financé ma thèse.

Je remercie les membres du laboratoire pour leur aide au cours de ces trois années et pour toutes les conversations de couloir que nous avons partagé. Je te remercie Pierre d'avoir supporté les changements de température intempestifs dans le bureau, Jean-Christophe après cette thèse je lirais tout ce que tu m'as conseillé. Plus spécifiquement, je tiens à remercier Yasmine Labiad et Mileydis Perez pour leur aide durant mes dernières semaines alors que j'écrivais cette thèse, elles m'ont fourni des commentaires, des conseils et un soutien sans lequel cette thèse aurait été beaucoup plus difficile à écrire. Je remercie également Aurélie Bergon et Fabrice Lopez pour leur participation au projet ReMap.

Je remercie mes amis Nono, Steph, Lila, Flo et Sephora qui ont été une source de soutien tout au long de ces années passées à étudier. Je tiens à remercier spécifiquement Salouha de m'avoir hébergée, aidée et avec qui j'ai passé les meilleurs moments. Tu as été une amie précieuse et j'espère que notre amitié durera pendant de longues années. Tu es et tu resteras la plus belle rencontre de ces trois années. Je tiens à te remercier Adama, pour ton aide, ta bonne humeur et tes mots gentils, j'espère te voir le jour de ma soutenance.

Enfin, ma famille a joué un rôle fondamental dans mon succès scientifique. Ma grand mère, Fadila, m'a toujours encouragé à aller aussi loin que possible dans la carrière que je choisirais et à travailler dur pour arriver au sommet. Mes parents chéris m'ont toujours soutenu et m'ont encouragé chaque jour malgré ma mauvaise humeur et mes sautes d'humeur. Je remercie ma maman pour ses nombreuses corrections syntaxiques et orthographiques. Enfin, mes soeurs adorées, Morjane, Shayma et Shérihane avec qui j'ai partagé les hauts et les bas de chaque doctorant, elles ont toujours été là pour me rappeler que je n'étais pas seule. Sans vous je ne serais jamais arrivé jusque là. Je tiens à remercier plus particulièrement ma soeur Shayma qui a passé de longues heures à travailler en même temps que moi pendant le COVID, ne t'inquiète pas ma soeur je te revaudrais toutes ces nuits blanches.

A mon Tonton Foued, le plus affectueux de mes tontons, celui que j'avais hâte de revoir mais que le COVID m'a pris et que je ne reverrai plus jamais. Je t'aime.

Table des matières

Table des figures

Liste des tableaux

Liste des acronymes

A. thaliana
Arabidopsis thaliana. 23, 117

ABD
Laboratoire Advanced BioDesign. 24

ALDH
Aldéhydes Déshydrogénases. 24, 100

AML
Leucémie aigüe myeloide. 24, 94, 95, 186, 230

ARNm
ARN messager. 29, 31

cCREs
candidat textitCis-regulatory elements. 141, 217, 218

CRE
textitCis-regulatory elements. 217, 219

CRM
Cis-regulatory Module. 127, 129, 140

DHS
Sites hypersensibles à la DNase. 65

ENCODE
Encyclopedia of DNA Elements. 23, 77, 82, 105, 126, 154

ER
Endocucléase de réstriction. 65

ERV
Rétrovirus endogène. 92

FRiP
Fraction of Reads in Peaks. 77, 130

GEO
Gene Expression Omnibus. 23, 83, 105, 126

GTF
Facteurs de transcrpiton généraux. 29, 48

HSC
Cellules souches hématopoïétiques. 101

IC
contenu informationnel. 75, 77

LINE
Eléments nucléaires intercalés. 88

LSC
Cellule souche de la leucémie. 101

LTR
Longues répétitions terminales. 88, 170

Mya
Million years ago. 180, 181, 184

NR
Non redondant. 165

NSC
Normalized Strand Correlation. 77, 130

ORF
Open Reading Frame ou Cadre de lecture ouverte. 88, 90

pb
Paires de bases. 34

PFM
Matrice de fréquence de position. 75, 77, 154

PIC
Complexe de préinitiation. 29

Pol II
ARN polymérase II. 26

PWM
Matrices de poids de position. 75, 77

RSC
Relative Strand Correlation. 77, 130

RT
Transcriptase inverse. 88

STR
Short Tandem Repeat. 85, 228

TE
Element(s) transposable(s). 23, 24, 84, 148, 224

TF
Facteurs de transcription. 25, 105, 148, 187, 219

TFBS
Site de fixation au facteur de la transcription. 23, 29, 51, 84, 148

TIR
Répétitions inverses terminales. 90, 91

TR
Régulateurs transcriptionels ou Transcriptional regulators. 122–125, 129, 151

TSS
Site d'initiation de la transcription. 29, 31

TTS
Site de terminaison de la transcription. 31

UTR
Région non traduite. 31, 88

Glossaire

BAM

Un fichier BAM (*.bam) est la version binaire compressée d'un fichier SAM qui est utilisée pour représenter des séquences alignées jusqu'à 128 Mb.. 72, 127

BED

Le format BED (ou format Browser Extensible Data) est un format de fichier texte utilisé pour stocker des régions génomiques sous forme de coordonnées ainsi que les annotations associées.. 117, 127

FASTQ

Format de fichier texte permettant de stocker à la fois une séquence biologique (généralement une séquence nucléotidique) et ses scores de qualité correspondants.. 71, 117, 127

PM

Le pourcentage de mottif (PM) est la proportion de séquences de TE présentant un ou des motifs identifié par fimo. Ce calcul est réalisé sur les paires TE/TF avec un enrichissement significatifs en site de fixation. Le PM évalue donc la distribution globale des motifs de fixation identifiés dans les séquences de TE par paire de TE/TF.. 170

SAM

Un fichier SAM (Sequence Alignment Map) est un format de texte utilisé pour stocker des séquences biologiques alignées sur une séquence de référence.. 72, 127

SAMtools

SAMtools est un ensemble d'outils permettant d'interagir avec des alignements de séquences d'ADN courtes en format SAM, BAM et CRAM, ainsi que de les traiter après leur alignement.. 72, 127

Préambule

Au cours de ma thèse, j'ai travaillé sur trois grands axes de recherche : la mise à jour du catalogue de régions régulatrices ReMap, l'analyse de l'impact des éléments transposables sur la régulation et enfin, la cartographie des éléments régulateurs du gène ALDH1A1.

La première partie de ma thèse se concentre sur le catalogue de régions régulatrices appelé ReMap. Ce catalogue, disponible depuis 2015, répertorie les données ChIP-seq de facteurs de transcription provenant de GEO et ENCODE. ReMap inclut également une annotation et une curation manuelle ainsi que des contrôles qualité. J'ai développé la quatrième version en 2022, en mettant à jour les catalogues pour l'Homme et pour *A. thaliana*. Le catalogue de l'Homme contient 279 millions de pics en 2022 contre 165 millions de pics en 2020, ce qui témoigne d'une augmentation significative du nombre de données disponibles. J'ai également inclu deux nouveaux catalogues de régions régulatrices pour la souris et la drosophile. Dans le chapitre portant sur le catalogue ReMap (chapitre 2), nous aborderons les différentes modifications que j'ai pu apporter lors de la mise à jour ainsi que les analyses effectuées sur les données obtenues au cours de ma thèse.

La deuxième partie de ma thèse, vise à étudier l'impact des éléments transposables (TE) sur l'insertion des TFBS dans le génome au cours de l'évolution. Une analyse d'enrichissement a été effectuée pour déterminer la présence des TFBS dans les séquences de TE des génomes de l'Homme. Les résultats montrent une association significative pour 15,441 paires TE/TF. Afin d'identifier par quels mécanismes les TF se fixaient aux TE, nous avons recherché les motifs de fixation des TF dans les séquences de TE associés. Cette analyse a été réalisée pour 7,757 paires TE/TF associées avec un enrichissement significatif. Les résultats révèlent la présence de motifs dans la séquence des TE. Nous avons également mis en valeur une spécificité d'association entre les familles de TE et les groupes de TF. En effet, les groupes de TF se fixent préférentiellement sur certaines familles de TE par l'intermédiaire de motifs de fixation (ex : GATA* avec le L1, NFY* avec les ERV1) . L'alignement des TE révéle une haute conservation des motifs dans les séquences de TE. Enfin, nous avons cherché à identifier une corrélation entre l'âge des éléments transposables avec la fixation des TF sur leur séquences. Ces analyses seront présentées dans le chapitre 3.

Mon projet de thèse a été financé par une bourse de l'INSERM-Région SUD, qui implique une collaboration avec un laboratoire privé. Ainsi, j'ai eu le plaisir de travailler avec le laboratoire Advanced BioDesign (ABD), pour étudier la régulation du gène ALDH1A1 dans les leucémies myéloïdes aiguës (AML). Cette étude s'est appuyée sur le catalogue de régions régulatrices ReMap ainsi que des données de génomiques (marques d'histones, chromatine ouverte, Hi-C). L'objectif de notre collaboration était d'identifier les régulateurs de la transcription des ALDH. Ce projet pourrait aboutir à l'identification d'une nouvelle cible thérapeutique, qui pourrait être utilisée en complément du DIMMATE, un inhibiteur des ALDH. Nous avons donc construit un trackhub multi-omique permettant de partager et de visualiser les données avec nos partenaires. Ce trackhub contient des données de ChIP-seq provenant de ReMap ainsi que des données d'ATAC-seq, de DNase-seq et de ChIP-seq d'histones. Grâce à notre analyse, nous avons identifié 11 régions qui pourraient agir en tant qu'enhancers et deux régions qui pourraient agir en tant que promoteurs autour du gène ALDH. Ce travail est en cours de validation par notre collaboratrice Yasmine LABIAD afin de tester les régions identifiées et les facteurs qui s'y fixent (chapitre 4).

Dans l'introduction nous nous efforcerons de présenter tout les aspects abordés lors de cette thèse. Le chapitre 1.1 introduira les principes de base de la régulation des eucaryotes. Cette section servira de référence pour le reste du manuscrit. Elle fournira un contexte biologique nécessaire à sa compréhension.

Nous présenterons ensuite dans les chapitres 1.2 et 1.3, les méthodes de sequencages utilisées dans le cadre de la thèse, ainsi que les analyses bioinformatiques ayant servi à traiter ces données. Le chapitre 1.4 présentera les grands consortiums et bases de données génomiques existantes.

Le chapitre 1.5.1 portera sur le deuxième axe de cette thèse, l'implication des TE sur la régulation. Nous décrirons les différents éléments transposables et leur classification. Et établirons également un état de l'art portant sur les éléments transposables dans le contexte de la régulation.

Enfin, le chapitre 1.6 portera sur le troisième axe de thèse, la régulation de l'ALDH1A1 dans les AML. Nous décrirons donc les leucémies et plus particulièrement les AML. Nous présenterons par la suite les ALDH ainsi que leur rôle dans les AML d'après la litérature.

1. Introduction

1.1. La régulation chez les eukaryotes

1.1.1. Introduction

Le nombre exact de gènes humains n'est toujours pas connu, mais les estimations actuelles se situent entre 20 000 pour les gènes codants et 26 000 pour les non codants [1]. Ce nombre relativement faible de gènes est suffisant pour maintenir tous les processus biologiques au cours du développement et de la vie humaine. Beaucoup de nos gènes ont été conservés tout au long de l'évolution et peuvent être trouvés dans des espèces éloignées telles que la levure, ce qui montre donc l'impact d'autres processus dans la diversité des espèces. Il est tout aussi important de savoir comment et quand les gènes sont utilisés. Ceci est régulé par un mécanisme complexe qui contrôle l'activité des gènes. Seul un sous-ensemble de tous les gènes est actif dans une cellule à un moment donné, et les protéines produites à partir de ces gènes déterminent essentiellement la fonction cellulaire. Le grand nombre de combinaisons possibles de gènes actifs donne aux cellules la flexibilité nécessaire pour soutenir tous les différents processus biologiques.

Au fur et à mesure que de nombreux génomes d'organismes différents ont été séquencés, il a été observé que la séquence d'ADN de ces gènes reste souvent très similaire, même entre des espèces éloignées. Il a été révélé par la suite qu'une grande partie des différences inter espèces, est expliquée par le réseau d'interactions des gènes entre eux : le réseau de régulation des gènes[37]. Des protéines appelées facteurs de transcription (TF) sont responsables d'une grande partie de la régulation. Les TF peuvent se lier aux régions régulatrices de la séquence appelée éléménts *cis*-régulateurs. Cette fixation permet d'activer ou de réprimer la transcription des gènes. Les protéines histones qui sont des composants de la chromatine jouent également un rôle majeur dans la régulation, puisque les modifications de ces protéines peuvent rendre l'ADN plus ou moins accessible à la machinerie transcriptionnelle. Ces notions seront détaillées dans les parties suivantes. Afin d'apprécier les différentes formes que peut prendre la régulation, il est instructif de présenter brièvement les différentes étapes du processus d'activation de l'expression des gènes. Nous allons donc présenter les différentes étapes de la transcription ainsi que les éléments *cis*-régulateurs impliqués dans la régulation génique.

1. https://www.ensembl.org/Homo_sapiens/Info/Annotation

1.1.2. Le Génome : de la découverte à la compréhension

En 1953 J.Watson et F.Crick ont décrit la structure de l'ADN [306]. Trois ans plus tard, en 1955 George Gamow découvre la synthèse des protéines à partir de l'ADN [96]. Ensuite en 1961, la nature des codons a été découverte par Marshall Nirenberg and Heinrich J. Matthaei [211]. C'est ainsi qu'il a été établi que les organismes vivants stockent les informations nécessaires à leur fonctionnement dans leur génome. L'origine du mot génome provient de la concaténation des termes gène et chromosome par le botaniste Hans Winkler en 1920, ce mot désigne donc l'ensemble des chromosomes et des gènes d'une espèce ou d'un individu. Les chromosomes (Figure 1.1 panel 6) sont des morceaux d'ADN enroulés qui stockent tout ou une partie de l'information génétique d'un organisme. Chaque cellule d'un organisme multicellulaire contient au moins une copie de son génome. Les cellules d'un organisme multicellulaire remplissent de nombreuses fonctions différentes grâce à leur capacité à se différencier en différents types cellulaires. Chaque type cellulaire produit un ensemble différent de protéines.

La chromatine (Figure 1.1 panel 2 et 3) fait référence à la structure complexe qui aide les cellules à stocker l'ADN. Le génome humain, du début à la fin, mesure environ un mètre de long. La chromatine permet le compactage du génome humain pour en insérer une copie dans chacune des millions de cellules qui composent le corps humain. L'élément de base de la chromatine est le nucléosome, environ 146 paires de bases d'ADN enroulées environ deux fois autour d'un complexe constitué de protéines d'histones. Chaque histone a une courte queue polypeptidique qui peut être modifiée chimiquement pour modifier la force de l'interaction entre l'histone et l'ADN (Figure 1.1 panel 2). En plus de ses fonctions de compactage de l'ADN, la chromatine contrôle également la transcription car l'ADN est étroitement enroulé autour des nucléosomes (Figure 1.1 panel 3) ne peut pas interagir avec l'ARN polymérase II (Pol II). Cependant, les nucléosomes sont des structures dynamiques et leur interaction avec l'ADN peut être contrôlée. Les protéines appelées facteurs de transcription peuvent se lier à des séquences d'ADN qui s'enroulent normalement autour des nucléosomes, soit en déplaçant les nucléosomes, soit en les affaiblissant l'interaction avec l'ADN.

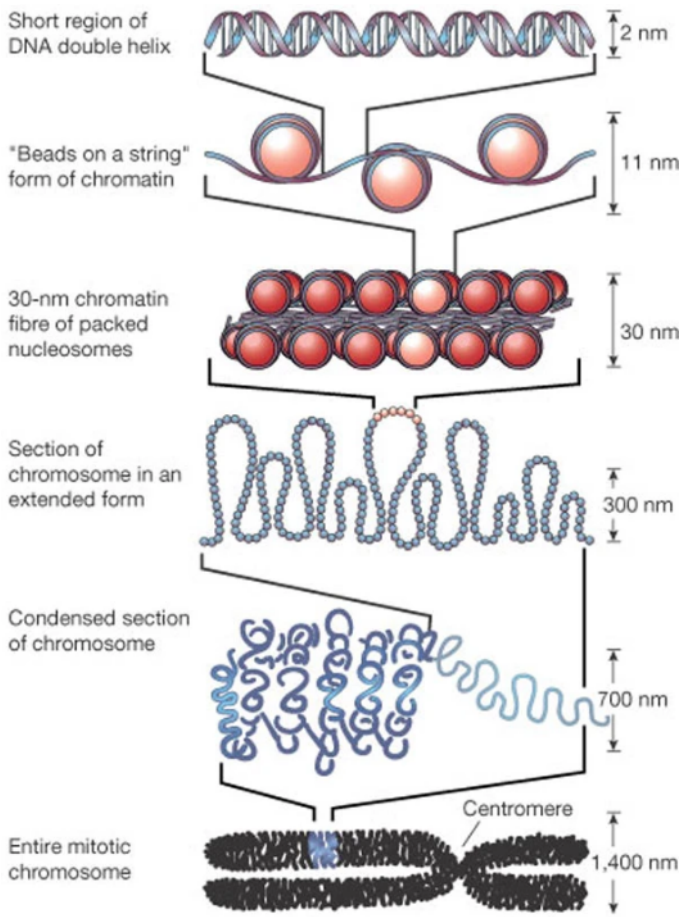


FIGURE 1.1. – ***Organisation de l'ADN dans la Structure Chromatine.*** *Le niveau d'organisation le plus bas est le nucléosome, qui est formé par l'enroulement de deux tours super-hélicaux d'ADN (165 paires de bases) autour d'un octamère d'histones. Les nucléosomes sont reliés par des sections courtes d'ADN de liaison. La chaîne de nucléosomes est pliée en une fibre d'environ 30 nm de diamètre à un niveau supérieur d'organisation et ces fibres sont encore pliées en structures de niveaux supérieures. Les détails de pliage au-delà du nucléosome sont incertains.* [1].

1. Felsenfeld, G., Groudine, M. Controlling the double helix. Nature 421, 448–453 (2003).

Le « dogme central » de la biologie moléculaire, qui n'est rétrospectivement considéré que comme une approximation grossière du processus par lequel les gènes sont utilisés pour produire des protéines, a été proposé pour la première fois par Francis Crick en 1958 [60], cinq ans après sa publication de la structure de l'ADN avec James Watson [306]. Cette théorie affirmait que les gènes de l'ADN codent pour l'ARN, qui à son tour code pour les protéines. Bien qu'il ne soit pas nécessairement incorrect dans l'interprétation la plus large, il a été révélé depuis [127] que ce processus implique de nombreuses autres étapes, y compris une grande quantité de détails non présents dans la théorie originale. Une vision plus moderne du processus par lequel l'ADN est utilisé pour produire la variété de produits géniques présents dans la cellule, ainsi que certains des moyens par lesquels il est régulé, est résumée ci-dessous, avec un accent particulier en rapport avec les travaux de ma thèse, sur la transcription, les éléments régulateurs, et les facteurs de transcription.

1.1.3. La transcription

La transcription est le processus biologique au cours duquel une molécule d'ARN est synthétisée à partir d'une molécule d'ADN. Il s'agit d'un processus nuancé, mais qui implique à sa résolution la plus grossière en trois étapes principales, appelées initiation, élongation et terminaison (Figure 1.2). Ces trois étapes seront détaillées dans les parties suivantes.

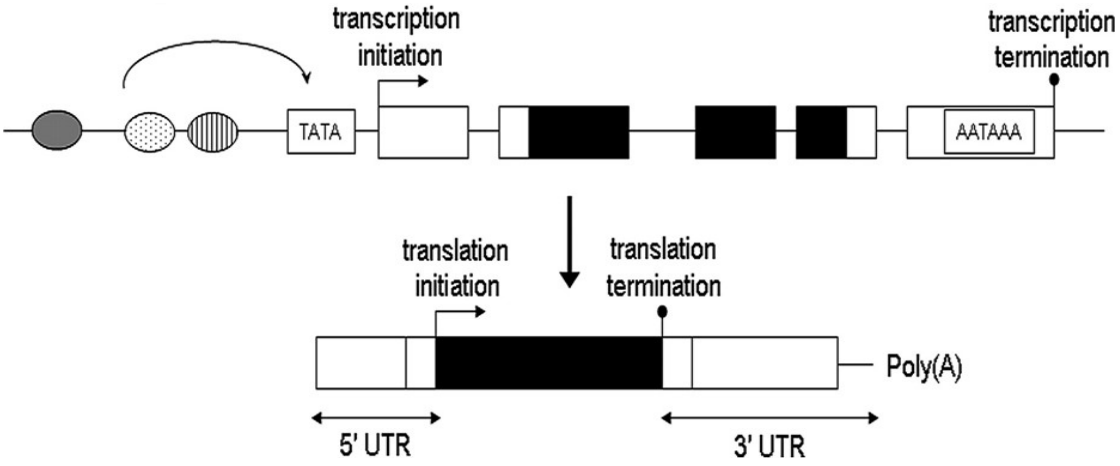


FIGURE 1.2. – ***Structures généralisées des gènes et des unités de transcription eucaryotes.*** *TATA désigne l'un des éléments promoteurs centraux eucaryotes possibles, auxquels les facteurs de transcription transmettent des informations, et poly(A) désigne l'ajout post-transcriptionnel d'une queue de poly(A). Les barres noires représentent l'ADN codant, les barres ouvertes représentent l'ADN transcrit mais non traduit, et les lignes fines dans les régions transcrites représentent les introns.*[2].

2. Lynch, Michael. "The origins of eukaryotic gene structure." Molecular biology and evolution 23.2 (2006).

1.1.3.1. L'initiation de la transcription

L'initiation est l'étape de la transcription au cours de laquelle la majorité de la régulation a lieu. Comme la transcription dans son ensemble, elle peut être décrite en trois étapes fondamentales. Le premier est la fixation d'une ARN polymérase à l'ADN dans la région promotrice du gène (en particulier, au site d'initiation de la transcription (TSS)). L'ARN polymérase est la molécule qui interprète l'ADN de la région codante du gène et transcrit l'ARN résultant. La région promotrice est une région d'ADN trouvée en amont du début de la région codante du gène. Pour faciliter cette fixation, le promoteur central du gène, situé à l'intérieur du promoteur, doit d'abord être fixé par un ensemble de protéines appelées facteurs de transcription généraux (GTF). Les GTF sont nécessaires pour recruter une ARN polymérase, formant (avec d'autres protéines et complexes) le complexe de préinitiation (PIC). Parallèlement aux GTF, les TF spécifiques aux gènes peuvent jouer un rôle clé dans l'initiation, via la fixation aux sites de fixation des facteurs de transcription (TFBS) [129]. Le reste de l'initiation est composé de l'ouverture, ou "fusion", de la double hélice d'ADN et le détachement de la polymérase des molécules du complexe de pré-initiation pour commencer à transcrire le gène et produire l'ARN messager (ARNm) résultant (Figure 1.3).

Lors d'une initiation réussie, la polymérase s'échappe de la région promotrice, se séparant des GTF, et passe à l'étape suivante de la transcription, connue sous le nom d'élongation, au cours de laquelle la molécule d'ARN produite par l'ARN polymérase est allongée d'un nucléotide à la fois jusqu'à ce qu'une copie complémentaire du gène entier soit construite. Enfin, le transcrit est clivé et une séquence d'adénines est placée à l'extrémité dans l'étape finale de la transcription, appelée terminaison.

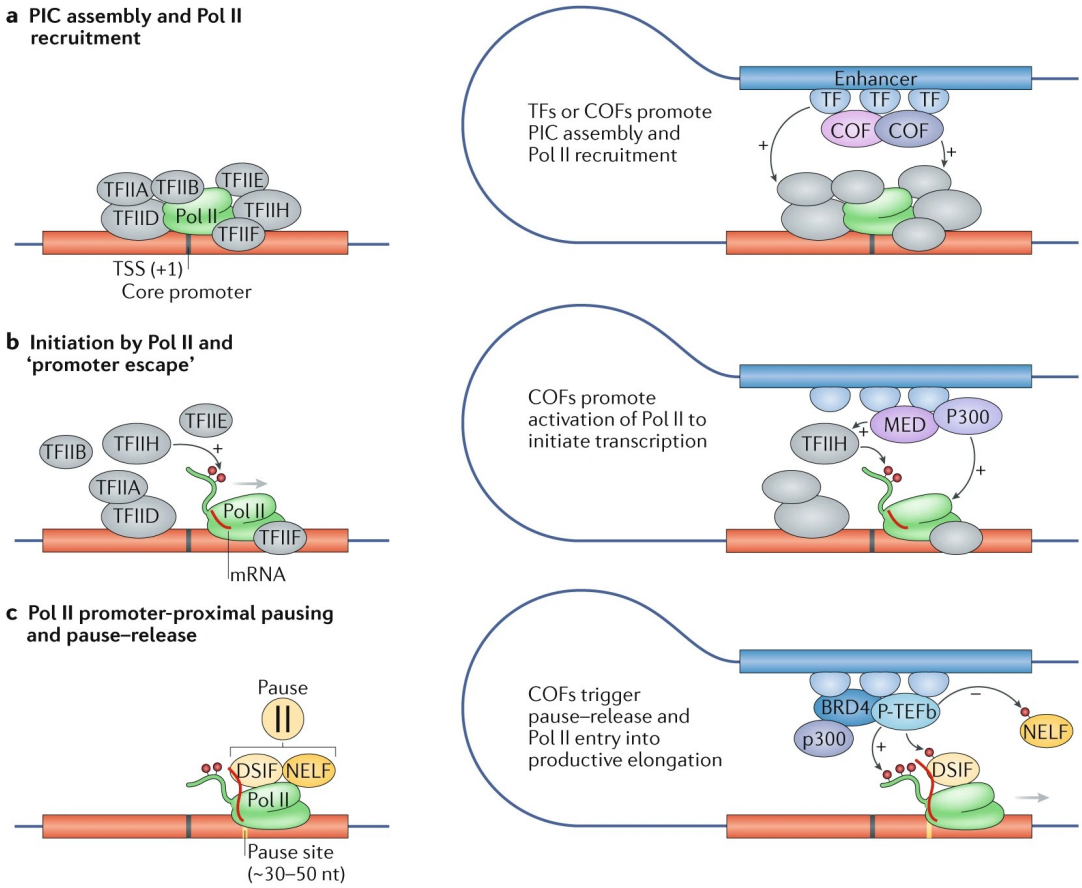


FIGURE 1.3. – ***Complexe de pré-initiation de la transcription.*** *a) Assemblage du PIC et recrutement de l'ARN polymérase II. La première étape de l'initiation de la transcription est l'assemblage du PIC, composé de la Pol II et de six TF généraux : le facteur de transcription IIA (TFIIA), TFIIB, TFIID, TFIIE, TFIIF et TFIIH (image de gauche). Les activateurs peuvent favoriser l'assemblage du PIC en recrutant des TF généraux et des cofacteurs qui interagissent directement avec les TF ou la Pol II (image de droite). b) Initiation par la Pol II et "échappement du promoteur". c) Pause de la Pol II en amont du promoteur et libération de la pause.*[3].

3. Haberle, Vanja, and Alexander Stark. "Eukaryotic core promoters and the functional basis of transcription initiation." Nature reviews Molecular cell biology 19.10 (2018).

1.1.3.2. L'élongation

Bien que la traduction de l'ARNm en protéine ou produit génique commence au codon d'initiation, le gène est transcrit à partir d'un emplacement en amont de ce site d'initiation de la traduction, le site d'initiation de la transcription (TSS). L'emplacement du TSS n'est pas facilement reconnu et il peut, en fait y en avoir plus d'un pour un gène donné. De même, la traduction se termine au codon stop mais la transcription se poursuit jusqu'au site de terminaison de la transcription (TTS). La séquence d'ADN du TSS au codon de départ et du codon d'arrêt au TTS sont respectivement désignées comme les régions non traduites 5' et 3' (5' et 3' UTR). De plus, le produit génique n'est souvent pas continu du début aux codons d'arrêt dans la séquence de pré-ARNm initialement transcrite. Les exons, les segments qui codent pour les acides aminés de la protéine, sont entrecoupés par longs introns non codants. Les deux UTR, ainsi que les introns, peuvent contenir des éléments régulateurs qui agissent pour influencer l'expression du gène correspondant ou d'autres gènes voisins. Ces éléments supplémentaires permettent une régulation fine de l'expression.

1.1.3.3. La terminaison

La terminaison de la transcription est différente pour les multiples polymérases. Contrairement aux procaryotes, l'élongation par l'ARN polymérase II chez les eucaryotes a lieu 1 000 à 2 000 nucléotides en aval de l'extrémité 3' du gène en cours de transcription. Cette queue de pré-ARNm est ensuite éliminée par clivage pendant la transcription de l'ARNm.

1.1.4. Les éléments régulateurs

1.1.4.1. Généralités

Comme mentionné brièvement ci-dessus, la régulation de la transcription se produit principalement lors de l'initiation de la transcription, et en grande partie via la présence (ou l'absence) de protéines de fixation à l'ADN appelées facteurs de transcription. En revanche, contrairement aux facteurs de transcription généraux qui sont indispensables à toute initiation, différentes associations de TF sont requises pour les différents promoteurs/enhancers (amplificateurs) et peuvent entraîner des résultats variés sur les niveaux de transcription. Les facteurs de transcription ont des préférences énergétiques pour (et contre) certaines séquences de nucléotides, et ces préférences aident à les guider vers leurs sites de fixation à l'ADN. Les TF peuvent se présenter sous forme de protéines uniques (ex : AP-1 et MEF-2) ou dans le cadre de complexes protéiques (ex : complexe d'activation du facteur nucléaire Kappa B (NF-κ B), qui comprend plusieurs protéines telles que p50 et p65), et se lient généralement à l'ADN sur un site de fixation spécifique (ex : CCAAT pour les NFY).

Les TF peuvent agir soit comme des activateurs, qui augmentent le taux de transcription, soit comme des répresseurs (silencers), qui réduisent le taux de transcription ou annulent complètement la transcription de certains gènes. Les activateurs augmentent la transcription via des mécanismes tels que la fixation directe à l'ADN et le recrutement et/ou l'activation ultérieurs de l'ARN polymérase, d'autres protéines pour activer la polymérase, la modification des protéines ou les changements de conformation de l'ADN. Les activateurs interagissent avec différentes protéines telles que le complexe médiateur, le complexe cohésine, les protéines ou les complexes de TF pour activer les gènes via des "interactions de longue portée". En outre, la figure 1.4 décrit différentes classes d'éléments *cis*-régulateurs dans le génome humain et leur distance relative par rapport au TSS des gènes codants pour les protéines. Les silencers diminuent la transcription via des mécanismes tels que l'intéraction en compétition thermodynamique avec les activateurs, l'encombrement stérique de la fixation de l'activateur et/ou de la polymérase, la destruction de l'activateur et le silencing génique via la modification de la chromatine. Les facteurs de transcription, étant le produit des gènes, sont eux aussi régulés, créant des boucles de rétroaction et augmentant considérablement la complexité du réseau de régulation des gènes.

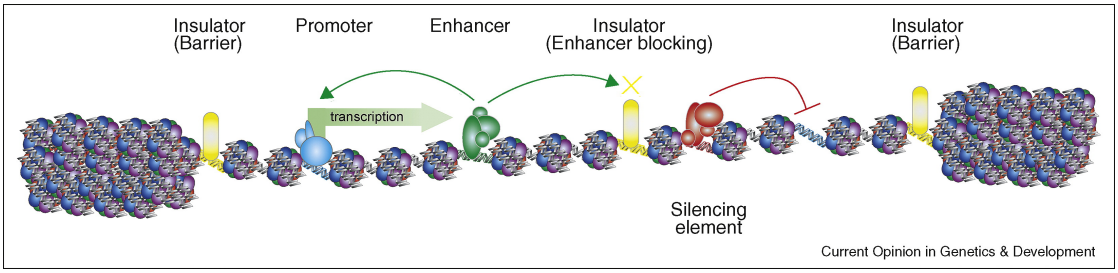


FIGURE 1.4. – ***Les différentes classes d'éléments cis-régulateurs dans le génome humain.*** *La transcription démarre au niveau des promoteurs (ADN bleu), qui sont activés par des enhancers (ADN vert) ou réprimés par des silencers (ADN rouge). L'activité des enhancers et des silencers peut être isolée par des insulators (ADN jaune), qui empêchent également la propagation de structures de chromatine répressive condensée (montrées à chaque extrémité de cette région chromosomique). Ce modèle représente les nucléosomes comme de l'ADN (hélice grise) enroulé autour de protéines d'histones (couleurs variées), qui sont moins denses au niveau des éléments régulateurs où se fixent les TF (ovales bleus), des protéines activatrices et répressives (ovales verts et rouges, respectivement) et CTCF (ovale jaune).*[4].

4. Heintzman, Nathaniel D., and Bing Ren. "Finding distal regulatory elements in the human genome." Current opinion in genetics and development 19.6 (2009).

Il existe donc plusieurs éléments impliqués dans la régulation des gènes appelées éléments *cis*-régulateurs. Dans les parties suivantes, nous détaillerons les cinq éléments principaux de la régulation : Les promoteurs, Les enhancers, Les silencers, les insulateurs et les facteurs de transcription.

1.1.4.2. Les promoteurs

Le promoteur fait généralement référence à une région d'ADN qui permet une initiation précise de la transcription d'un gène [268]. Le promoteur central est une étendue minimale de séquences d'ADN (par exemple, la boîte TATA, l'initiateur et l'élément promoteur central en aval) entourant le site de démarrage de la transcription qui interagit directement avec les composants de la machinerie de transcription basale, y compris l'ARN polymérase II. Bien que les séquences ou motifs d'ADN comprenant la région promotrice centrale des gènes individuels puissent être structurellement et fonctionnellement divers, nous pensons que son rôle universel est de conduire une initiation précise de la transcription [268]. Les facteurs de transcription qui se lient à environ 100 à 200 pb en amont du promoteur central peuvent augmenter le taux de transcription en facilitant le recrutement ou l'assemblage de la machinerie de transcription basale sur le promoteur central ou en favorisant le recrutement de séquences d'ADN régulatrices distales spécifiques au promoteur central.

Le promoteur d'un gène actif est généralement situé dans une région sans nucléosome flanquée de deux nucléosomes bien positionnés (nucléosomes +1 et -1). Cette région sans nucléosome rend le promoteur plus accessible et facilite l'assemblage de la machinerie de transcription. Ces nucléosomes associés à des promoteurs présentent des variants d'histone spécifiques et des modifications d'histone. En particulier, chez les mammifères, les deux nucléosomes adjacents sont enrichis en variants d'histone H3.3/H2A.Z [135]. Une autre caractéristique importante des promoteurs flanquant les nucléosomes est qu'ils sont marqués par des marques d'histones spécifiques, dont il a été démontré qu'elles sont en corrélation avec l'activité du promoteur. Alors que la marque H3K4me3 est associée à des promoteurs actifs, H3K27me3 est associée à des promoteurs réprimés [20, 170]. H3K27ac est une autre marque bien étudiée qui est également associée à des promoteurs actifs [113] (Figure 1.5). Chez les mammifères, la grande majorité des promoteurs chevauchent les îlots CpG [170]. L'avènement du séquençage à haut débit nous a permis de cartographier l'initiation de la transcription avec une sensibilité et une résolution sans précédent. Cela a révélé que les éléments *cis*-régulateurs sont généralement associés à des sites d'initiation transcriptionnelle flanquant les séquences régulatrices.

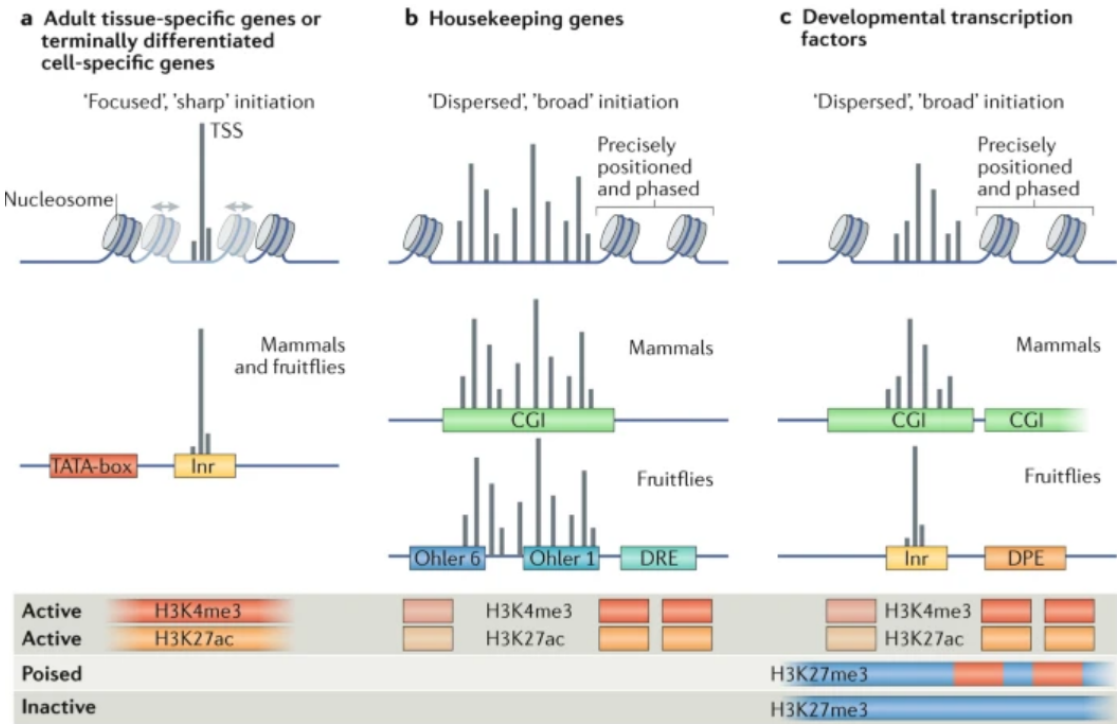


FIGURE 1.5. – ***Initiation de la transcription et types de promoteur central.*** *Trois types principaux de promoteurs ont été proposés chez les métazoaires en fonction de différentes propriétés telles que le motif d'initiation, la composition en séquences, la configuration de la chromatine et la fonction génique. a) Le premier type se caractérise par un motif TATA-box et Inr, une initiation précise et des nucléosomes imprécisément positionnés. b) Le deuxième type correspond aux gènes de ménage ("Housekeeping genes") qui sont associés à une initiation de transcription dispersée, une région sans nucléosomes définie flanquée de nucléosomes marqués par H3K4me3 et H3K27ac. c) Le troisième type concerne les promoteurs de facteurs de transcription clés dans le développement, qui sont marqués de manière bivalente avec H3K4me3 et H3K27me3, et entourés d'éléments non-codants conservés.*[5].

5. Haberle, Vanja, and Alexander Stark. "Eukaryotic core promoters and the functional basis of transcription initiation." Nature reviews Molecular cell biology 19.10 (2018).

Les Epromoteurs

Les Epromoteurs sont une nouvelle classe d'éléments *cis*-régulateurs définis comme un élément promoteur qui présente une activité enhancer et qui peut réguler l'expression d'un gène éloigné à travers une interaction promoteur-promoteur. Ces interactions ont été identifiées pour la première fois par des méthodes basées sur la chromatine, suggérant que ce réseau régulateur est courant dans les cellules mammifères [221]. Des Epromoteurs ont été identifiés dans certaines études in vitro chez la drosophile [321],la souris [210] et l'homme[250], et certains ont validé leur activité fonctionnelle in vivo [65]. (Figure 1.6).

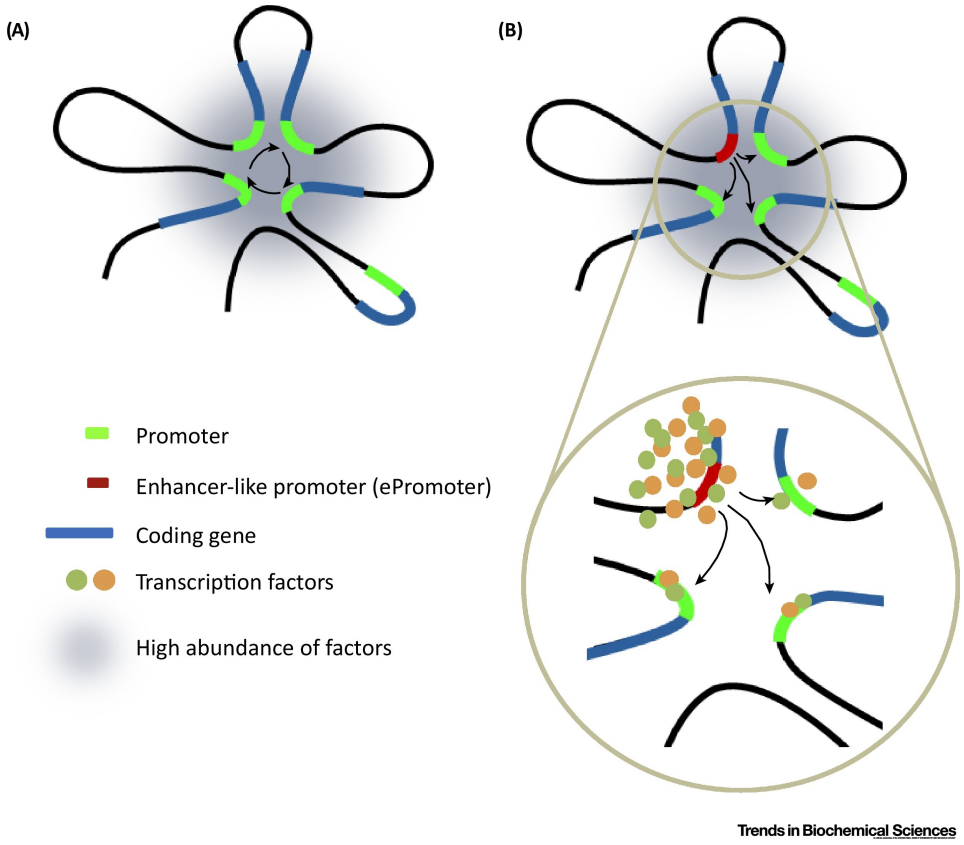


FIGURE 1.6. – ***Modèle d'intéraction Epromoteurs et de la régulation génique.*** *(A) Les interactions chromosomiques placent les promoteurs en étroite proximité physique (usines de transcription), facilitant le recrutement des facteurs de transcription et de la Pol II nécessaires à la transcription de leurs gènes associés. (B) La présence d'un Epromoteur à l'intérieur de l'usine de transcription pourrait favoriser le recrutement de niveaux élevés de facteurs de transcription et de la Pol II.* [6].

6. Medina-Rivera, Alejandra, et al. "Widespread enhancer activity from core promoters." Trends in biochemical sciences 43.6 (2018).

1.1.4.3. Les Amplificateurs/"Enhancers"

Tentative de définition

Fournir une définition précise des enhancers n'est pas une tâche facile car ils peuvent avoir des rôles différents selon l'état cellulaire. En effet, ils peuvent être actifs ou inactifs, ou peuvent assumer une fonction de non-enhancers. Leur mécanisme fonctionnel tel que dérivé des procédures expérimentales n'est pas encore parfaitement connue [229]. Pour expliquer en détail, on peut voir sur la Figure 1.7 (A-D) une illustration de différents modèles d'interaction des éléments régulateurs proximaux et distaux reliés par la boucle de la chromatine.

Les enhancers peuvent être défini comme les acteurs des régions régulatrices de l'ADN qui augmentent la production transcriptionnelle dans les cellules. Typiquement, les enhancers présentent les propriétés suivantes : Les enhancers peuvent résider en amont ou en aval du TSS de leurs gènes cibles ou ils peuvent même être sur différents chromosomes par rapport à leurs cibles [229]. Ils peuvent jouer un rôle clé dans l'expression des gènes spécifiques aux tissus ainsi que manifester des propriétés distinctes à travers différents tissus, organes et conditions cellulaires. Un enhancer peut initier la transcription de l'ARN polymérase II, produisant une nouvelle classe d'ARN non codants, non épissés et non polyadénylés, appelés eRNAs [153]. Dans les parties suivantes, nous verrons plus en détail la définition et le fonctionnement des enhancers.

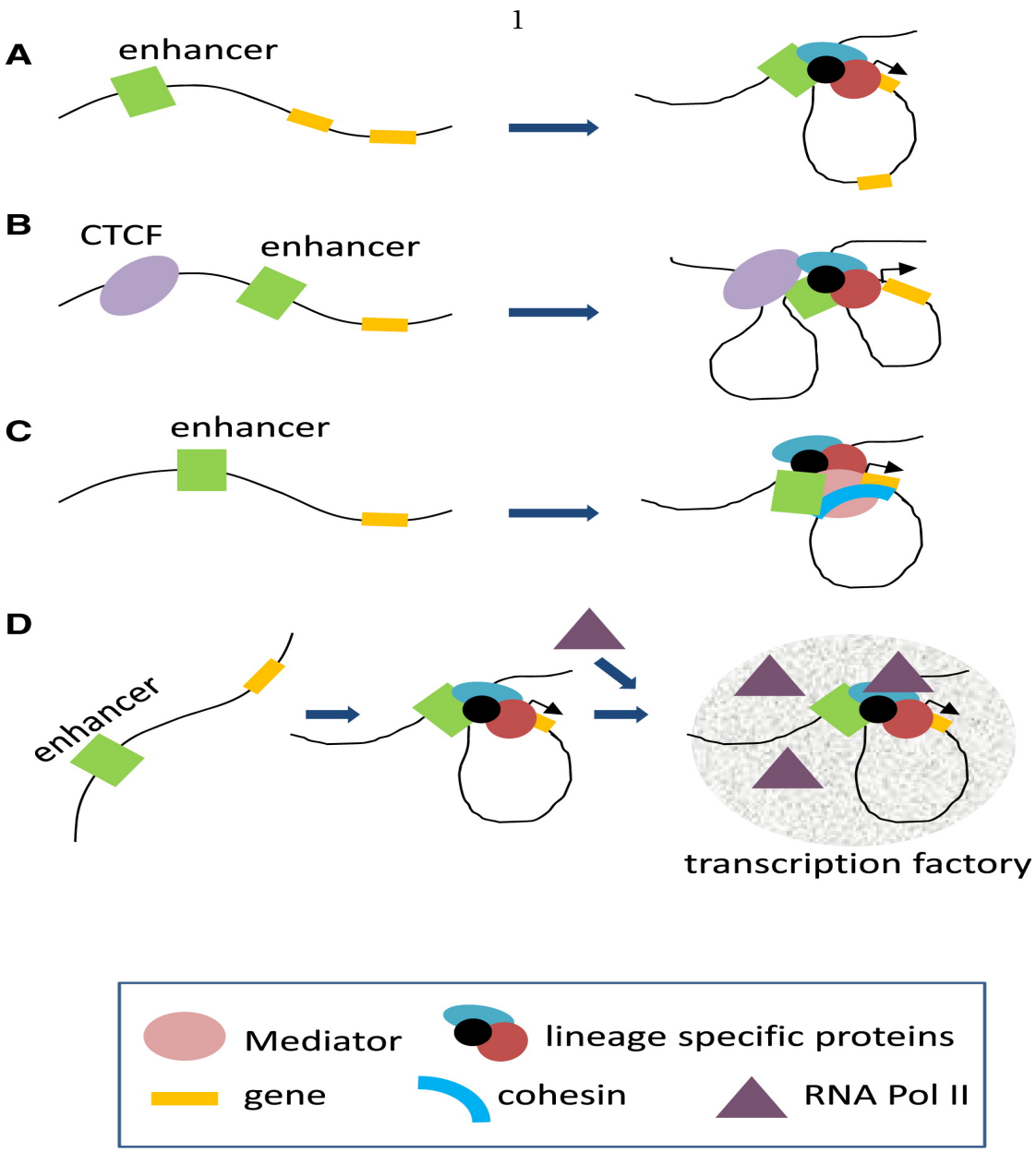


FIGURE 1.7. – ***Les Enhancers et les Promoteurs interagissent par des boucles de chromatine.*** *(A) Les TF favorisent l'interaction à longue distance entre l'enhancer et l'un des gènes par interaction protéique homotypique et/ou hétérotypique.(B) Le gène est activé par des activateurs qui cooptent CTCF en interaction à longue distance avec le gène. (C) L'enhancer est relié au promoteur du gène par le médiateur et la cohésine avec la participation de TF activant le gène. (D) La boucle enhancer-gène est représentée comme étant médiée par des activateurs avant l'accumulation de la Pol II et l'apparition d'une usine de transcription*[7].

7. Plank, Jennifer L., and Ann Dean. "Enhancer function : mechanistic and genome-wide insights come together." Molecular cell 55.1 (2014).

Découverte des enhancers

En 1981, le terme "enhancer" a été inventé suite à l'observation d'un élément d'ADN du virus simien (SV40) pouvant conduire à l'expression génique exogène d'un gène rapporteur β-globine et antigène T de lapin cloné [18]. Il avait été précédemment observé que la transcription des gènes précoces du SV40 dépendait d'une séquence répétée en tandem de 72 paires de bases (pb) à environ 200 pb en amont du site d'initiation de la transcription [25, 111]. Banerji et al. ont ensuite montré que cette répétition en tandem de 72 pb pouvait améliorer l'expression du gène de la β-globine dans les cellules HeLa [18]. Cet « enhancer viral » augmenterait la transcription à de très grandes distances du site d'initiation de la transcription et fonctionnait indépendamment de son orientation, à la fois en amont et en aval du promoteur de la β-globine.

Des travaux ultérieurs ont identifié de nouveaux éléments enhancers dans plusieurs autres virus animaux partageant des propriétés similaires [252, 275]. Ces analyses ont mesuré l'impact des enhancers sur l'expression de gènes rapporteurs. Il est rapidement devenu évident que l'activité enhancer dans ces systèmes de gènes rapporteurs dépendait en partie de la spécificité de la cellule hôte. Par exemple, les enhancers identifiés dans le papillomavirus de bovin et le sarcoma virus de murin ont montré une préférence marquée pour les cellules bovines et les fibroblastes de souris respectivement. Ces résultats suggèrent que les facteurs de régulation de la cellule hôte déterminent la spécificité de l'activité de l'enhancer envers le type cellulaire.

Peu de temps après la découverte des enhancer viraux, des séquences similaires ont été identifiées dans les génomes de mammifères, initialement dans le locus de la chaîne lourde des immunoglobulines (IgH). Avec la découverte d'éléments enhancer dans les virus, ces régions préservées de l'ADN ont été identifiées comme des éléments enhancer aux propriétés similaires [17]. Après cette découverte, plusieurs enhancer endogènes ont été identifiés et caractérisés sur la base de l'impact fonctionnel sur l'expression du gène cible [42, 199]. Collectivement, ces travaux ont établi les enhancers en tant qu'éléments fonctionnellement distincts des promoteurs et capables d'activer l'expression d'un gène cible. Cette activation est possible à partir de très longues distances et fonctionne indépendamment de l'orientation. Cela a permis de déterminer les propriétés de base des éléments enhancer afin de comprendre leur rôle dans la transcription.

Notre compréhension actuelle de l'activité de l'enhancer est fondée sur le rôle critique des facteurs de transcription au niveau des éléments enhancer. Il est devenu de plus en plus clair que les TF se lient aux éléments enhancer et recrutent d'autres facteurs pour contrôler l'expression du gène cible. De plus, les régions enhancers se trouvent généralement dans les régions de la chromatine ouverte accessible aux TF et permettent une conformation favorable de la chromatine (boucle) pour l'activité enhancer.

Fixation des facteurs de transcription sur les enhancers

Peu de temps après la découverte des enhancers, il a été montré que des facteurs de transcription se lient spécifiquement à de courtes séquences d'ADN appelées motifs de fixation [169]. En 1987 Lee et al. ont montré que le facteur de transcription AP1 se lie à un enhancer du promoteur de la métallothionéine humaine ainsi qu'à l'enhancer SV40, augmentant ainsi de manière significative l'expression génique [169]. Ces découvertes ont établi l'importance de la spécificité de la séquence du motif de fixation des facteurs de transcription au sein de ces éléments *cis*-régulateurs afin d'activer l'enhancer [276].

Les facteurs de transcription se fixent de manière générale à de petites séquences d'ADN dégénérées longues de 6 à 12 paires de bases avec une spécificité de séquence relativement faible grâce à des domaines de fixation. Les enhancer contiennent des clusters de plusieurs sites de fixation de facteurs de transcription différents pour activer et réprimer les TF (Figure 1.8). La présence de plusieurs motifs de fixations de facteurs de transcription suggère que l'activité de l'enhancer est contrôlée par l'interaction de plusieurs facteurs de transcription plutôt que par la simple affinité d'un seul facteur de transcription pour son motif de fixation respectif. En effet, il existe plusieurs exemples dans lesquels la fixation des TF aux enhancer permet un contrôle précis et dynamique de la transcription. Par exemple, chez *Drosophila melanogaster*, la spécificité de la voie RAS est déterminée en partie par la fixation d'un complexe facteurs de transcription (Mad, dTCF, Tinman, Twist) à un enhancer qui contrôle le facteur de transcription Even skipped (Eve) [115]. Chez l'humain, le TF p53 se fixe également sur une région enhancer de gènes cibles suppresseurs de tumeur [198].

Les TF interagissent également directement et indirectement les uns avec les autres pour faciliter davantage l'activité. Les mécanismes de coopérativité des TF dépendent de l'emplacement des motifs de fixation au sein de l'enhancer. Les positions des motifs de fixation du facteur de transcription dans un enhancer donné dictent l'arrangement spatial de la fixation du TF et donc son potentiel d'interaction. L'importance de l'arrangement des motifs de fixation du facteur de transcription d'un enhancer est illustrée par l'enhancer bien caractérisé de l'interféron-β (IFN-β). L'IFN-β est un exemple merveilleusement complexe du chevauchement spécifique des motifs de fixation des TF. Cette fixation d'un complexe de TF permet d'entraîner une activité enhancer fonctionnelle [223]. L'organisation précise de l'enhancer IFN-β illustre également des cas de fixation coopérative indirecte entre facteurs de transcription. Dans certains cas, les facteurs de transcription peuvent fixer leurs motifs respectifs au sein d'un enhancer et recruter un cofacteur (Figure 1.8). La fixation du TF au niveau des enhancers tire parti de l'organisation de leurs motifs de fixation pour stabiliser l'affinité avec la région enhancer.

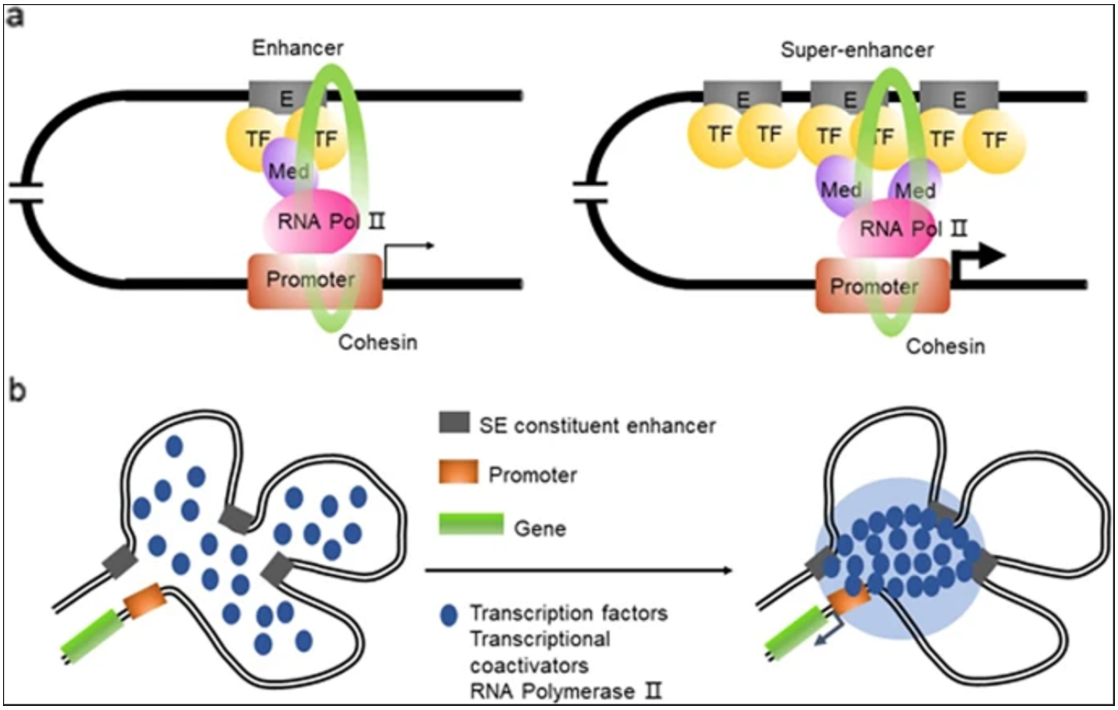


FIGURE 1.8. – *a) Les Enhancers et super-enhancers (SEs) sont occupés par une forte densité de régulateurs transcriptionnels, y compris des TF, des coactivateurs et le complexe de la Pol II. b) Un modèle de séparation de phase pour l'activation des SEs. Des interactions à haute densité entre les régulateurs transcriptionnels forment des complexes multimoléculaires séparés par phase au locus SE, conduisant à la transcription des gènes dirigés par SE.* [8].

Il existe d'autres formes de coopérativité entre TF comme le recrutement de TF dans des complexes de remodelage de la chromatine pour déclencher le repositionnement des nucléosomes et découvrir des sites de fixation au sein d'un enhancer. Par exemple, dans la lignée épithéliale mammaire murine, le facteur de transcription AP1 se lie à son motif d'ADN et recrute des remodeleurs de chromatine pour exposer le site de fixation du TF collaborateur GR. Ces facteurs de transcription, appelés facteurs pionniers, ont été identifiés la première fois chez la levure et ont depuis été caractérisés à plusieurs reprises [8]. Il est important de noter que les facteurs pionniers seuls ne sont généralement pas suffisants pour former un complexe enhancer au niveau des enhancer. Cela suggère que l'activité enhancer, avec l'action de plusieurs facteurs de transcription, permet un contrôle dynamique de l'expression du gène cible.

8. Jia, Qunying, et al. "Oncogenic super-enhancer formation in tumorigenesis and its molecular mechanisms." Experimental and Molecular Medicine 52.5 (2020).

Structure de la chromatine au niveau des enhancers

Pour que les facteurs de transcription se lient aux enhancers et médient l'activité des enhancers, leurs motifs de fixation doivent être accessibles. En effet, les enhancers actifs se trouvent généralement dans les régions de chromatine ouverte dépourvues de nucléosomes et accessibles aux facteurs de transcription. Ainsi, la structure de la chromatine fournit une couche supplémentaire de régulation de l'activité de l'enhancer en déterminant la disponibilité des régions régulatrices de l'enhancer pour les facteurs de transcription. Il est devenu de plus en plus clair que l'état de la chromatine change de manière fluide au cours du développement et en réponse à des stimuli externes [202]. De plus, les états de la chromatine reflètent également les programmes d'expression génique particuliers associés à différents types cellulaires [81].

L'unité fondamentale de la chromatine est le nucléosome : un octamère d'histone contenant deux copies de quatre protéines histones (H3, H4, H2A, H2B) dans lesquelles 147 paires de bases d'ADN sont enroulées. Le positionnement global de nucléosomes est le principal déterminant de l'accessibilité de l'ADN. Un aspect clé régulant le positionnement des nucléosomes et donc l'état global de la chromatine est la présence de modifications covalentes sur les queues d'histones. Il existe plusieurs types de modifications, notamment l'acétylation, la méthylation, la phosphorylation, l'ubiquitylation et la sumoylation. Chacune de ces modifications chimiques covalentes porte une charge chimique (Figure 1.9) qui peut influencer les intéractions nucléosomiques et ainsi modifier la structure globale de la chromatine. Par exemple, l'acétylation des résidus de lysine de queue d'histone entraîne la perte de charge positive et donc le relâchement de l'ADN lié à la chromatine [3]. Ces diverses modifications de queue d'histone jouent un rôle essentiel dans la conformation de la chromatine et ont été associées à plusieurs processus, notamment la transcription, la réplication de l'ADN et la réparation de l'ADN. Des mesures à l'échelle du génome de ces modifications d'histones ont révélé des modèles particuliers d'enrichissement qui correspondent à des régions actives de la chromatine (promoteurs et enhancer) ou à des régions silencieuses de la chromatine.

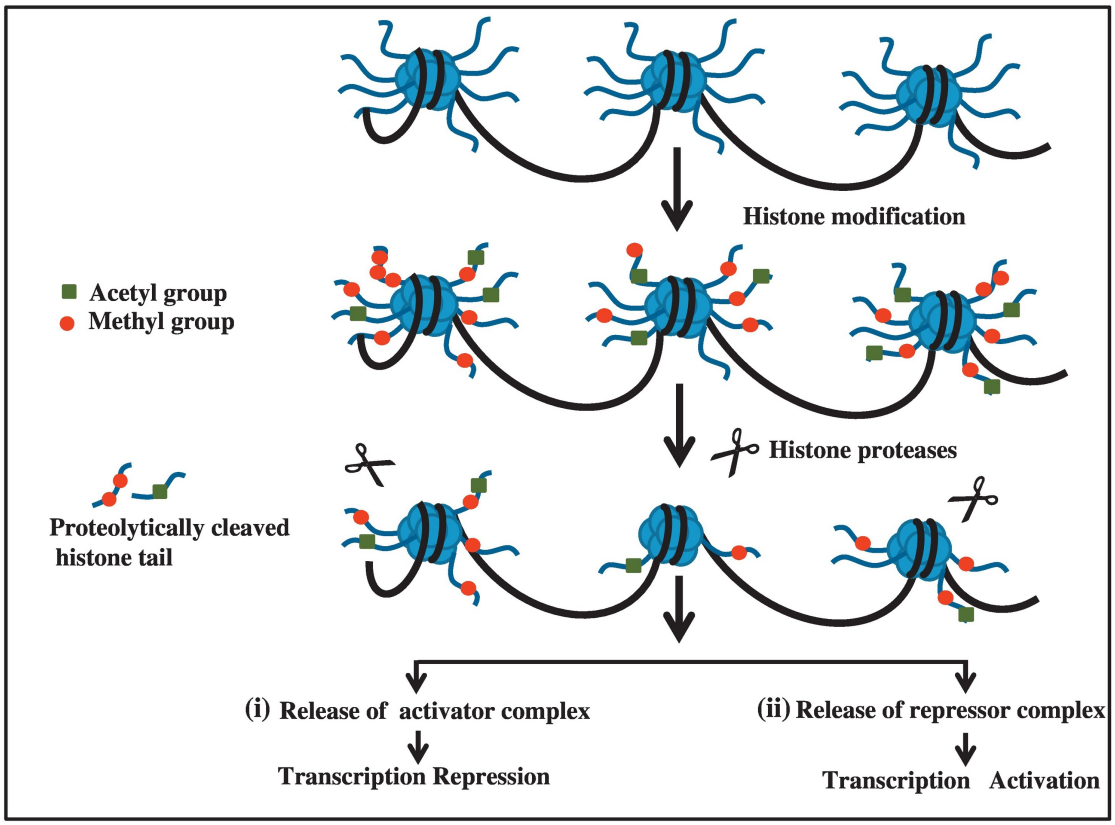


FIGURE 1.9. – ***L'effet du clivage de la queue des histones sur la transcription.*** *Le modèle proposé démontre que le clivage de la queue des histones peut affecter la transcription de deux manières. (i) Si un complexe activateur de transcription est lié à un nucléosome avec une queue d'histone, alors il favorisera la transcription. Dans ce cas, le clivage de la queue d'histone entraînera la libération du complexe activateur du nucléosome, ce qui entraînera une répression de la transcription. (ii) Si un complexe répresseur de transcription est lié à un nucléosome avec l'aide de la queue d'histone, alors il réprime la transcription. Dans ce cas, le clivage entraînera la libération du complexe répresseur du nucléosome, ce qui entraînera une activation de la transcription. Par conséquent, le clivage d'histones peut activer ou réprimer la transcription.*[9].

9. Azad, Gajendra Kumar, et al. "Modifying chromatin by histone tail clipping." Journal of molecular biology 430.18 (2018).

En plus des effets sur les intéractions nucléosomiques, les modifications sur les queues d'histones servent également à recruter des facteurs de chromatine avec une activité enzymatique pour modifier davantage la chromatine. Les facteurs de chromatine appelés « writer » déposent des modifications d'histone, tandis que les facteurs de chromatine appelés « eraser » suppriment ces modifications. À leur tour, les protéines « reader » de la chromatine contiennent des domaines de liaison qui reconnaissent chacune de ces marques d'histones spécifiques. Collectivement, la combinaison des modifications d'histones telles que régulées par ces trois classes de protéines (writer, eraser et reader) forment ce que l'on a appelé le « code des histones ». Ce code histone influence l'état de la chromatine et contribue à un ensemble de mécanismes intervenant dans l'expression génique d'une cellule. Par exemple, l'acétylation des résidus de lysine sur l'histone H3 joue un rôle critique dans la régulation des éléments enhancer. L'acétylation des histones a été particulièrement bien étudiée et est fondamentalement liée à l'ouverture de la chromatine et donc à l'activation des gènes [3]. Les histones acétyltransférases (HAT) déposent des groupes acétyle, les histones désacétylases (HDAC) éliminent ces groupes acétyle et les protéines contenant du bromodomaine reconnaissent les groupes acétyle. Il est également important de noter que la mono-méthylation marque également les enhancers et peut caractériser davantage l'état (actif, prêt ou amorcé) d'un élément *cis*-régulateur donné [120].

En conclusion, le dépôt, l'élimination et le recrutement de modifications d'histones jouent un rôle essentiel dans l'activité de l'enhancer. Pour que les facteurs de transcription se lient à leurs sites de fixation, la chromatine doit être accessible. Ainsi, les TF et les modifications d'histones servent de base à la régulation de l'activité de l'enhancer et travaillent pour dicter l'activité de l'enhancer.

Boucles chromosomiques entre enhancers et promoteurs

Comme cela a été décrit précédemment, les facteurs de transcription se lient aux enhancers afin d'activer la transcription des gènes cibles. La transcription repose sur le recrutement et l'allongement de la Pol II à partir du site d'initiation de la transcription. Étant donné que les enhancers peuvent souvent être situés à de grandes distances du promoteur de leur gène cible, on pense que les enhancer sont amenés à proximité de leurs promoteurs centraux cibles par formation d'une boucle [75, 94]). Ce mécanisme entraîne une interaction spatiale pour permettre aux enhancers d'activer leurs gènes cibles (Figure 1.10 section 1.1.4.5). En plus de recruter des cofacteurs de la chromatine, les facteurs de transcription se lient également aux composants du complexe médiator et contribuent ainsi à la création de boucles chromosomiques. Par exemple, dans les cellules souches embryonnaires, le facteur de transcription ELK1 se lie à une sous-unité Sur2 du complexe médiator afin d'activer l'expression génique à la signalisation ERK [279].

En plus du rôle central que jouent les facteurs de transcription dans la boucle, les cofacteurs de la chromatine tels que BRD4 (bromodomain contenant la protéine, 4) qui se lient aux lysines acétylées se lient également au complexe médiator pour aider à la formation de la boucle promoteur-enhancer [315]. Le développement de technologies pour mesurer la conformation tridimensionnelle (3D) des chromosomes a contribué à une meilleure compréhension de la formation d'une boucle promoteur-enhancer. En 2009, une technique connue sous le nom de Hi-C a été introduite [175]. Cette technique a été adaptée à partir de la technique de capture de conformation de la chromatine (3C) [68] pour une application à l'échelle du génome. De plus, cette boucle formée entre les enhancers et les promoteurs centraux est encore stabilisée par des facteurs clés, notamment le complexe cohésine et les facteurs de charge associés tels que NIPBL (Nipped-B). Le complexe cohésine est de forme cylindrique et encercle deux nucléosomes pour aider à stabiliser l'interaction promoteur-enhancer [75]. Ainsi, les facteurs de transcription, les facteurs de chromatine ("eraser") et les modifications d'histone contribuent tous collectivement à l'interaction physique promoteur-enhancer pour permettre à la Pol II d'effectuer une transcription active.

Identification et caractérisation des super enhancers

La capacité de faire des prédictions à l'échelle du génome des éléments enhancers a abouti à l'identification d'un grand nombre d'éléments *cis*-régulateurs dans un type cellulaire donné. Par exemple, les prédictions d'enhancer basées sur les modifications d'histones peuvent identifier de l'ordre de centaines de milliers d'éléments *cis*-régulateurs putatifs [176]. Compte tenu de la nature prédictive de cette approche, il est hautement improbable que tous ces enhancers soient fonctionnels. Ces dernières années, il est devenu de plus en plus clair que les approches adoptées pour détecter davantage d'éléments *cis*-régulateurs peuvent donner des informations clés sur le contrôle transcriptionnel éxercé par l'enhancer [12]. Par exemple, les éléments *cis*-régulateurs ont été davantage caractérisés en grandes régions (contenant souvent plusieurs enhancers) qui activent l'expression de gènes liés, appelées régions de contrôle de locus (LCR) [271]. De même, l'identification et la caractérisation d'un sous-ensemble d'enhancers avec des niveaux exceptionnellement élevés de facteurs d'enhancer, appelés Super Enhancers (SEs) fournissent une classe d'enhancers avec des fonctions particulièrement importantes [181, 310]

1.1.4.4. Les silencers

Une compréhension de la répression transcriptionnelle est également essentielle pour une compréhension complète de la structure du promoteur et de la régulation de l'expression des gènes. La répression transcriptionnelle chez les eucaryotes est obtenue grâce à des "silencers", qui sont des éléments spécifiques à une séquence qui induisent un effet négatif sur la transcription de son gène particulier [218]. Il existe deux principaux types de silencieux, à savoir les « éléments silencers » et les « éléments de régulation négative ». Les éléments silencers sont des éléments classiques indépendants de la position qui dirigent un mécanisme de répression actif, et les éléments de régulation négative sont des éléments dépendants de la position qui dirigent un mécanisme de répression passif. Les silencers font partie intégrante de nombreux locus eucaryotes. La connaissance de leur rôle interactif avec les promoteurs et les enhancers, ainsi qu'avec d'autres éléments transcriptionnels, est essentielle pour notre compréhension de la régulation des gènes chez les eucaryotes. Plusieurs travaux récents permettent de les identifier [290, 250].

1.1.4.5. Les insulateurs

L'organisation des domaines chromosomiques et la restriction des enhancers nécessitent d'établir des limites. En sachant que les gènes pourraient être activés de manière non spécifique par des enhancers à proximité ou réprimés par la compaction de l'hétérochromatine, le génome semble avoir développé certains éléments d'ADN pour contrer ces effets, ces éléments sont appelés insulateurs.

Un insulateur est défini comme un élément de séquence d'ADN qui a la capacité de protéger un gène de son environnement chromosomique. Il effectue deux activités isolantes : une barrière à l'hétérochromatine et un blocage des enhancers. Deux expériences de gènes rapporteurs ont été utilisées pour identifier le rôle des insulateurs [22]. Ils agissent également comme "barrière à l'hétérochromatine" en protégeant un gène contre l'effet de la répression chromosomique, en empêchant l'extension de la formation d'hétérochromatine à partir de régions adjacentes[19].

Une deuxième activité des insulateurs est le blocage des enhancers, où l'activité de l'insulateur interrompt la communication entre un promoteur et un enhancer spécifiquement lorsqu'il est positionné entre les deux. Au contraire ce type d'insulateur n'interfère pas avec l'interaction entre le promoteur et l'enhancer lorsque l'insulateur est situé de chaque côté de la paire promoteur-enhancer.

Le complexe CTCF/Cohesin

CTCF est la protéine architecturale du génome se liant aux insulateurs la mieux caractérisée chez les mammifères. Grâce à ses 11 domaines en doigts de zinc, CTCF se lie de manière directionnelle et dynamique à des dizaines de milliers de sites de fixation. Le CTCF, associé à la cohésine, un complexe en forme d'anneau englobant l'ADN, assure la médiation des interactions de la chromatine à longue portée à l'échelle du génome (Figure 1.10). Les sites de fixation au CTCF se situant aux limites des domaines de chromatine peuvent fonctionner comme des isolants pour bloquer l'extrusion de la boucle de cohésine [94].

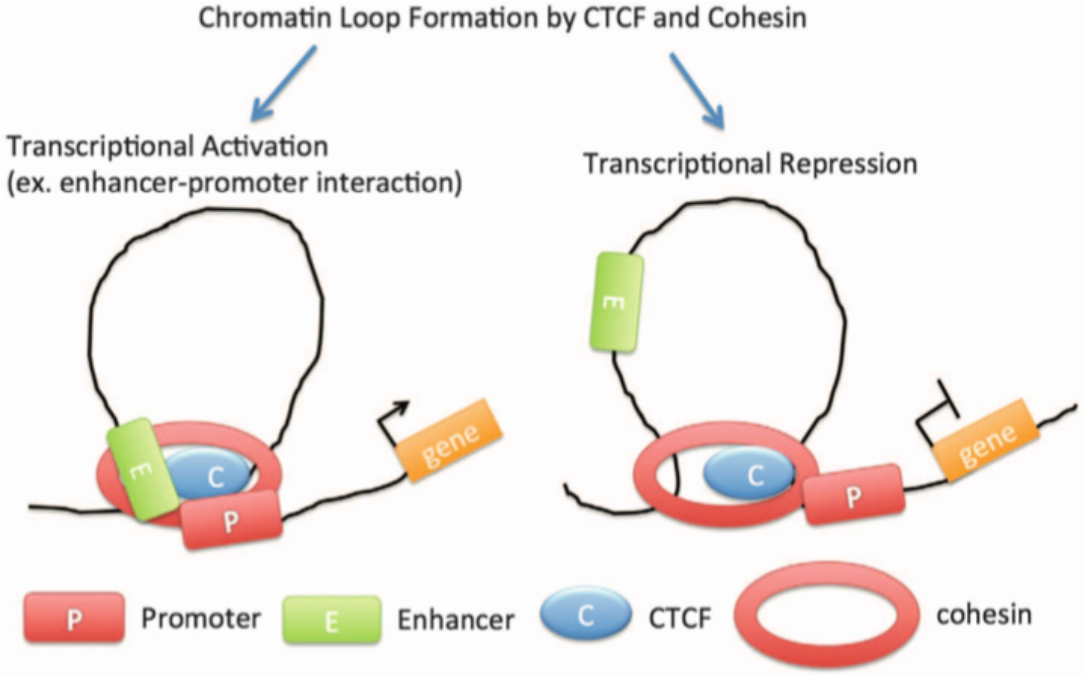


FIGURE 1.10. – ***CTCF et le complexe de cohésine peuvent entraîner une activation ou une répression transcriptionnelle.*** *Par exemple, lorsque CTCF et la cohésine se fixent à leur TFBS et créent une boucle de chromatine qui englobe l'enhancer et le promoteur, une activation transcriptionnelle se produit. En revanche, si CTCF et la cohésine forment une boucle de chromatine qui empêche l'enhancer d'atteindre le promoteur, l'expression génique est réprimée.* [10].

10. Kim, Somi, Nam-Kyung Yu, and Bong-Kiun Kaang. "CTCF as a multifunctional protein in genome regulation and gene expression." Experimental and molecular medicine 47.6 (2015).

1.1.4.6. Les facteurs de transcription

Les facteurs de transcription sont des protéines qui peuvent lier des séquences d'ADN spécifiques via leurs domaines de fixation à l'ADN. Il est bien établi que les facteurs de transcription jouent un rôle crucial dans la détermination du type cellulaire [285] et qu'une grande partie de tous les TF sont en fait exprimés dans la plupart des types cellulaires [294].

Caractéristiques basiques des TF

En général, il existe deux types de TF, les TF généraux et les TF spécifiques. Les TF généraux (GTF) sont un groupe de protéines responsables de la reconnaissance des sites spécifiques au sein des promoteurs centraux des gènes transcrits et du recrutement de l'ARN polymérase II pour former le complexe de pré-initiation pour l'initiation de la transcription. Chez les eucaryotes, les TF généraux comprennent six membres, à savoir TFIIA, TFIIB, TFIID, TFIIE, TFIIF et TFIIH, qui travaillent ensemble pour recruter l'ARN polymérase II pour initier la transcription. Les TF spécifiques, sont capables d'interagir avec l'ADN via leurs domaines de fixation à l'ADN avec une spécificité élevée. Ils se fixent généralement à des régions régulatrices spécifiques (y compris des promoteurs et des enhancers) du génome pour activer (ou réprimer) la transcription de gènes particuliers dans différents types cellulaires. Par rapport aux GTF, les TF spécifiques représentent une grande majorité des TF et possèdent des spécificités dû aux domaines protéiques. De plus, étant donné que les GTF sont universellement exprimés dans tous les types cellulaires alors que les TF spécifiques sont plus restreints dans leur expression, il est donc plausible que des TF spécifiques jouent un rôle plus important dans la détermination de l'identité cellulaire.

Les facteurs de transcription interagissent de manière non covalente avec l'ADN, principalement par des liaisons hydrogène et des forces de Van der Waals. Le contact entre les TF et l'ADN contient à la fois des interactions spécifiques et non spécifiques. Les interactions spécifiques se produisent entre les résidus d'acides aminés spécifiques du TF et les bases nucléotidiques dans les sites de fixation des séquences d'ADN ; les interactions non spécifiques se produisent généralement entre le TF et le squelette phosphate de l'ADN.

Les spécificités de liaison de plus de 1000 TF de plus de 131 espèces ont été déterminées et classées en 54 groupes en fonction de leurs domaines de liaison à l'ADN (Figure 1.11), ce qui indique que de nombreux TF partagent des domaines de fixation à l'ADN de structure similaire. De plus, les domaines de fixation TF sont conservés entre espèces même distantes ; par exemple, presque tous les TF de Drosophila partagent des domaines fixations similaires à leurs orthologues mammifères [214].

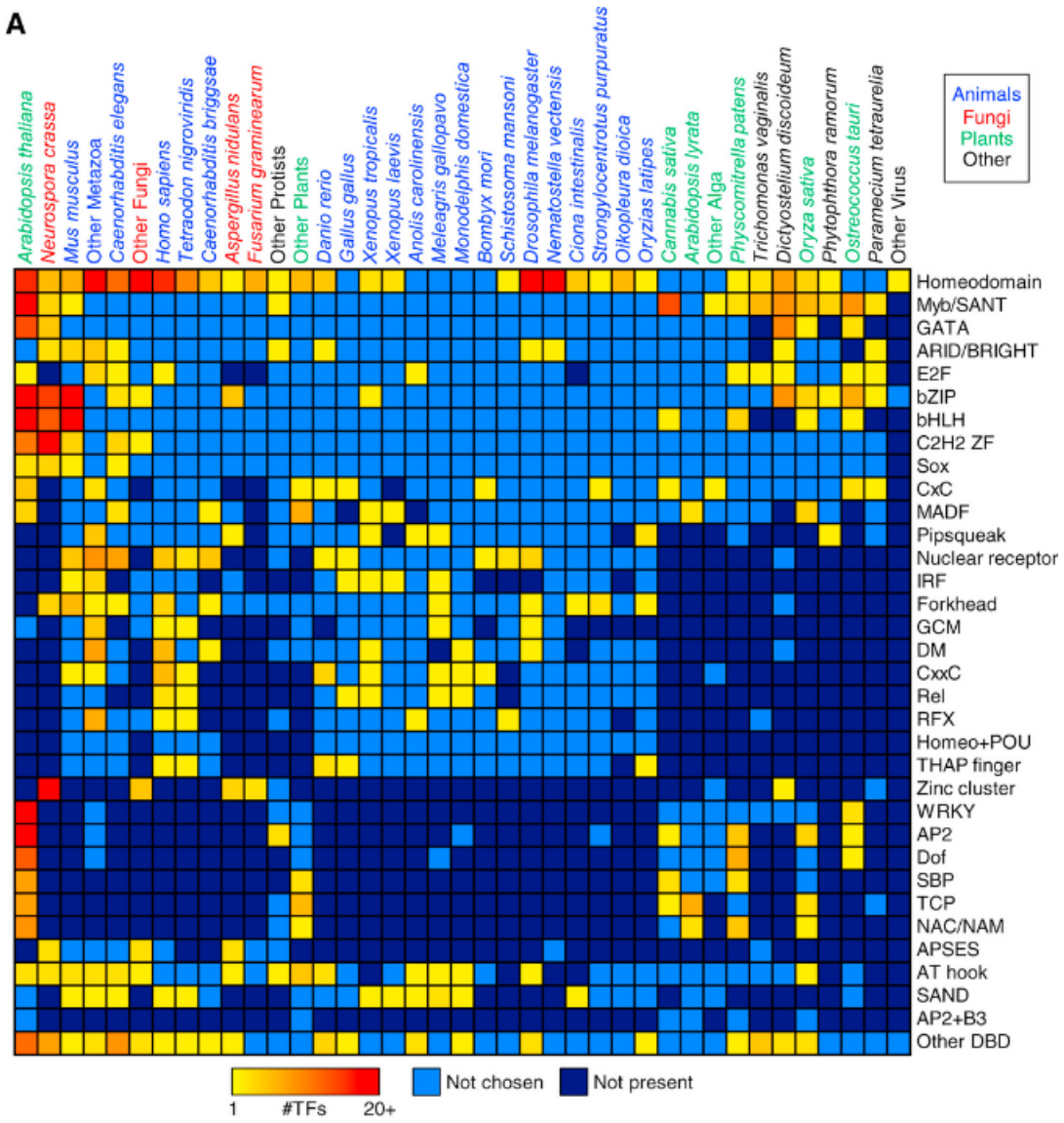


FIGURE 1.11. – ***omaine de fixation à l'ADN des TF.*** *(A) Les facteurs de transcription sont présentés par espèce et par classe de domaine de fixation à l'ADN. Les TF ayant plusieurs classes de domaine de fixation à l'ADN sont indiqués par un "+" (par exemple, AP2+B3). Les classes de domaine de fixation à l'ADN et les espèces contenant moins de cinq membres sont regroupées sous "Autres". Les espèces sont classées selon le nombre total de TF avec des motifs caractérisés.* [11].

11. Weirauch, M. T. et al. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell 158, 1431-1443, doi :10.1016/j.cell.2014.08.009 (2014).

En plus d'interagir avec les séquences d'ADN elles-mêmes, un TF peut également interagir avec une séquence d'ADN particulière en combinaison avec d'autres TF, soit par des interactions protéine-protéine, soit par des changements conformationnels de l'ADN. Les TF sont responsables du recrutement d'autres types de régulateurs transcriptionnels ainsi que de l'ARN polymérase, conduisant à la répression ou à l'activation de l'expression du gène cible. La spécificité cellulaire des TF est donc essentielles pour déterminer l'identité cellulaire. Ainsi les perturbations de l'expression des TF spécifiques peuvent altérer l'identité cellulaire. Par exemple, l'épigénome et le transcriptome du fibroblaste de souris peuvent être reprogrammés pour être identique à celui des cellules souches pluripotentes induites (iPSC) [285]. Ce processus est possible grâce à la surexpression d'un petit ensemble de TF (Figure 1.12) . Les fibroblastes peuvent également être directement différentiés en cellules musculaires en introduisant le TF MyoD [67]. Par conséquent, l'introduction d'un petit ensemble de TF dans une cellule est suffisante pour modifier son identité. Ce qui indique qu'une base composée d'un petit ensemble de TF est responsable de la spécificité du réseau de régulation de la transcription de la cellule.

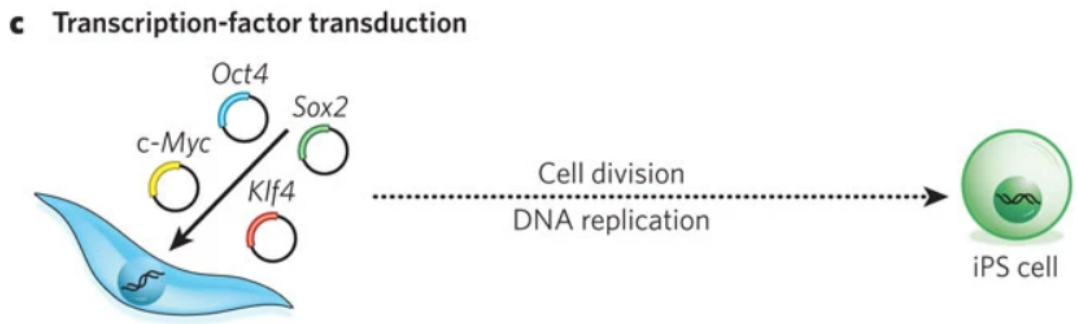


FIGURE 1.12. – ***Transduction de facteurs de transcription.*** *Cette approche peut être utilisée pour former des cellules souches pluripotentes induites, qui ont des propriétés similaires aux cellules souches embryonnaires et peuvent être générées à partir de presque tous les types de cellules du corps grâce à l'introduction de quatre gènes (Oct4, Sox2, Klf4 et c-Myc) à l'aide de rétrovirus. L'état pluripotent est maintenu de manière héréditaire et un grand nombre de cellules peuvent être générées, ce qui rend cette approche avantageuse pour les applications cliniques.* [12].

12. Yamanaka, Shinya, and Helen M. Blau. "Nuclear reprogramming to a pluripotent state by three approaches." Nature 465.7299 (2010).

Les sites de fixation au facteurs de transcription

En raison de leur capacité à modifier la chromatine et à contrôler la transcription, les TF jouent un rôle central dans la régulation de la transcription des gènes. Pour que les TF contrôlent la transcription de gènes spécifiques, ils doivent se lier à des régions spécifiques du génome. Cette spécificité est possible car certains TF possèdent un domaine de fixation à l'ADN ayant une structure tridimensionnelle possédant une affinité pour des séquences spécifiques d'ADN (Figure 1.13) [183]. Ces séquences d'ADN sont courtes (6-20pb chez les eucaryotes) et dégénérées et peuvent être représentées par une séquence consensus ou un ensemble de séquences consensus. Cet ensemble est appelé le motif de fixation, ou bien TFBS. Les approches permettant d'identifier les motifs de fixation des TF sera discuté dans la partie 1.3.4.

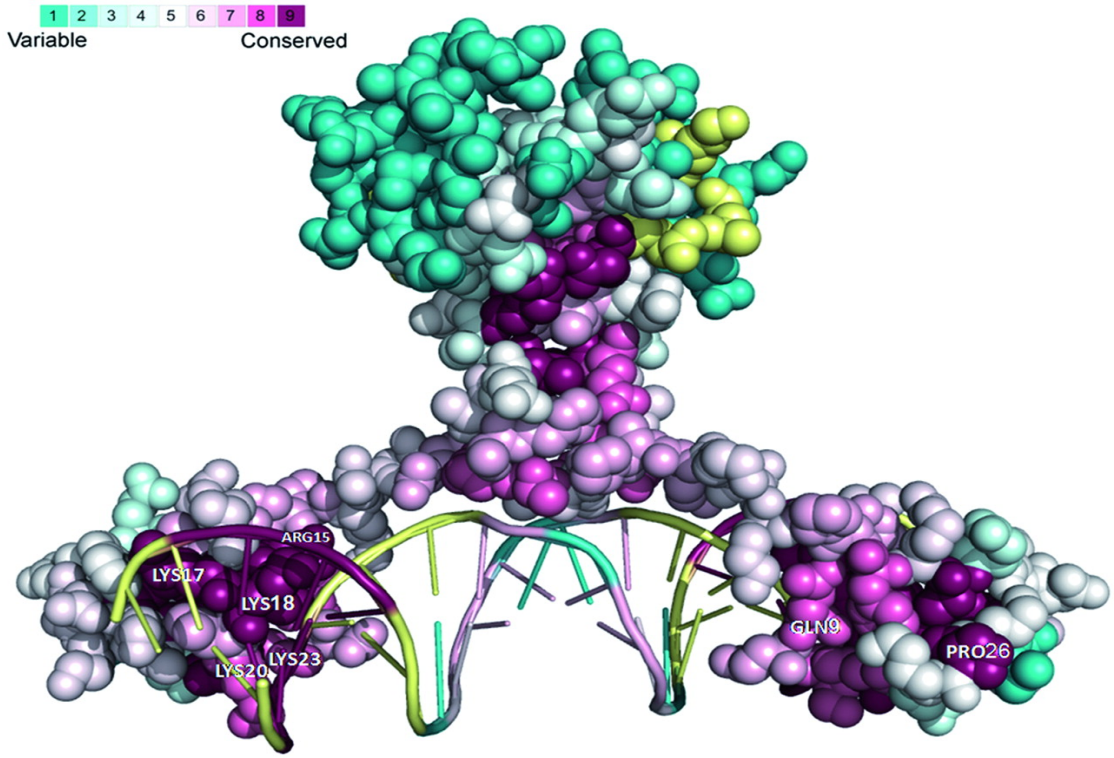


FIGURE 1.13. – ***Une analyse ConSurf pour le facteur de transcription GAL4 et son site de fixation à l'ADN.*** *La région N-terminale du TF GAL4 chez la levure fixée à l'ADN est présentée. Les acides aminés et les nucléotides sont colorés en fonction de leur degré de conservation en utilisant la barre de codage couleur, avec du turquoise au pourpre indiquant des degrés de conservation variables. Les positions pour lesquelles le niveau de conservation inféré avec un niveau faible de confiance sont marquées en jaune clair. La figure révèle que les régions fonctionnellement importantes à la fois sur l'ADN et la protéine sont très conservées.* [13].

13. Ashkenazy, Haim, et al. "ConSurf 2010 : calculating evolutionary conservation in sequence and structure of proteins and nucleic acids." Nucleic acids research 38.suppl_2 (2010)

Fonctionnalité des TFBS

Des analyses à l'échelle génomique de l'occupation des TF ont rapporté quelques centaines à plusieurs dizaines de milliers de sites de fixations des TF le long des génomes eucaryotes [53, 296]. Une étude in vitro combinant l'inactivation du TF et le profiling de l'expression génique a révélé une faible corrélation entre le TFBS et les gènes exprimés de manière différentielle lors de l'inactivation du TF [63]. Par ailleurs, la présence de TFBS peut expliquer qu'une fraction modeste de la conservation de l'activité des enhancers [301]. Selon ces observations, la majeure partie du TFBS serait non fonctionnelle puisque seuls quelques sites ont un impact sur l'expression des gènes. Plusieurs études évolutives ont tenté de déchiffrer la fonction des TFBS. En effet, si la fixation d'un TF donné à un certain endroit est fonctionnellement pertinente, il est raisonnable de supposer que le site de fixation du TF est sous une forte contrainte évolutive. Une comparaison de deux TF spécifiques au foie CEPBα et HNF4α sur cinq vertébrés séparés jusqu'à 80 millions d'années a montré une forte divergence dans leurs sites de fixation bien que le motif soit hautement conservé [254].

La plupart des écarts entre les TFBS liés aux espèces pourraient s'expliquer par la séquence sous-jacente, puisque 60 à 85% des pertes de liaison étaient liées à des événements de substitution, d'insertion ou de délétion. Cependant, 40 à 50% des TFBS perdus ont été compensés par un autre événement de liaison (binding event) dans les 10 kb. Des observations similaires ont été observées lors de la comparaison de l'occupation de trois TF d'hépatocytes sur cinq espèces de souris séparées jusqu'à 20 millions d'années. Plus les espèces sont éloignées, plus la divergence des sites de fixation du TF est élevée [277]. Néanmoins, les régions liées par plusieurs TF sont plus conservées parmi les espèces de rongeurs, ce qui suggère une pression de sélection plus forte pour les éléments d'ADN avec une affinité avec les TF plus élevée. En revanche, la drosophile présente un niveau de conservation généralement plus élevé dans leur TFBS par rapport aux mammifères [224]. Bien que cela puisse être corrélé avec une fréquence plus élevée d'éléments conservés dans les génomes de Drosophila (37-53 % contre 3-8 % pour les mammifères) [302], de manière similaire aux mammifères, les régions avec un TFBS en commun et situé proches de gènes sont globalement mieux conservées au cours de l'évolution. Ces travaux suggèrent que la régulation coopérative des TF spécifiques aux tissus et la codépendance de leur fixation décrivent la fonctionnalité du TFBS. En revanche, de nombreux TFBS semblent plutôt non fonctionnels et ont tendance à évoluer rapidement.

D'autre part, des analyses approfondies sur un nombre plus élevé de TF ont en effet révélé que les TF ont généralement tendance à se fixer sous formes de clusters dans les régions cibles à occupation élevée (High Occupancy Target (HOT)) [204]. Les régions HOT semblent être un bon indicateur de la fonctionnalité du TFBS car elles sont corrélées avec les locus interactants, le recrutement de la Pol II et l'expression spatio-temporelle des gènes. En revanche, les régions cibles à faible taux d'occupation ont tendance à être plutôt non fonctionnelles [276].

Complexes de TF

Les complexes TF sont formés par l'interaction de plusieurs régulateurs transcriptionnelle, chacun ayant une fonction spécifique. Les protéines qui composent ces complexes comprennent les TF eux-mêmes, ainsi que d'autres protéines tel que des co-facteurs qui les aident à se lier à l'ADN ou à recruter d'autres protéines pour former des complexes plus larges. Les cofacteurs TF ne se lient pas nécessairement à l'ADN et sont donc recrutés par le biais d'interactions protéine-protéine. Une telle spécificité latente a également été caractérisée dans la spécification des cellules hématopoïétiques. Le cofacteur FOG-1 non lié à l'ADN interagit avec GATA-1 pour réguler positivement la différenciation des mégakaryocytes et des érythrocytes [292]. Les complexes de TF sont également régulés par des modifications post-traductionnelles (Figure 1.14). Ces modifications peuvent affecter la stabilité, la localisation et l'activité des complexes des facteurs de transcription, modifiant ainsi leur capacité à réguler la transcription des gènes.

Enfin, les complexes des facteurs de transcription peuvent interagir avec d'autres complexes protéiques, tels que les complexes de remodelage de la chromatine, pour modifier l'accessibilité de l'ADN et réguler ainsi la transcription des gènes. Par exemple, MyoD coopère avec l'homéobox Pbx1 pour recruter le complexe de remodelage SWI/SNF au niveau du promoteur de Myog pour augmenter la fixation du TF et induire la transcription du gène [28]. De même, Pax7 interagit avec le complexe histone méthyltransférase Wdr5-Ash2L-MLL2 au voisinage de Myf5. Son recrutement induit la triméthylation de H3K4 suivie de l'expression transcriptionnelle de Myf5 [196].

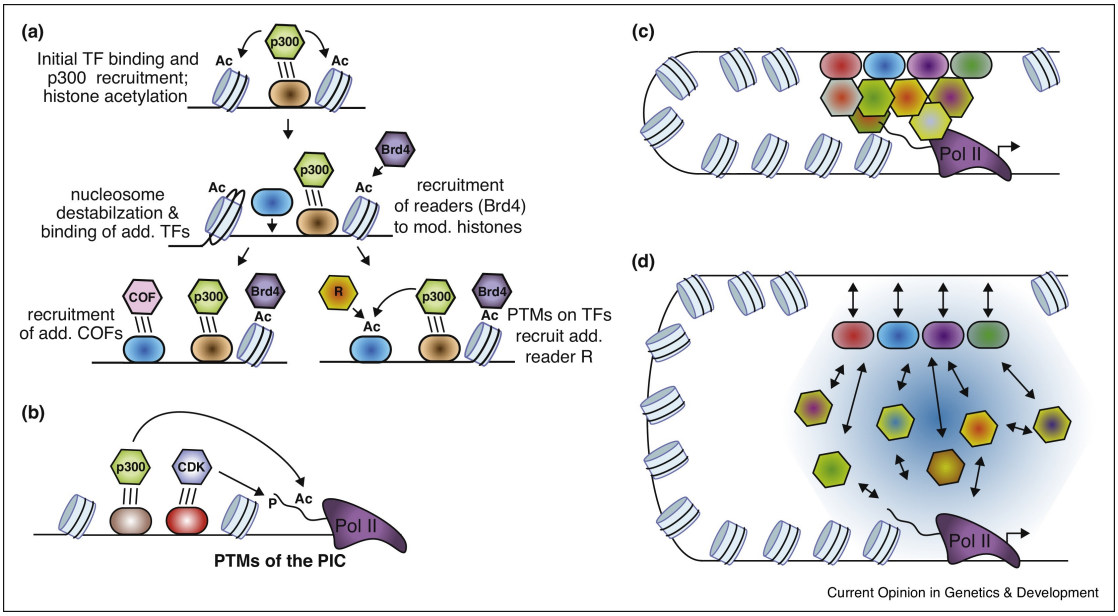


FIGURE 1.14. – ***Interaction des co-facteurs et recrutement de la Pol II.*** *(a) La coopérativité au niveau des enhancers peut résulter de l'activité de cofacteurs tels que p300 et les modifications post-traductionnelles qu'ils catalysent. (b) Les facteurs de transcription recrutent différents COFs tels que p300 et CDK qui peuvent modifier post-traductionnellement la Pol II et d'autres protéines liées au promoteur central pour réguler la transcription. (c) Modèle de régulation transcriptionnelle dans lequel les TF, les COFs et les protéines des promoteurs centraux forment des complexes protéiques relativement stables. (d) Modèle plus souple dans lequel les TF et les cofacteurs interagissent les uns avec les autres et avec le complexe de pre-initiation de manière plus dynamique, via des interactions protéine-protéine transitoires, des concentrations locales accrues et des PTMs.* [14].

14. Reiter, Franziska, Sebastian Wienerroither, and Alexander Stark. "Combinatorial function of transcription factors and cofactors." Current opinion in genetics and development 43 (2017).

1.1.4.7. Topologically Associating Domains (TADs)

Contexte

Dans le contexte de ma thèse, j'ai utilisé des données de Hi-C, pour l'analyse de la régulation de l'ALDH1A1 (chapitre 4), afin de délimiter les bordures de TADs autour de ce gène. Nous allons donc pésenter dans ce chapitre l'état des connaissances scientifiques sur les TADs.

Définition

Au début des découvertes liées aux Hi-C (techniques permettant de détecter les régions du génome interagissant ensemble, voir chapitre 1.2.4.3), les chercheurs ont découvert que le génome pouvait être séparé en deux compartiments A et B, ou les régions contenues dans un compartiment interagissent préférentiellement (régions de A entre elles et régions de B entre elles) [175]. Les TAD sont des structures plus petites que les compartiments qui ont été largement explorées et bien caractérisées. Les limites de TAD sont dynamiques mais se sont avérées en grande partie conservées entre cellules individuelles [207], à travers les espèces et les types cellulaires [70]. Certaines de ces frontières peuvent être partagées par des compartiments, alors qu'il s'agit de structures distinctes. Contrairement aux TAD, le niveau de conservation des compartiments est discutable, car leurs limites semblent similaires à travers les lignées cellulaires, mais montrent des variations autour de gènes clés quant à l'activation et la répression de ces gènes [255]. En raison de leur faible niveau de conservation, par rapport aux TAD, les compartiments ont été suggérés comme une entité statistique reflétant des contacts préférentiels entre les TAD, plutôt que comme des entités physiques.

Protéine de structure des TADs

De nombreuses protéines nucléaires aident à stabiliser les différentes structures de la chromatine telles que les TAD et les boucles de chromatine. En effet, on pense que la formation de TAD est pilotée par un modèle d'extrusion en boucle : la chromatine est tirée à travers un facteur d'extrusion en boucle (complexe protéique en forme d'anneau) jusqu'au éléments aux limites du TAD [94] (Figure 1.15). Le principal facteur d'extrusion de boucle impliqué dans la formation de TAD est la cohésine, et deux protéines CTCF positionnées à chaque extrémité du TAD servent généralement de frontière. Les TAD ont donc également été définis comme des domaines d'interaction formés par extrusion de boucle et délimités par des protéines architecturales[94].

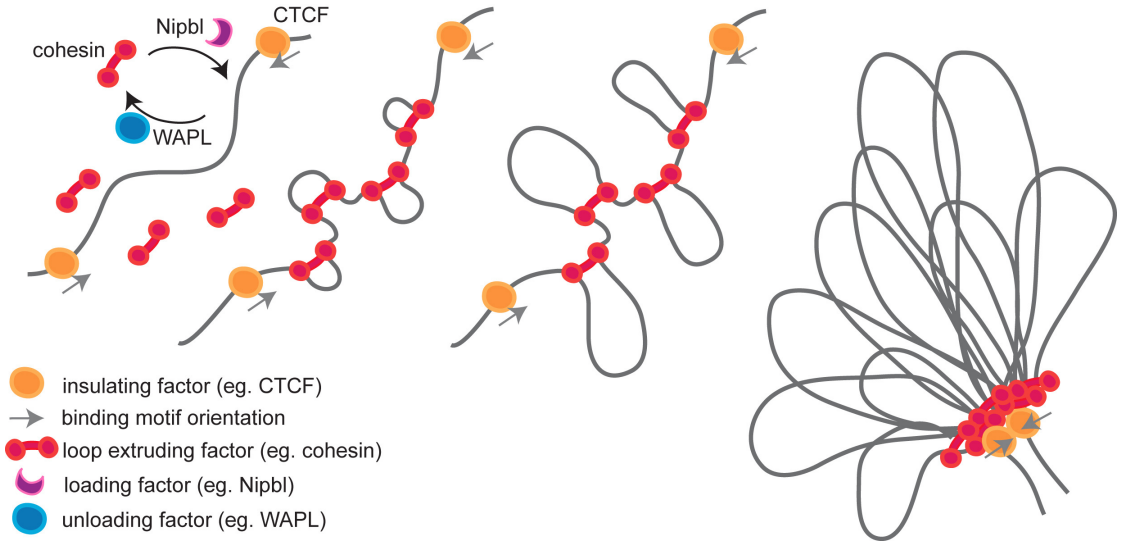


FIGURE 1.15. – ***Extrusion de boucle comme modèle pour la formation de TAD.*** *La cohésine est chargée et déchargée en continu de la chromatine et extrude de manière bidirectionnelle les boucles tout en étant engagée sur la chromatine. Les éléments formant un obstacle à l'extrusion, tels les sites fixés au CTCF, deviennent une barrière plus stable pour le TAD extrudé.* [15].

15. Sikorska, Natalia and Sexton, Tom. (2019). Defining Functionally Relevant Spatial Chromatin Domains : It is a TAD Complicated. Journal of Molecular Biology.

1.2. Méthode de séquençage à haut débit

1.2.1. Contexte

Dans ce chapitre, nous présentons différentes méthodes de séquençage qui ont été utilisées tout au long de ma thèse pour explorer différents aspects de la régulation génique. Nous décrirons notamment la technique de séquençage ChIP-seq, qui permet d'identifier les sites d'interaction entre les TF et l'ADN. Nous décrirons également d'autres méthodes tel que ATAC-seq, DNase-seq et Hi-C qui ont été utilisées pour la construction d'une cartographie multi-omique de la région régulatrice de l'ALDH1A1.

1.2.2. Le séquençage à ADN

La première méthode de séquençage de l'ADN a été développée par Ray Wu en 1971 [314] puis optimisée par Frederick Sanger en 1975 [249]. Cette technique, connue sous le nom de séquençage de Sanger, tire parti du système de réplication de l'ADN, qui est utilisé lors de la division cellulaire. Cette méthode nécessite quatre échantillons différents contenant chacun : un seul brin d'ADN à séquencer, une amorce (une courte séquence d'ADN, généralement d'une douzaine de bases, qui correspond au début du brin d'ADN cible), une ADN polymérase et des nucléotides avec chacune des bases (ATCG), toutes sont nécessaires à la réplication de l'ADN. En plus de cela, des nucléotides d'ADN modifiés qui ne contiennent pas les propriétés chimiques (groupe 3'-OH) à lier par des nucléotides supplémentaires (terminateurs) sont ajoutés à chaque mélange, mais uniquement des nucléotides avec un type particulier de base par mélange. Chaque mélange crée des répliques de longueur variable qui se terminent toujours par une base spécifique. La comparaison des longueurs de tous les différents fragments créés (électrophorèse sur gel) permet de déterminer la séquence exacte du brin initial d'ADN. Cette technique donne des résultats de séquençage de haute qualité et peut être utilisée pour séquencer l'ADN jusqu'à environ 800 paires de bases de long. Bien que cette méthode soit relativement ancienne, coûteuse (par paire de base séquencée) et chronophage, elle est encore largement utilisée aujourd'hui dans des contextes nécessitant de longues reads de séquençage. Le séquençage Sanger est également utilisé dans de nombreux projets qui ne nécessitent pas de grandes quantités d'ADN à séquencer. La plupart des méthodes de séquençage d'ADN reposent sur le même principe sous-jacent et utilisent l'ADN polymérase. Pour cette raison, le séquençage de l'ADN se produit généralement des fragments d'ADN de l'extrémité 5' à l'extrémité 3'.

1.2.3. Les techniques NGS (Next Generation Sequencing)

Les NGS (nouvelle génération de séquençage) font référence aux technologies modernes de séquençage à haut débit qui permettent le séquençage de grandes quantités d'ADN en peu de temps. Plusieurs méthodes ont été développées et sont décrites ci-dessous. Elles ont chacune des propriétés différentes ayant des avantages et des inconvénients. Cependant, ces dernières années, une technologie particulière, appelée séquençage Illumina (du nom de la société qui a acheté la technologie "Solexa"), qui a été développée à Cambridge au milieu des années 1990 par Shankar Balasubramanian et David Klenerman, a complètement dominé le domaine des NGS [27]. Le séquençage Illumina [201] repose sur le même principe sous-jacent que le séquençage Sanger. Cependant, cette technologie permet désormais de fixer différents colorants fluorescents sur les différentes bases et, par conséquent, différents mélanges ne sont plus nécessaires. Un autre élément nouveau est que les terminateurs sont temporaires, ce qui signifie qu'ils peuvent être supprimés et que le processus de synthèse de l'ADN peut continuer. Une caméra est utilisée pour identifier quelle étiquette a été attachée avant le retrait des terminaisons. Pour faciliter l'automatisation et augmenter le débit du processus, les fragments d'ADN à séquencer sont attachés à une surface (appelée cellule d'écoulement ou Flow Cell), de manière ordonnée. La Flow Cell est similaire à une grille dans laquelle chaque carré peut contenir un fragment d'ADN spécifique à séquencer. Avant le début du séquençage, un processus de clonage a lieu pour augmenter le nombre de copies des fragments d'ADN par cellule (appelées clusters). Une fois le séquençage commencé, et après suffisamment de temps pour que les nucléotides se lient, une caméra capture une image, avant que les terminateurs ne soient supprimés et que le séquençage ne se poursuive. Chaque image prise par la caméra montrera, pour chaque cluster, une couleur particulière, correspondant à la base attachée à ce moment-là.

Bien que certaines technologies soient meilleures que le séquençage Illumina sur certains aspects, cette méthode est le meilleur compromis (qualité, débit, coût, etc.) et est la principale raison pour laquelle elle est si répandue aujourd'hui. Néanmoins, le principal inconvénient d'Illumina (et de la plupart des technologies NGS) est la longueur maximale des fragments d'ADN pouvant être séquencés (actuellement de l'ordre de quelques centaines de paires de bases) [154]. Cela signifie que des séquences d'ADN plus longues sont décomposées en petits fragments à séquencer et que les courts fragments séquencés résultants, appelés reads de séquençage, doivent ensuite être assemblés ou alignés sur un génome de référence.

1.2.3.1. RNA-sequencing

Le séquençage à ARN (RNA-seq) est une méthode expérimentale qui tire parti de la technologie NGS pour effectuer une analyse de l'expression génique qui a été publiée pour la première fois en 2008 [177]. Il vise à séquencer le contenu en ARN des cellules. Comme presque toutes les technologies de séquençage d'ADN reposent sur le séquençage par synthèse, il n'est pas possible de séquencer directement des transcrits d'ARN avec les mêmes protocoles. L'ARN-seq repose sur une enzyme appelée transcriptase inverse, principalement présente dans les virus, qui transcrit l'ARN en ADN. Grâce à cette enzyme, il est possible de créer des fragments d'ADN complémentaires aux transcrits d'ARN ; ceux-ci sont appelés ADNc. Les fragments d'ADNc peuvent ensuite être séquencés par des méthodes de séquençage NGS "standard". De la même manière que le ChIP-seq (voir section 1.2.3.2), les reads de séquençage produites peuvent être allignés sur le génome d'une espèce et les gènes transcrits montreront des enrichissements en read.

Étant donné que de nombreux transcrits d'ARN sont modifiés après la transcription, l'alignement des reads sur le génome de référence peut s'avérer difficile. Une autre solution consiste à aligner les reads sur un transcriptome de référence [323]. Comme pour les génomes de référence, les transcriptomes de référence sont également produits et conservés. L'assemblage des read d'ADNc "de novo" est également possible pour collecter les séquences d'ARN [105]. Une autre nouveauté récente a été la réalisation du protocole RNA-seq sur des cellules individuelles (single-cell RNA-seq), et les résultats sont prometteurs notamment pour comprendre l'expression génique au niveau des cellules uniques [117].

1.2.3.2. ChIP-seq

Le ChIP-seq (Chromatine ImmunoPrecipitation sequencing) est une méthode moderne utilisée pour détecter des régions génomiques fixées par des protéines telles que les TF ou les histones. Cette méthode s'appuie sur un certain nombre d'étapes [226]. Cette technique est basée sur le concept d'immunoprécipitation, un processus utilisé pour purifier les protéines. Les séquences d'ADN fixées par le TF sont isolées à l'aide de l'anticorps spécifique au TF d'intérêt. Ces séquences subissent ensuite un procédé de précipitation.

Étant donné que les protéines d'une cellule se lient et se délient en permanence à l'ADN, la première étape de l'immunoprécipitation de la chromatine consiste à réticuler (fixer) toutes les protéines fixées à l'ADN. Cela se fait généralement avec du formaldéhyde et avec un échantillon contenant un grand nombre de cellules vivantes. Pour permettre la précipitation, la deuxième étape du processus consiste à rompre tous les liens ADN-protéines. Étant donné que les protéines protègent l'ADN, des méthodes qui coupent l'ADN de manière aléatoire peuvent être utilisées (sonication), ce qui donne un mélange de protéines toujours liées à l'ADN et de petits fragments d'ADN. Une immunoprécipitation est ensuite effectuée pour précipiter uniquement les protéines d'intérêt ainsi que l'ADN auquel elles sont fixées. Après l'étape de précipitation, la défixation des protéines libère tout le matériel ADN précédemment lié qui peut être ensuite analysé.

La première version de l'immunoprécipitation de la chromatine utilisait une technique basée sur les puces à ADN, pour identifier cet ADN les chercheurs l'hybridaient à un ensemble de sondes cette technique est appelée ChIP-Chip ou ChIP-on-chip [30]. Elle a permis l'identification des motifs de fixation des TF dans le génome de référence afin d'étudier les modèles de fixation de protéine à l'ADN à l'échelle du génome. La version moderne de ce test publiée pour la première fois en 2007 utilise le séquençage utilisant la technologie NGS au lieu d'une puce à puce, et est donc appelée immunoprécipitation de la chromatine suivie d'un séquençage (ChIP-seq) [20]. L'ensemble du processus est illustré à la Figure 1.16.

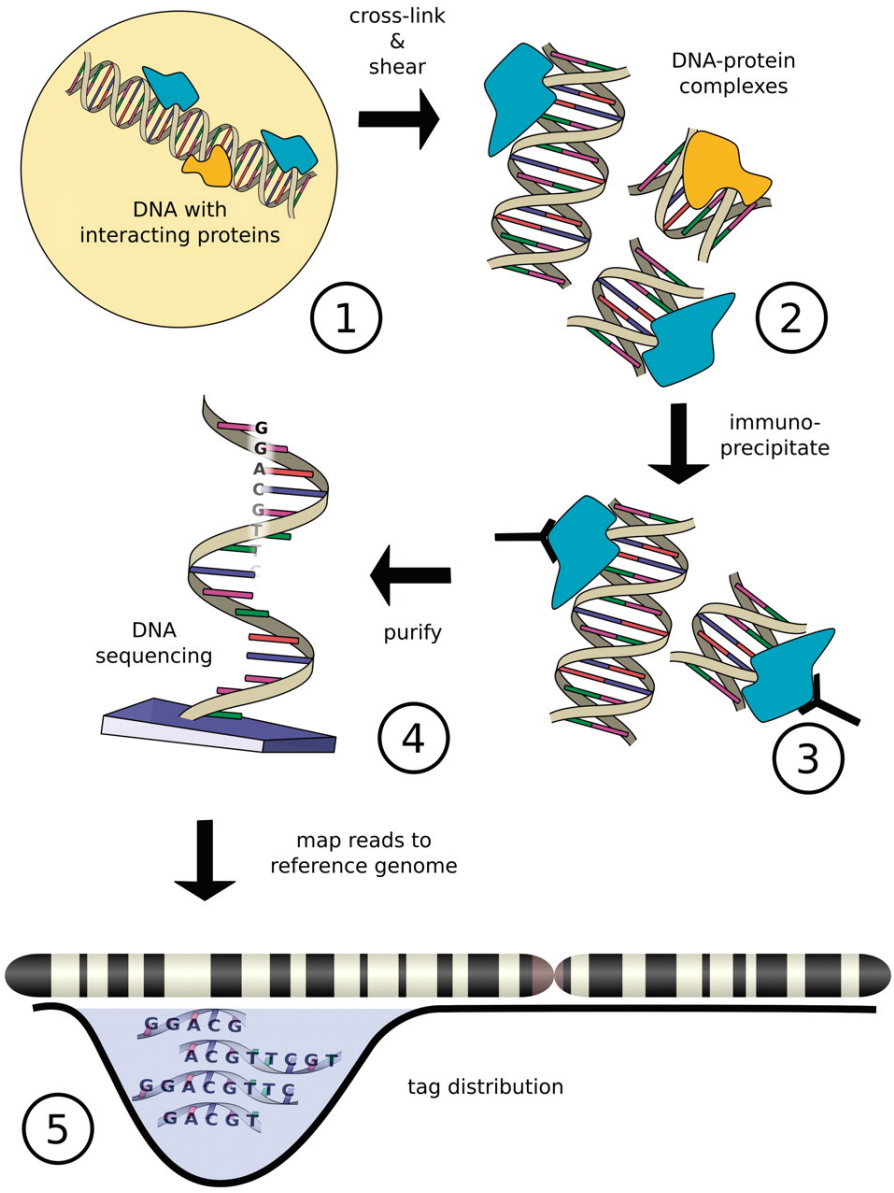


FIGURE 1.16. – ***Workflow d'une analyse ChIP-seq.*** *La chromatine dans le noyau (1) est fixée chimiquement et fragmentée (2), suivie de l'enrichissement des complexes contenant la protéine cible en utilisant l'immunoprécipitation (3). Les "short reads" obtenues à partir du séquençage massivement parallèle (4) sont cartographiées sur un génome de référence (5), ce qui donne une distribution de marqueurs sur le génome.* [16].

16. Szalkowski, Adam M., and Christoph D. Schmid. "Rapid innovation in ChIP-seq peak-calling algorithms is outdistancing benchmarking efforts." Briefings in bioinformatics 12.6 (2011).

Les fragments de lecture, produits à partir du ChIP-seq, sont ensuite alignés sur le génome de l'espèce étudiée. Cela crée des amas de reads appelées pics ChIP-seq qui s'alignent sur les régions génomiques fixées à des protéines. A partir de ces données, il est possible d'analyser les profils de fixation de protéines spécifiques à l'échelle du génome, grâce à un "Peak Caller". Cette étape sera détaillé dans la partie 1.3.3.

Bien que le ChIP-seq soit le protocole le plus largement utilisé pour identifier les sites de fixation des facteurs de transcription in vivo, il présente certains inconvénients. La première contrainte est d'avoir un anticorps de haute qualité spécifique au facteur de transcription d'intérêt, qui peut ne pas toujours être disponible. La détection de la fixation est possible que pour un seul TF (ChIP-seq) ou deux TF ou co-facteur (avec la technique ChIP-reChIP), il serait donc nécessaire de réaliser des centaines d'expériences pour caractériser de manière exhaustive les sites de fixation pour tous les facteurs de transcription actifs dans un type cellulaire donné. De plus, en raison de la nature du protocole, la résolution des pics est faible (200-300pb) ; ainsi, nous ne pouvons pas déterminer avec précision quelle séquence d'ADN exacte se fixe au TF. Des travaux ont montré que certaines régions de la chromatine interagissent avec un grand nombre de TF de manière non spécifique et que des corps de gènes subissent des niveaux élevés de transcription et présentent un enrichissement en facteurs de transcription inattendus [225]. Ces régions sont considérées comme des artefacts associés aux biais liés à ce protocole. Cependant, le ChIP-seq reste toujours la technique de choix dans le domaine pour caractériser les TFBS.

En réalité, l'étape de précipitation des anticorps n'est pas efficace à 100% et les fragments d'ADN résultants qui sont ensuite séquencés couvrent des régions beaucoup plus grandes que les régions fixées à des protéines. Néanmoins, les régions fixées à des protéines sont enrichies en reads alignés et les pics sont néanmoins détectables si l'expérience est de bonne qualité. Des logiciels utilisant divers modèles mathématiques ont été développés pour détecter ces régions enrichies en reads [322]. L'un des biais majeurs des expériences de ChIP-seq sont les régions de chromatine ouverte, qui sont également enrichies en reads. Il est donc courant d'effectuer une expérience ChIP "nue" (pas de réticulation ni de précipitation) ou une expérience "mock ChIP" (utilisation d'un anticorps "faux" qui ne se fixe à aucune protéine) en parallèle avec le même échantillon que l'expérience ChIP-seq normale [246]. Ce type d'expérience de contrôle ou input permet la comparaison entre les deux pistes (track) de pic résultantes et donc permet la suppression de tout pic qui n'est probablement pas le résultat d'une région liée à une protéine [173]. La plupart des programmes d'analyse ChIP-seq sont conçus pour utiliser de telles expériences de contrôle.

1.2.3.3. ChIP-exo

Le ChIP-exo a été développé à l'Université de Pennsylvanie en 2012 [241] et adapté plus tard pour les plateformes Illumina [262] (Figure 1.17). Le protocole a été mis à jour pour nécessiter moins d'étapes de traitement enzymatique permettant de plus petites quantités d'ADN en input [119], et récemment, il a été suffisamment raffiné pour permettre l'échantillonnage de populations aussi faibles que 27 000 cellules [244].
ChIP-exo est une méthode basée sur l'immunoprécipitation de la chromatine pour cartographier les emplacements auxquels une protéine d'intérêt (TF) se fixe sur le génome. Cette technique s'inspire du protocole ChIP-seq, améliorant la résolution des sites de fixation de centaines de pb à près d'une pb. Il utilise une exonucléase pour dégrader les brins d'ADN lié à la protéine en direction 5'-3'. Cette dégradation se fait très proche du site de fixation de la protéine (quelques nucléotides) [241]. Les nucléotides des extrémités traitées par l'exonucléase sont fixés à l'aide d'une combinaison de séquençage d'ADN, de puces à ADN et de PCR. Ces séquences sont ensuite alignées sur le génome pour identifier les sites de fixation des TF.

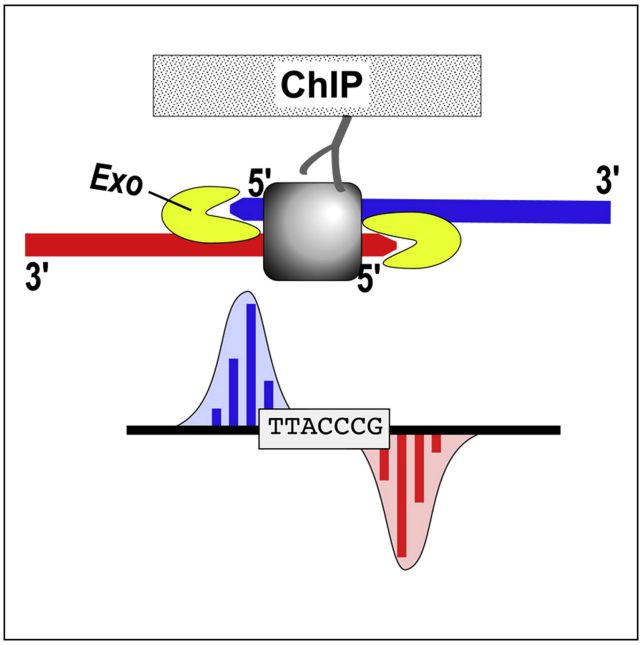


FIGURE 1.17. – ***Figure illustrant la technique ChIP-exo.*** *Des anticorps ciblant une protéine sont utilisés pour fixer les deux brins d'ADN (sens et antisens) (en bleu et rouge). Des exonucléases (en jaune) pour éliminer ensuite les extrémités non protégées de l'ADN. Le ChIP-exo permet d'identifier des pics de signal (en bas) à la fois en amont et en aval de la séquence d'ADN fixée à la protéine.*[17].

17. Rhee, Ho Sung, and B. Franklin Pugh. "Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution." Cell 147.6 (2011).

1.2.3.4. DAP-seq

Le DAP-seq (DNA affinity purification sequencing)(Figure 1.18) est une technique récemment développée pour découvrir des sites de fixation de TF produisant des datasets du même type que le ChIP-seq. Le DAP-seq utilise des TF exprimés de manière exogène (TF construit in vitro) pour interroger directement l'ADN génomique, sans avoir besoin de lignées transgéniques étiquetées ou d'anticorps spécifiques au TF, ce qui est pratique pour les espèces avec peu d'anticorps (exemple *A. thaliana* dans ReMap). En combinant avec la technologie de séquençage à haut débit, les fragments d'ADN produits après DAP sont séquencés et analysés pour trouver les sites de fixation de la protéine à l'ADN. Le DAP-seq peut résoudre les limitations de la technologie ChIP qui sont notamment le manque de d'anticorps spécifiques pour certaines protéine cible.

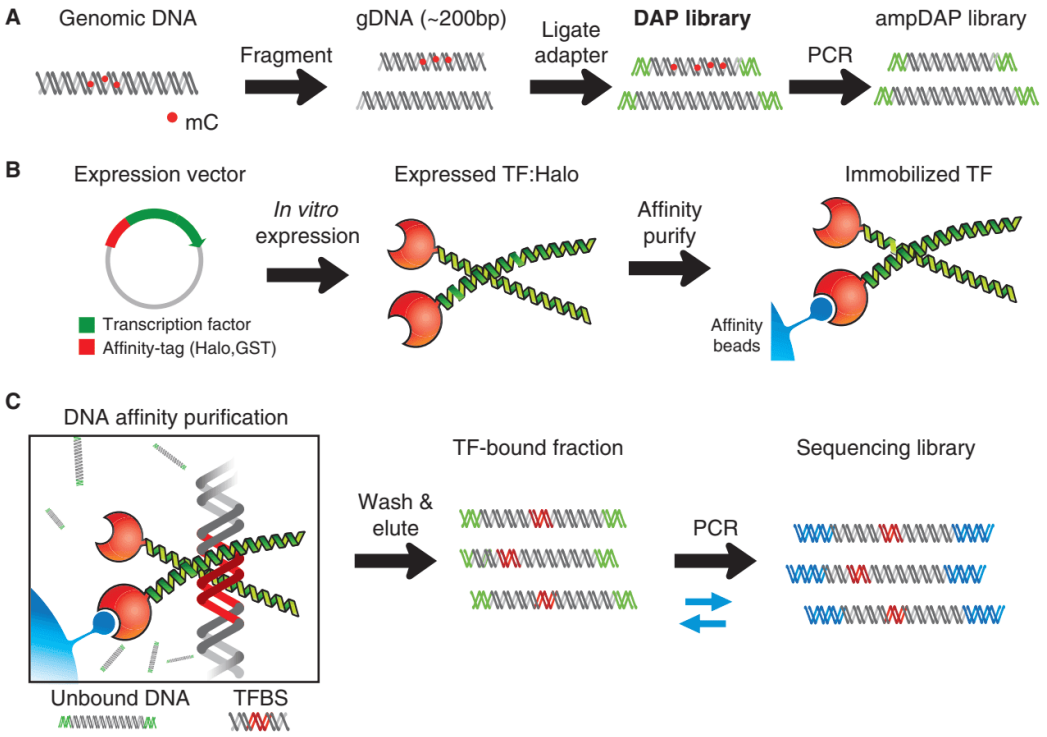


FIGURE 1.18. – ***Workflow de la techinque DAP-seq.*** *A) Préparation de la librairie. B) Expression de la protéine. C) Purification de l'ADN.* [18].

Le DAP-seq peut-être utilisé pour plusieurs applications, l'analyse des sites de fixation et les fonctions des facteurs de transcription. Il permet également de déterminer le type de modification d'histones à une position spécifique du brin d'ADN ainsi que pour l'étude de l'impact des modifications d'histones sur l'expression des gènes.

18. https://www.cd-genomics.com/dap-seq-service.html

1.2.4. Techniques pour profiler la chromatine ouverte

1.2.4.1. DNase-seq

On sait que les éléments régulateurs résident dans les régions du génome ouvertes et sans nucléosomes. Ces régions sont hypersensibles à l'attaque des nucléases. La digestion avec la nucléase DNase I, couplée au séquençage à haut débit (DNase-seq), est la première technique établie à l'échelle du génome pour sonder de telles régions de chromatine ouvertes [36, 122], et est largement appliquée dans des consortiums de recherche tels que ENCODE [53] ou Roadmap Epigenomics [163].

Initialement isolée du pancréas bovin [164], la DNase I est une nucléase qui peut cliver les molécules d'ADN en hydrolysant les liaisons phosphodiester du squelette phosphate de la molécule de sucre. Au milieu des années 70, des travaux ont démontré que la DNase I clive préférentiellement la chromatine transcriptionnellement active. Plus précisément, les cellules dans lesquelles des loci de gènes sélectionnés [308] et ovalbumine [100] sont actifs sur le plan de la transcription sont soumises à une digestion par la DNase I. Cet effet n'est pas observé pour les cellules où les gènes ne sont pas transcrits, ce qui laisse supposer que l'activité transcriptionnelle est associée à une conformation altérée de la chromatine, plus sensible au clivage par la DNase I. Ces observations ont été étendues quelques années plus tard, dans des travaux sur le virus simien 40 [260] et la chromatine de Drosophila [313], qui ont démontré que la DNase I clive la chromatine sous-jacente d'une manière spécifique à la position. Ces travaux ont estimé que deux régions de préférence digérées par la DNase I étaient plus courtes que la longueur de l'ADN enroulé autour de 2 nucléosomes et mesurant environ 140 paires de bases chacunes. De plus, les deux articles ont discuté du fait que ces régions pourraient éventuellement être dépourvues de nucléosomes. Ils ont défini pour la première fois ce que nous savons maintenant être les sites hypersensibles à la DNase (DHS).

Historiquement, les DHS étaient cartographiés à l'aide d'une méthode appelée étiquetage indirect [312]. La chromatine est d'abord digérée par la DNase I, puis l'ADN isolé est ensuite clivé par une endonucléase de restriction (ER) spécifique à la séquence. Les fragments résultants sont séparés par électrophorèse sur gel, transférés sur une membrane et hybridés à des sondes spécifiques aux séquences se trouvant aux flancs des sites de clivage de l'ER. Les longueurs de fragment déterminées fournissent une mesure directe de la distance entre les sites de clivage des ER et la DNase I, à partir de laquelle les DHS peuvent être repérés. Cette méthode est à faible débit et ne peut être appliquée qu'à un nombre limité de locus à la fois, car elle nécessite que les loci d'intérêt soient préalablement caractérisées (par exemple, connaissance de la séquence et des sites de reconnaissance des ER). Cependant les DHS sont maintenant facilement cartographiés à l'échelle du génome grâce à la DNase-seq à haut débit.

Il existe deux variantes de DNase-seq qui sont généralement appelées protocoles single-hit (ou end capture) [36, 273] et double-hit [122], car les fragments résultants représentent une DNase I coupée au niveau d'une ou aux deux extrémités, respectivement. Dans le protocole single-hit (Figure 1.19, à gauche), la digestion par la DNase I est effectuée et le premier lieur (linker) hébergeant un site de restriction de l'enzyme de restriction MmeI est ligaturé aux extrémités digérées par la DNase I. MmeI coupe ensuite 20 pb en aval de son site de restriction, où le deuxième lieur est ensuite ligaturé. Dans le protocole double-hit (Figure 1.19, à droite), la chromatine digérée par la DNase I est soumise à un fractionnement pour obtenir des fragments de 100 à 500 pb. Les adaptateurs Illumina sont ensuite ligaturés aux extrémités des fragments. Dans les deux protocoles, les fragments linker-flanked sont amplifiés, purifiés et séquencés par PCR en suivant le workflow de séquençage Illumina.

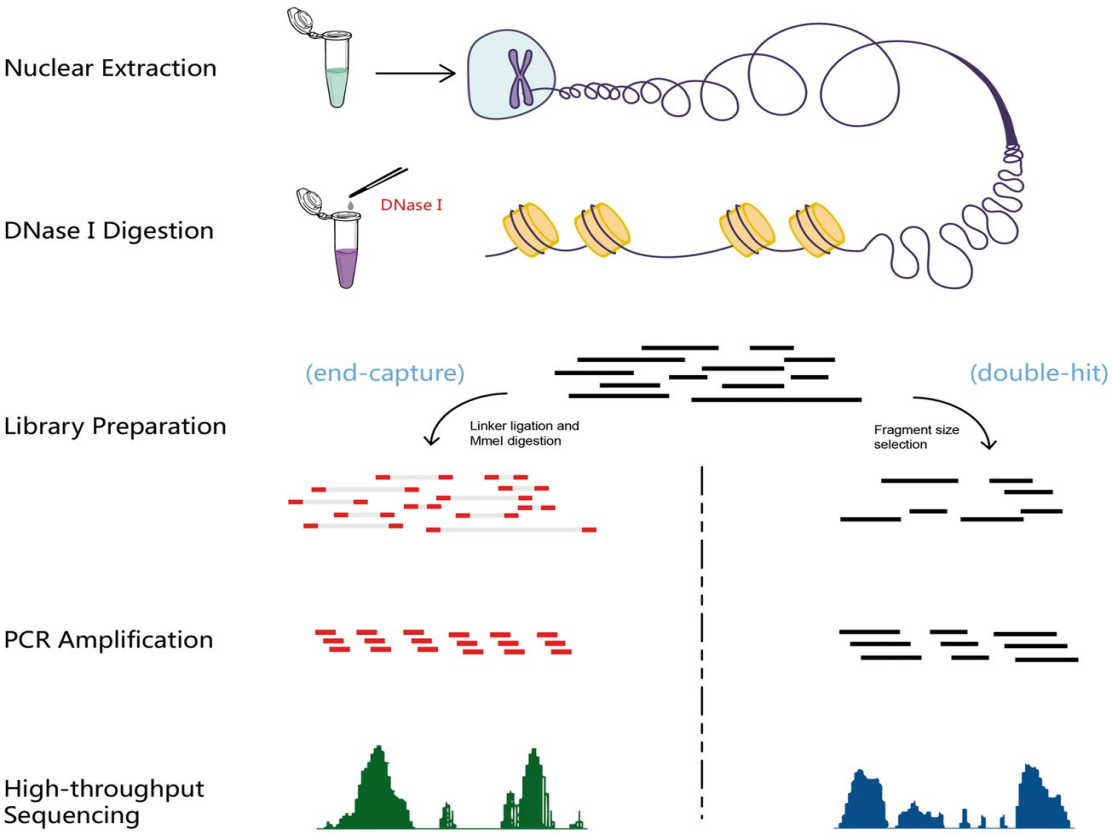


FIGURE 1.19. – ***Aperçu des protocoles expérimentaux DNase-seq.*** *Les étapes de base nécessaires pour une expérience DNase-seq sont présentées dans la figure, comprenant l'extraction nucléaire, la digestion par la DNase I, la préparation de la bibliothèque, l'amplification par PCR et le séquençage à haut débit. Différentes stratégies de préparation de bibliothèques sont adoptées par les méthodes d'end-capture et de double-hit.* [19].

19. Liu, Yongjing, et al. "A practical guide for DNase-seq data analysis : from data management to common applications." Briefings in bioinformatics 20.5 (2019).

1.2.4.2. ATAC-seq

Une technique plus récente pour profiler les régions de chromatine ouvertes est le dosage de la chromatine accessible par transposase en utilisant le séquençage (ATAC-seq) [40]. Au lieu d'une nucléase comme la DNase I, l'ATAC-seq utilise des enzymes transposases Tn5. Les transposases contribuent aux réarrangements génomiques et par conséquent à l'évolution du génome, en mobilisant des éléments d'ADN appelés transposons [206]. Tn5 est un transposon bactérien qui confère une résistance aux antibiotiques à l'hôte, à travers les trois gènes de résistance qu'il abrite (kanamycine, bléomycine et streptomycine). La mobilisation du transposon Tn5 est réalisée via un mécanisme "couper-coller" où il est clivé de son emplacement d'origine et inséré dans l'ADN cible par la transposase Tn5. Cela dépend de l'interaction spécifique de la transposase avec les séquences de 19 pb aux deux extrémités du transposon Tn5. Le complexe transposon-transposase Tn5 se lie alors à l'ADN cible et via une attaque nucléophile, les extrémités 3' du transposon se lient de manière covalente aux extrémités 5' de l'ADN cible clivé. Au cours de ce processus, le brin antisens de l'ADN cible est clivé à une position 9pb en aval du brin sens, ce qui conduit à la duplication de ces 9pb de part et d'autre du transposon inséré. Une meilleure compréhension du fonctionnement des composants de ce système ainsi que des modifications ont conduit à son utilisation comme outil in vitro. Il s'agit notamment de rendre la transposase Tn5 hyperactive par le biais de mutations et d'utiliser une version modifiée de la séquence terminale de 19 pb appelée extrémité mosaïque avec une plus grande efficacité de transposition [325].

En outre, il a été constaté que le préchargement in vitro des transposases Tn5 hyperactives avec des adaptateurs de séquençage hébergeant des séquences des extrémités mosaïque, sans l'ADN de transposon intermédiaire, est suffisant pour la transposition [7]. Puisqu'il n'y a pas d'ADN intermédiaire, cette réaction de transposition altérée conduit à la fragmentation de l'ADN cible via l'attaque et le clivage de la transposase tandis que les fragments résultants sont simultanément marqués par ligature de l'adaptateur aux extrémités 5', un processus connu sous le nom de "tagmentation". Ces développements et les rapports de transposons s'intégrant préférentiellement dans les régions sans nucléosome [97], ont ouvert la voie à ATAC-seq en tant que méthode basée sur la transposase Tn5 pour profiler la chromatine ouverte [40]. L'ATAC-seq a un protocole rapide et simple qui comprend l'isolement des cellules/noyaux, la lyse, la tagmentation, l'amplification par PCR et le séquençage (Figure 1.20). Beaucoup moins de matériel de départ est nécessaire pour l'ATAC-seq par rapport à la DNase-seq, 500-50 000 [40] contre 1-10 millions [273] de cellules ou noyaux, respectivement, bien que des variations récentes du protocole permettent d'appliquer les deux techniques au niveau d'une seule cellule [41].

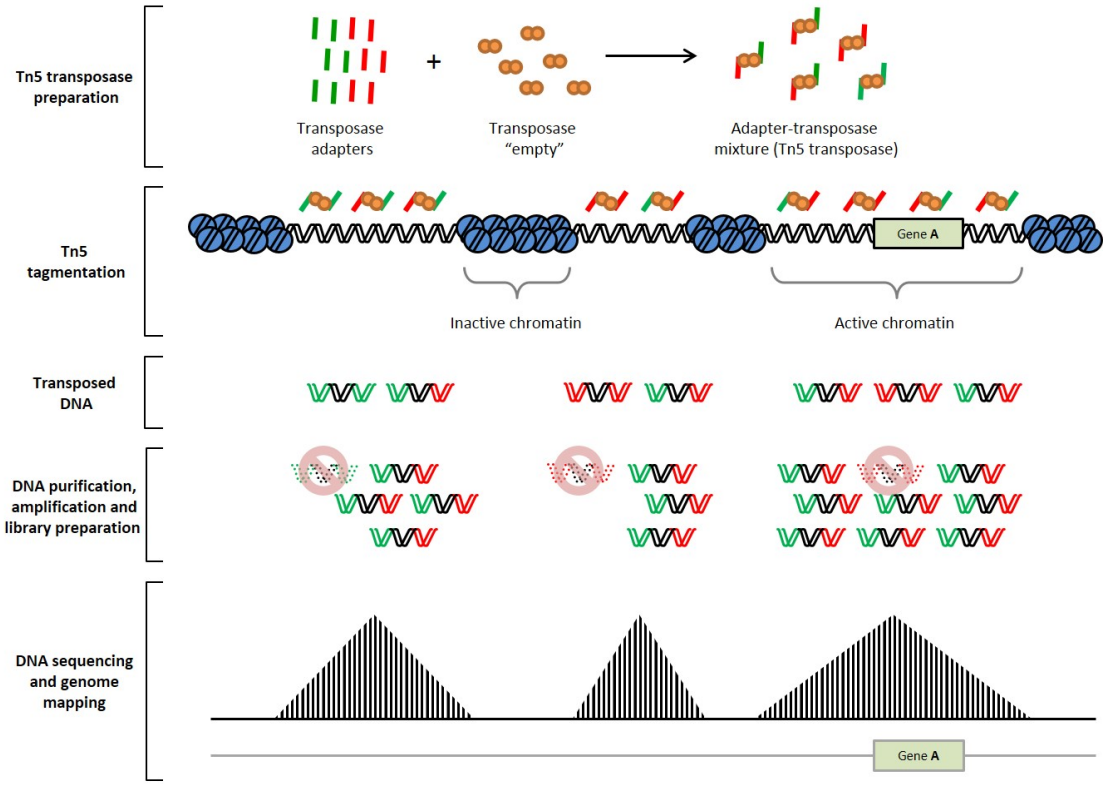


FIGURE 1.20. – ***Figure illustrant les cinq étapes de l'ATAC-seq.*** *Préparation de la transposase Tn5 pour l'ATAC-seq. Puis la transposition de Tn5 dans les cellules, permettant ainsi la fragmentation et le marquage des fragments d'ADN accessibles. Ensuite la purification des fragments d'ADN transposés marqués. Amplification par PCR des fragments d'ADN purifiés pour la préparation de la bibliothèque. Et enfin, séquençage des fragments d'ADN marqués et cartographie sur le génome.*[20].

20. https://eciofishr.wordpress.com/2019/04/22/technical-section-atac-seq/

1.2.4.3. Hi-C : une méthode de capture de la conformation de la chromatine

Hi-C est une technologie de capture de la conformation de la chromatine à l'échelle du génome dérivée de la technique 3C (capture conformation de la chromatine) [68] (Figure 1.21). Tout d'abord, les cellules sont réticulées afin que les contacts entre les régions d'ADN, avec ou sans l'aide d'une protéine nucléaire, soient stabilisés. L'ADN est ensuite digéré à l'aide d'une enzyme de restriction et les fragments sont liés. Après ligation, la réticulation est inversée. Le but est d'obtenir des fragments d'ADN circulaires contenant les deux régions qui étaient en contact dans le noyau. Dans le protocole Hi-C, des nucléotides biotinylés sont ajoutés à la jonction des fragments, de sorte que l'ADN participant aux contacts puisse être extrait de tous les fragments. Les fragments extraits sont ensuite directement séquencés. Lors de l'alignement suivant, les reads doivent s'aligner sur deux régions du génome, ce qui signifie que ces deux régions étaient en contact étroit dans le noyau. Cependant, comme le nombre de contacts avec la chromatine dans le génome est difficile à déterminer et qu'un nombre minimum d'observations doivent être faites pour distinguer les vrais contacts du bruit, un compromis doit être fait entre la mise à l'échelle et la profondeur. Toutes les interactions avec bins de longueur fixe sont additionnées, et la longueur des bins correspond à la résolution de l'expérience Hi-C, de sorte qu'un nombre élevé correspond à une résolution inférieure [165]. Par conséquent, les expériences à haute résolution (bins de 10 kb ou moins) nécessitent une profondeur de séquençage très élevée, mais permettent de trouver des structures à une échelle fine, tandis que les expériences à basse résolution sont plus faciles à produire mais ne collectent que de grandes structures.

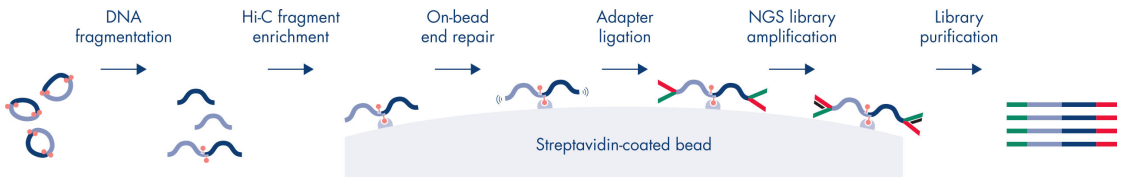


FIGURE 1.21. – ***Un aperçu du workflow de la technique Hi-C workflow.*** [21].

21. https://www.qiagen.com/us/products/discovery-translational-research/epigenetics/epitect-hi-c-kit-us

Les expériences Hi-C sont généralement visualisées grâce à des heatmaps. Les axes x et y de la heatmap correspondent aux bins définis par la résolution. A leur intersection, l'intensité de la couleur correspond à la force de l'interaction : blanc s'il n'y a pas d'interaction enregistrée entre les régions, rouge s'il y a une forte interaction. Le nombre d'interactions qu'il faut observer pour être qualifiées de "fortes" peut varier d'une expérience à l'autre et selon la résolution. Les heatmaps Hi-C sont symétriques, car les interactions ne sont pas orientées (le nombre d'interactions de la région 1 à la région 2 est le même que le nombre d'interactions de la région 2 à la région 1). Par conséquent, les heatmaps Hi-C sont souvent coupées le long de la diagonale et représentées sous forme de triangles, assis sur leur hypoténuse. Les bins sont marqués le long de l'hypoténuse. En raison de leurs modèles d'interaction, les compartiments créent un motif en forme de damier sur les heatmaps Hi-C, les TAD peuvent être vus comme des triangles et les boucles de chromatine sont représentées par un seul point rouge [237] (Figure 1.22).

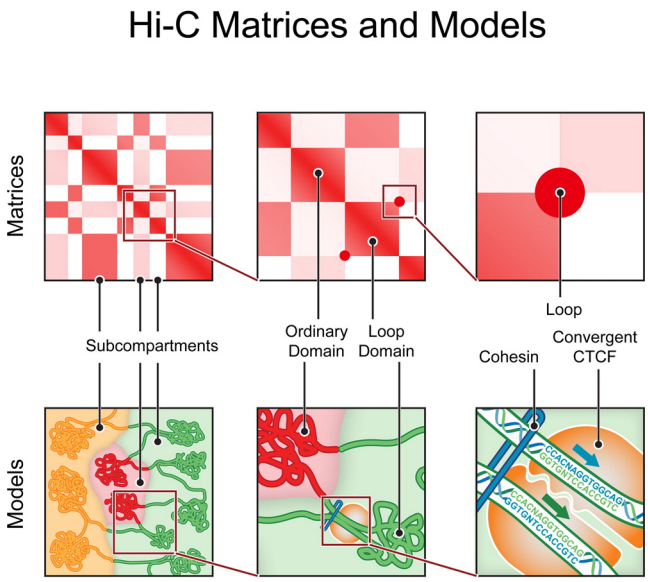


FIGURE 1.22. – ***Aperçu des caractéristiques révélées par des cartes Hi-C.*** *A gauche : le schéma de contact à longue distance d'un locus (en haut) indique son compartiment nucléaire (en bas). Six sous-compartiments ont été détectés. Milieu : les carrés de fréquence de contact renforcée le long de la diagonale (en haut) indiquent la présence de petits domaines de chromatine condensée, dont la longueur médiane est de 185 kb (en bas). A droite : les pics dans la carte de contact (en haut) indiquent la présence de boucles (en bas). Ces boucles ont tendance à se situer aux frontières des domaines et à lier CTCF.*[22].

22. Rao, Suhas SP, et al. "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." Cell 159.7 (2014).

1.3. Analyses bioinfo des techniques de séquençage

1.3.1. Contexte

Dans la première partie de ma thèse (section 2), j'ai collecté annoté et traité des données ChIP-seq afin de construire les catalogues ReMap de régions régulatrices. Ce processus a été réalisé avec un workflow, permettant de traiter ces données de la donnée brut (FASTQ) au fichier avec les pics (BED). Les étapes permettant le traitement de ces données seront décrites dans ce chapitre.

1.3.2. Traitement, alignement et filtrage des fragments de lecture

Les technologies de séquençage de nouvelle génération produisent des séquences de longueur définies, également appelées fragments de lectures ou "reads", appartenant à des fragments d'intérêt d'une configuration expérimentale donnée. La sortie à ce stade est généralement au format FASTQ, qui comprend à la fois les informations de séquence, ainsi qu'un contrôle qualité sur les nucléotides pour chaque read. Un choix populaire pour le contrôle qualité est appelé le score phred [83], qui est essentiellement une mesure de la probabilité que l'appel de base pendant le séquençage soit correct pour chaque nucléotide. Les fragments de lecture peuvent contenir des séquences d'adaptateur qui doivent être retirées, afin d'obtenir un alignement correct sur le génome de référence. Cette étape est réalisée par des outils de découpage d'adaptateurs (trimming tools) [189, 259, 157, 132]. Dans le cadre de ma thèse, nous avons été amené à utiliser l'outil Trim galore [160], qui utilise Cutadapt [189] et FastQC [11] afin de supprimer les adaptateurs.

L'étape suivante consiste à aligner les fragments de lecture sur le génome de référence. Cette étape permet de déterminer de quelle région génomique le fragment provient sur le génome de référence tout en tenant compte des décalages et des lacunes. Cette étape peut amener des erreurs d'appariements [239]. Les principales raisons de ce problème sont les erreurs de séquençage et les différences entre l'échantillon testé et le génome de référence dues à la variation de séquence.

Une autre considération à prendre en compte est la quantité de données. Aligner des millions de reads sur un génome de référence long de plusieurs millions de bases nécessite des algorithmes efficaces. Deux idées algorithmiques utilisées par les aligneurs sont le filtrage et l'indexation [239]. Le filtrage élimine les régions du génome de référence où une correspondance n'est pas attendue pour un fragment donné, en comparant les sous-séquences plus courtes au sein de la lecture au génome de référence. L'indexation, d'autre part, fait référence au prétraitement du génome de référence pour permettre d'interroger les séquences correspondantes beaucoup plus rapidement. Les principales approches d'indexation utilisées sont le tableau de suffixes amélioré [4] et le FM-index [88], qui est basé sur la transformée de Burrows-Wheeler [44]. Les aligneurs Bowtie2 [167] et BWA [171], utilisés dans les analyses présentées dans cette thèse, intègrent le FM-index. La sortie est au format SAM qui peut être convertie au format binaire BAM.

Une fois l'alignement effectué, il est souvent nécessaire de traiter le fichier BAM. En effet, certains fragments de lectures sont alignés à plusieurs positions sur le génome. Il est possible de filtrer ces fragments pour conserver uniquement ceux alignés sur une seule position, ce qui augmente la fiabilité mais diminue la couverture des régions. En outre, la plupart des méthodes de préparation de bibliothèque incluent une étape d'amplification par PCR, ce qui peut créer plusieurs copies du même fragment. Le séquençage par extrémités appariées (paired-end) permet d'éliminer les doublons de manière plus fiable, car il prend en compte les extrémités 5' et 3'. Plusieurs outils permettent de filtrer ces réplicats PCR (ex : FastUniq [317]), lors de ma thèse nous avons utilisé l'outil rmdup de la suite SAMtools [172].

1.3.3. Recherche des pics de fixation

Lors de l'alignement des reads sur le génome, il apparaît un enrichissement des fragments d'ADN autour des sites de fixation des protéines. Les sites d'enrichissement résultants sont généralement plus larges (> 100 pb) par rapport au site de fixation réel (6-20 pb) [226]. Les régions qui présentent un enrichissement statistiquement significatif en fragments d'ADN (appelées « pics ») sont identifiées à l'aide d'approches informatiques (le "peakcalling"). Comme les fragments d'ADN ChIP-seq sont généralement séquencés de l'extrémité 5' vers l'extrémité 3', les fragments seront en fait légèrement décalés vers l'extrémité 5' sur chaque brin, résultant en deux pics décalés sur chaque brin qui entourent le "vrai" site de fixation. La distance entre les paires de pics est généralement appelée « taille de décalage ». Étant donné que la taille de ce décalage dépend de la taille moyenne des fragments séquencés, les logiciels modernes de peak calling estiment la valeur de ce décalage, puis déplacent les pics en conséquence pour former un seul enrichissement en reads sur le site de fixation correct (Figure 1.23). Certains "peak-caller" [322, 137, 243, 246, 151, 296, 86, 141, 131] tentent de rendre compte de cette propriété : SPP modélise la taille du décalage en calculant des valeurs d'intercorrélation [151]. Le peak calling des expériences de ChIP-seq récolté dans le cadre de cette thèse ont été effectués avec le logiciel MACS [322].

MACS (Model-based Analysis of ChIP-seq) est un programme conçu pour identifier les régions liées aux protéines sur le génome, à l'aide de données ChIP-seq. Il effectue deux étapes clés pour déterminer précisément chaque emplacement lié. La première consiste à évaluer la distance de décalage des pics entre les pics des deux brins d'ADN avant de décaler les reads en conséquence pour former des pics uniques couvrant les sites de fixation des protéines. Dans la deuxième étape de l'analyse, MACS utilise une distribution de Poisson pour modéliser le nombre de reads le long du génome et calculer une valeur p par pic potentiel que l'utilisateur peut ensuite utiliser pour filtrer les pics improbables.

Pour identifier la taille du décalage, MACS utilise d'abord une approche de fenêtre glissante pour identifier les régions avec un enrichissement de fragments significatif par rapport aux données de contrôle. Il sélectionne ensuite au hasard 1 000 de ces régions hautement enrichies et fait l'hypothèse que celles-ci correspondent à des sites liés aux protéines. Chacune de ces régions est composée de fragments de brins directs et indirects, chaque ensemble formant des pics différents. L'alignement de ces fragments décalées entraîne la formation de deux distributions de fragments. MACS utilise ensuite la distance entre les sommets de ces distributions comme distance de décalage de crête d, comme illustré sur la Figure 1.23. Enfin, il décale toutes les fragments alignées vers l'extrémité 3' de la moitié de la distance de décalage estimée, d/2 [187].

Après avoir effectué le décalage des fragments, MACS utilise une fenêtre glissante de longueur 2d pour parcourir l'ensemble du génome et identifier les régions avec un fort enrichissement en fragments. MACS utilise une distribution de Poisson pour modéliser le nombre de fragments (une distribution de probabilité discrète qui est souvent utilisée pour modéliser le nombre de fois où quelque chose se produit dans un intervalle de temps ou d'espace). Pour chaque fenêtre, MACS calcule la valeur p de la fenêtre pour déterminer si une région a un enrichissement significatif. L'ensemble des pics détectés correspond à ces régions significativement enrichies (les régions adjacentes à fort enrichissement en read sont fusionnées).

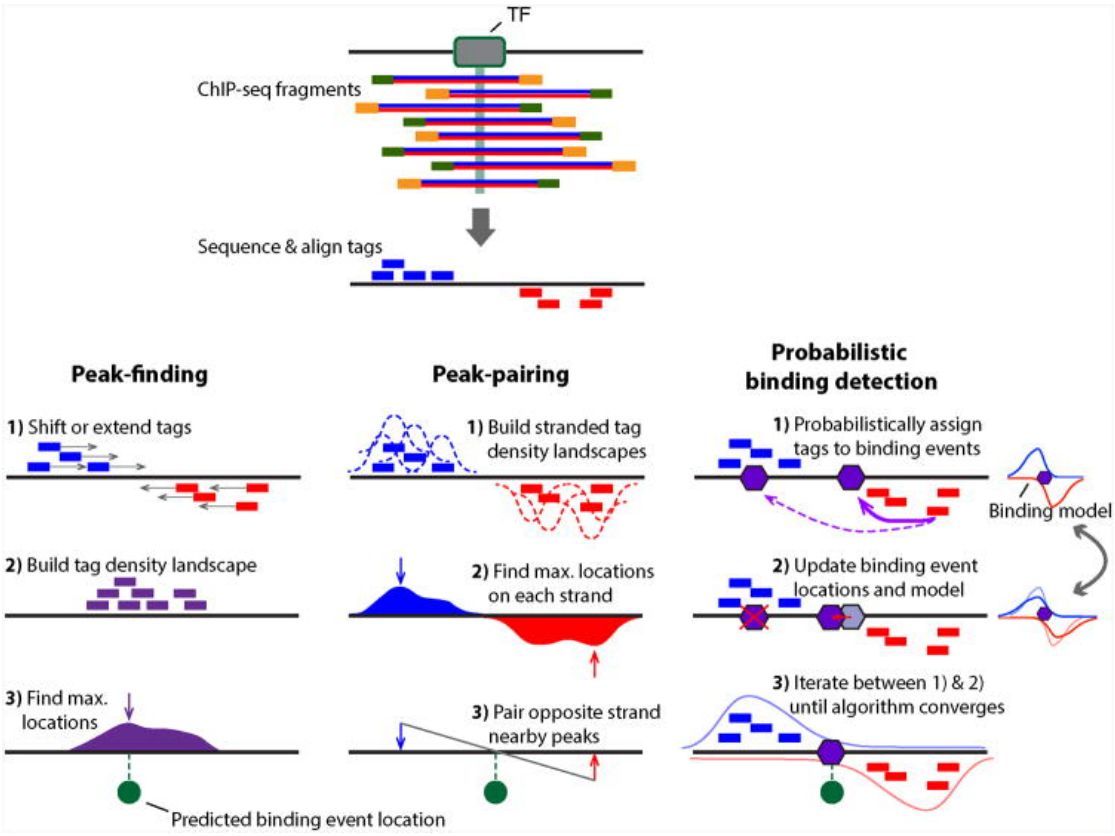


FIGURE 1.23. – ***Méthodes de détection des événements de fixation des TF à l'ADN à partir de données de séquençage à haut débit.*** *Les méthodes de prédiction se basent sur l'observation suivante. L'événement de liaison se situe aux positions où la densité des fragments de librairie est la plus élevée. 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). La méthode d'appariement des pics (peak-pairing). La méthode probabiliste (Probabilistic binding detection).*[23].

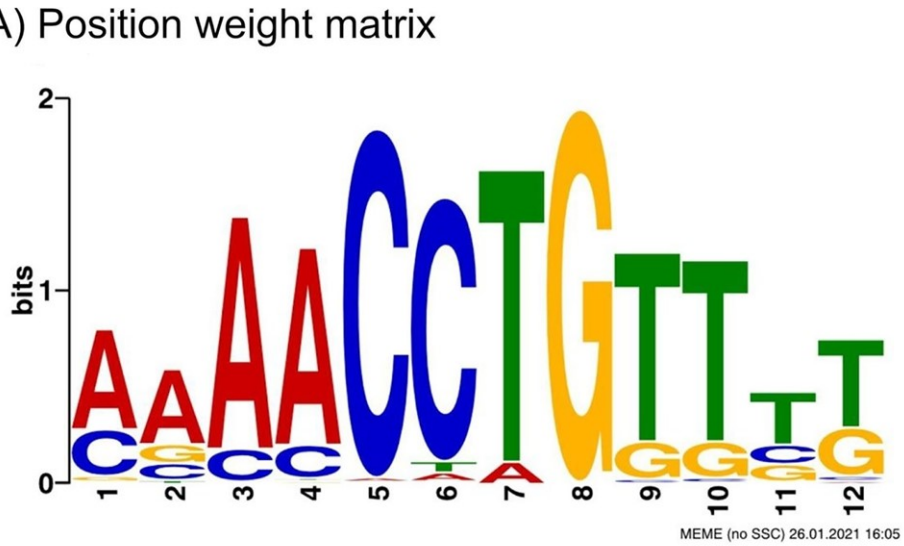
23. Mahony, Shaun, and B. Franklin Pugh. "Protein–DNA binding in high-resolution." Critical reviews in biochemistry and molecular biology 50.4 (2015).

1.3.4. Identifier les motifs de fixation des TF

Comme l'illustre la section précédente, le peak calling sur les données ChIP-seq TF permet d' identifier les régions liées par les TF à l'échelle du génome. Cependant, cette analyse a une résolution relativement faible (c'est-à-dire que les régions identifiées sont beaucoup plus grandes que les sites de contact TF-ADN réels, 100-200pb contre 6-15pb, respectivement).

Les TF se fixent à des séquences courtes (généralement 6-15pb) et spécifiques dans tout le génome appelées sites de fixation. Les sites de fixation d'un TF donné ne sont pas toujours identiques, mais montrent des préférences nucléotidiques spécifiques à la position, également appelées motifs de fixation. Les motifs sont le plus souvent représentés via des matrices de poids de position (PWM) [281]. Comme le montre la Figure 1.24, chaque colonne d'un PWM représente une position de base, les lignes donnant les poids de chacun des quatre nucléotides à cette position. Les pondérations sont généralement les probabilités logarithmiques d'observer un nucléotide donné à une position donnée. Une représentation équivalente à la PWM est appelée la matrice de fréquence de position (PFM) cette matrice utilise des probabilités ou des fréquences plus simples.

Les PWM (et PFM) peuvent également être représentés visuellement, à l'aide de logos de motifs (Figure 1.24), où les nucléotides attendus à une position donnée sont dessinés proportionnellement à leurs poids respectifs, la hauteur totale représentant le contenu informationnel (IC). L'IC d'une position donnée équivaut à la vraisemblance logarithmique (à l'échelle log2) de chaque nucléotide multipliée par sa fréquence, additionnée sur les quatre nucléotides. Ainsi, il va de 0, désignant aucune spécificité, à 2, où un seul nucléotide est spécifiquement préféré à une position donnée. Un biais du modèle PWM est qu'il suppose que toutes les positions de base sont indépendantes les unes des autres, ce qui n'est pas vrai pour tous les TF, où des approches plus complexes peuvent être plus appropriées [280] Néanmoins, les modèles PWM simples sont les plus largement utilisés à ce jour.



A	211	213	283	272	4	10	23	0	1	2	8	7
C	98	48	40	46	320	300	0	2	6	7	69	11
G	12	54	0	4	1	3	0	324	53	53	62	109
T	5	11	3	4	1	13	303	0	266	264	187	199

FIGURE 1.24. – ***Matrice de poids position d'un facteur de transcription obtenue à partir de la découverte de motifs de novo fournie par MEME.*** *(A) Matrice de poids de position d'un TF dérivée de la découverte de motif de novo fournie par MEME.*[24].

24. Leiz, Janna, et al. "Technologies for profiling the impact of genomic variants on transcription factor binding." Medizinische Genetik 33.2 (2021).

Lorsque les séquences de plusieurs sites de fixation pour un TF donné sont connues, et si les positions correspondantes dans les sites de fixation respectifs peuvent être alignées pour représenter le motif de fixation, un PFM peut être construit simplement en calculant les fréquences nucléotidiques par position, qui pourraient alors être converti en PWM. Comme indiqué dans les sections précédentes, le ChIP-seq est une méthode in vivo qui fournit des sites de fixation pour un génome TF donné, qui peuvent ensuite être utilisés pour rechercher le modèle PWM sous-jacent. Cela constitue le problème de découverte de motif de novo, où ni les emplacements précis des sites de fixation dans les pics de ChIP-seq, ni les paramètres de motif attendus ne sont connus, et les algorithmes tentent généralement de trouver les motifs qui maximisent l'IC [280]. Les PWM sont construits à partir de ces approches in vitro, en utilisant des méthodes de calcul sur mesure. Des bases de données telles que UniPROBE [126] et JASPAR [46] fournissent des collections PWM complètes issues de ces efforts. Les PWM de JASPAR ont été obtenue à partir de diverses sources, notamment de la base de données ReMap (Annexe B). Il existe également des méthodes in vitro qui évaluent la liaison TF-ADN, tel que le SELEX à haut débit (HT-SELEX), dans lesquels des séquences longues de 10 à 40 pb sont soumises à des cycles successifs de liaison au TF, conduisant à une spécificité accrue à chaque cycle [139].

Si le ou les modèles de liaison d'un TF sont facilement disponibles, il devient possible de scanner un ensemble de séquences ou l'ensemble du génome, en utilisant le modèle PWM, pour trouver des correspondances de motifs qui constituent des sites de fixation TF putatifs. Des suites d'outils tel que RSAT [288], TRANSFAC[192] et FIMO [106] permettent notamment de parcourir des séquences d'intérêts afin d'identifier des motifs de fixation.

1.3.5. Contrôle qualité

Le consortium ENCODE a défini une série de métriques permettant d'effectuer un contrôle qualité visant à valider les profils de liaison ChIP-seq TF avant l'interprétation biologique [166]. La qualité des données TF ChIP-seq a été évaluée selon les recommandations du consortium ENCODE [53] :

Le Normalized strand correlation (NSC $\geq$ 1,05), déterminée par l'analyse de corrélation croisée et correspondant au rapport entre le pic de la puce et le signal de fond. Une faible valeur NSC indique un faible enrichissement. Le Relative strand correlation (RSC $\geq$ 0,8), déterminée par l'analyse de corrélation croisée et correspondant au rapport entre le pic ChIP et le pic fantôme (pic de longueur de read). Une faible valeur RSC indique un faible rapport signal/bruit.Et enfin, le FRiP (Fraction of Reads in Peaks) est une mesure de qualité permettant de quantifier la proportion de reads qui correspondent à des pics significatifs par rapport au nombre total de reads dans l'échantillon. Plus le FRiP est élevé, meilleure est la qualité des données. Un FRiP élevé indique que l'expérience a réussi à enrichir les régions d'intérêt avec une faible quantité de bruit de fond.

1.3.6. Workflow

Un workflow est un ensemble ordonné d'étapes logicielles qui permettent d'automatiser les processus de traitement de données complexes en bioinformatique. Les workflows sont souvent utilisés pour gérer les tâches qui nécessitent des étapes de traitement interdépendantes, telles que l'analyse de séquences génomiques. Les workflows peuvent être implémentés à l'aide de différents outils de développement de logiciels, tels que Snakemake [158] et Nextflow [69].

Snakemake est un outil populaire pour la création de workflows en bioinformatique basé sur Python. Il est capable d'exécuter plusieurs étapes en parallèle avec n'importe quel outil installé ou service web disponible ayant des formats d'entrée et de sortie bien définis. Snakemake utilise une méthode de reconnaissance de dépendance pour garantir que les étapes sont exécutées dans le bon ordre et que les données nécessaires sont disponibles avant de lancer une étape donnée. Il est facile à utiliser et peut être intégré à de nombreux environnements informatiques. Pour nos analyses nous avons activement utilisé Snakemake comme outil de gestion de workflow.

Nextflow est également un outil populaire pour la création de workflows. Il utilise une syntaxe de programmation déclarative pour définir les étapes du workflow et les dépendances entre elles.

1.3.7. Navigateurs de Génomes

Les navigateurs de génomes sont des outils informatiques utilisés pour explorer et visualiser les données de génome humain d'autres organismes. Ils sont généralement basés sur des bases de données génomiques tel que, les gènes, les régions codantes et les régions régulatrices.

Il existe plusieurs types de navigateurs de génomes, chacun conçu pour répondre à des besoins différents. Les navigateurs de génomes populaires incluent UCSC Genome Browser [147], Ensembl genome browser [87], NCBI Genome Viewer [236] et Gbrowse [74]. Chacun de ces navigateurs propose des fonctionnalités différentes, telles que la possibilité de rechercher des séquences spécifiques, de visualiser des données de séquençage de haute résolution, de comparer des génomes d'organismes différents et de partager des données avec d'autres utilisateurs.

Les fichiers de données génomiques peuvent être organisés pour une visualisation facile et interactive dans un navigateur de génome, sous forme de “trackhub” (Figure 1.25). Il peut contenir des fichiers de séquences, des annotations génomiques, des fichiers ChIP-seq, des fichiers RNA-seq et d’autres types de données génomiques. Les données sont organisées en pistes (track) qui peuvent être activées ou désactivées individuellement pour permettre une visualisation personnalisée des données. Les trackhubs sont créés en téléchargeant les données sur un serveur externe et en les organisant dans un format spécifique qui peut être lu par les navigateurs génomiques. Les utilisateurs peuvent ensuite accéder aux données en utilisant un URL qui pointe vers le trackhub. Les trackhubs sont un moyen pratique de partager les données de génomes, car ils permettent de visualiser les données dans le contexte du génome complet et de comparer les données entre elles.

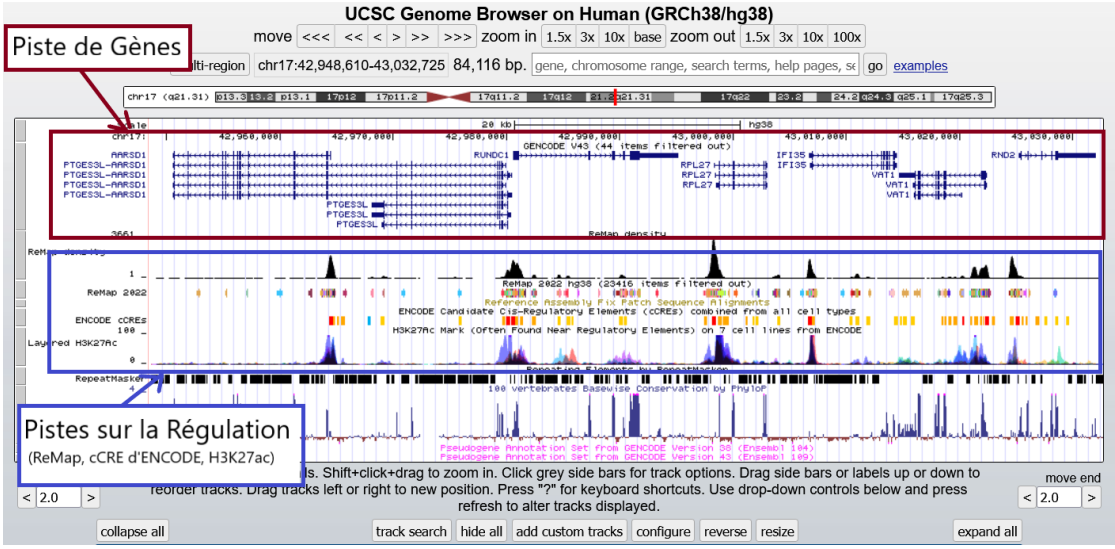


FIGURE 1.25. – ***Exemple de Trackhub.*** *Ce trackhub a été obtenue à partir du Genome Browser de UCSC. Région chr17 :42,948,610-43,032,725 de l’Homme (hg38)* [25].

25. https://genome.ucsc.edu/index.html

1.4. Base de données génomiques et grand consortiums internationaux

1.4.1. Contexte

Au cours de ma thèse, j'ai intégré des données ChIP-seq disponible publiquement pour construire un catalogue de régions régulatrices (chapitre 2) permettant d'annoter les génomes de quatre espèces. Dans cette partie nous décrirons les grands consortiums et bases de données éxistantes pour mieux contextualiser mes résultats.

1.4.2. Le challenge du big data

Depuis l'arrivée des machines hautes performances pour séquencer l'ADN, les scientifiques du monde entier séquencent en continu. L'augmentation de la quantité totale de données générées par séquençage croît de plus en plus rapidement (Figure 1.26). Par ailleurs, la quantité de données de séquençage génétique stockées à l'Institut Européen de Bioinformatique (EMBL-EBI) prend moins d'une année à doubler de taille [190].

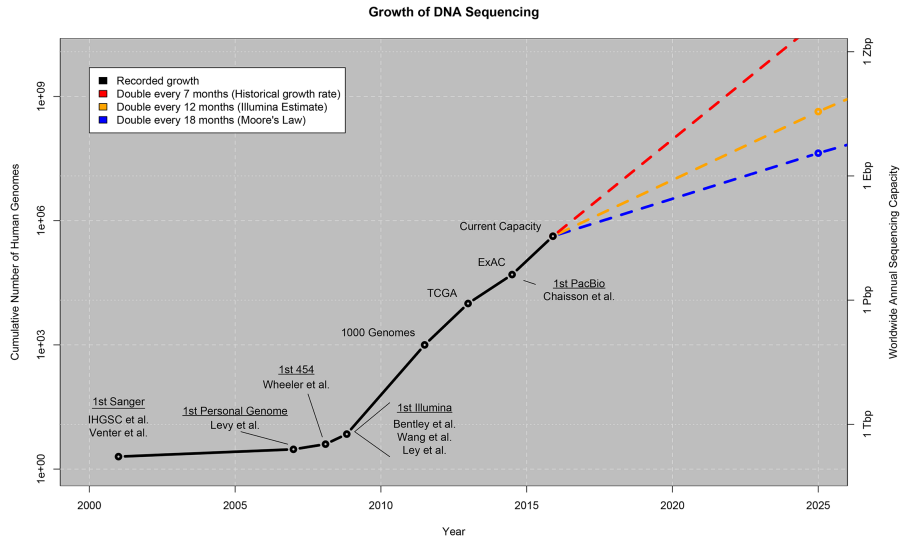


FIGURE 1.26. – ***Croissance du séquençage de l'ADN.*** *Le graphique montre la croissance du séquençage de l'ADN à la fois dans le nombre total de génomes humains séquencés (axe de gauche) ainsi que dans la capacité annuelle de séquençage dans le monde entier (axe de droite : Tera-paires de base (Tpb), Peta-paires de base (Ppb), Exa-paires de base (Epb), Zetta-paires de base (Zpbs)). Les valeurs jusqu'en 2015 sont basées sur les publications historiques, avec des étapes importantes dans le séquençage, ainsi que trois projets exemplaires utilisant le séquençage à grande échelle.Les valeurs au-delà de 2015 représentent une projection selon trois courbes de croissance possibles*[26].

26. Stephens, Zachary D., et al. "Big data : astronomical or genomical?." PLoS biology 13.7 (2015).

Dans ce contexte, le domaine de la génomique est souvent comparé à d'autres grands domaines ou entreprises générant des données et devrait devenir l'un des domaines générateur de données les plus importants d'ici quelques années [278]. Cela crée de sérieux défis pour les scientifiques, car ils doivent analyser de plus en plus de données, qui doivent également être stockées efficacement pour permettre un accès rapide.

D'aprés Kahvejian et al. [142], on pourrait affirmer que le plus grand aspect transformateur du projet du génome humain n'a pas été le séquençage du génome lui-même, mais le développement de nouvelles technologies qui en a résulté, le séquençage à haut débit a radicalement changé la recherche en sciences de la vie.

À mesure que la quantité de données génomiques augmente, il devient nécessaire de trouver des moyens de réutiliser ces données. Au cours des dernières décennies, des initiatives ont été consacrées à rendre les données bioinformatiques susmentionnées accessibles au grand public. Étant donné que le coût de la réalisation d'expériences à grande échelle reste élevé, des consortiums se sont formés pour pallier à ces coûts. L'objectif final est de centraliser les données et d'aider la communauté scientifique à étudier la régulation génomique. Dans cette section, nous présenterons certains de ces efforts.

1.4.3. Archivage des séquences et du génome

L'International Nucleotide Sequence Database Collaboration (National Institutes of Health. "International nucleotide sequence database collaboration.") est une initiative entre l'Institut Européen de Bioinformatique (EMBL-EBI, Union Européenne), le National Center for Biotechnology Information (NCBI, USA) et la DNA Data Band of Japan (DDBJ, Japon) visant à offrir une archive de données d'expériences de séquençage génomique à haut débit brutes et leurs métadonnées, accessibles via des archives telles que SRA. Les assemblages de génomes eux-mêmes sont gérés par le Genome Reference Consortium. L'assemblage actuel du génome humain, GRCh38, a été publié en décembre 2013 et a été depuis, corrigé pour des erreurs d'assemblage (problèmes de read court) [257] ainsi que pour l'ajout d'haplotypes alternatifs. Le T2T consortium est à l'origine d'un projet ambitieux visant à produire des assemblages de génome de haute qualité pour de nombreuses espèces, en utilisant les technologies les plus avancées de séquençage et d'assemblage. Il repose sur une approche de séquençage en longueur qui permet de produire des lectures de plusieurs kilobases, permettant ainsi de franchir les zones répétitives et les régions centromériques, souvent difficiles à assembler avec les techniques classiques. L'objectif est de produire des génomes complets, avec des séquences nucléotidiques précises et bien ordonnées, permettant une analyse plus fine des variations génétiques, des régions régulatrices et des éléments transposables. Le projet T2T a déjà produit d'un assemblage de qualité supérieure pour l'homme (T2T-CHM13) [216]. Les données sont accessibles publiquement via la bases de données GenBank.

1.4.4. Annotation des éléments *cis*-régulateurs

Le projet ENCyclopedia of DNA Elements (ENCODE) a été lancé en 2003 avec l'objectif ambitieux d'identifier tous les éléments fonctionnels de la séquence du génome humain [274]. Il a également permis de rendre disponible, de regrouper et de retraiter les données en les soumettant à un protocole de contrôle qualité normalisé afin d'étudier la régulation génomique. Leur objectif est de constituer une encyclopédie complète des sites de fixation TF, des marques d'histone, et plus généralement d'étudier les marqueurs de la chromatine. Le projet s'est déroulé en 4 phases dont la première (phase pilote) s'est déroulée jusqu'en 2007 pour identifier les méthodes les plus prometteuses. En effet, ENCODE a été à l'origine du développement de nombreux outils et méthodes bioinformatiques. Dans cette phase, les régions liées par plus de 150 différentes protéines régulatrices ont été identifiées dans 1% de la séquence d'ADN humain. Le catalogue ENCODE donne une liste de 1,3 millions d'éléments *cis*-regulateurs (CRE) putatifs pour 600 types cellulaires [274] sur la base de leurs données centralisées. On pourrait également citer GENCODE [93], un sous-projet d'identification et de classification des gènes. Ses annotations proviennent principalement d'Ensembl [61].

Le projet FANTOM (Functional Annotation of the Mammalian Genome) porte sur l'étude du transcriptome. Dans sa troisième phase, FANTOM a développé la méthode CAGE pour étudier l'initiation de la transcription et les promoteurs, en se concentrant sur l'extrémité 5' de l'ARNm mature. La phase 5 a permis de construire des "atlas" de promoteurs, d'enhancers, miRNAs et d'ARN longs non codants (lncRNA). La phase la plus récente (FANTOM6) s'est concentrée sur l'annotation des lncRNA [5].

ENCODE n'est pas le seul groupe à se concentrer sur l'annotation épigénomique. En effet, le projet BLUEPRINT [6] a été lancé en 2011 et terminé en 2016. Ce projet avait pour but d'étudier l'épigéome des cellules sanguines. Ces données sont devenues une ressource importante pour la recherche en épigénétique et en génomique des cellules sanguines, offrant la possibilité d'approfondir notre compréhension des mécanismes de régulation de l'expression génique et de la biologie des cellules sanguines.

1.4.5. Autres bases de données biologiques

Un grand nombre de bases de données biologiques, d'outils, de ressources et de services Web sont disponibles grâce aux données provenant d'expériences en sciences de la vie de différents laboratoires et centres de recherche du monde entier. Ces informations ont été systématiquement collectées et organisées dans un référentiel spécifique en fonction de la nature des données.

D'autres initiatives peuvent être mentionnées ici. Pour la publication d'un article scientifique, les données doivent être rendues accessibles. L'archivage de données expérimentales pour microarray et NGS est proposé par GEO (NCBI) et ArrayExpress (EBI). Les données brutes de séquençage peuvent être stockées au SRA et à l'ENA. Malheureusement, contrairement à des projets comme ENCODE, l'annotation et le traitement ne sont pas uniformes. Par ailleurs, Gene Ontology (GO) a développé une nomenclature applicable à tous les eucaryotes pour décrire les fonctions des gènes, sous forme de graphe acyclique, une approche également utilisée par KEGG. Quelques exemples de telles bases de données sont listés ci-dessous :

- Genes/Proteines : Ensembl [61], EMBL [146], GenBank [26], EntrezGene [186], UniProt [57], RefSeq [234], HGNC [293], UCSC [208], KEGG [144], GeneCards [247].
- Maladie : OMIM [9].
- Structures des protéines : PDB [43].
- Caractéristiques des protéines : Pfam [203], InterPro [227].
- Pathways : KEGG [145], PANTHER [287], Reactome [103].
- Ontologies : GeneOntology (GO) [56].
- Literature : PubMed/MEDLINE [309].
- Archivage : GEO [51], ArrayExpress [155].

Ces ressources sont maintenues et partagées par des organisations telles que l'Institut européen de bioinformatique (EMBL-EBI [1]) et le National Center for Biotechnology Information (NCBI [2]) pour faire progresser la science et la santé en donnant accès à ces ressources d'information biomédicales et génomiques aux chercheurs du milieu universitaire et de l'industrie.

Cette vaste quantité de connaissances publiques est très utile. Le principal défi des ressources de données publiques est la croissance exponentielle et ces ressources sont hétérogènes tant en contenu qu'en format.

1. https://www.ebi.ac.uk
2. https://www.ncbi.nlm.nih.gov

1.5. Les éléments transposables dans le contexte de la régulation

1.5.1. Contexte

Lors de la deuxième partie de ma thèse, j'ai étudié le rôle des éléments transposables dans la régulation de l'expression des gènes. Mon étude a porté sur l'implication des TEs dans l'insertion de TFBS dans les génomes humain, et a couvert plusieurs niveaux de la classification des TE. Dans ce chapitre, nous allons donc présenter les TE ainsi que leur classification, afin d'expliquer leur diversité et leur complexité. Ensuite, nous aborderons l'état de l'art de la recherche sur l'implication des TE dans la régulation, afin de faciliter la compréhension de mes travaux dans le chapitre 3.

1.5.2. Découverte des éléments transposables

Nous savons maintenant que les gènes ne constituent qu'une petite partie des génomes (environ 5% du génome humain [216] et que le reste a été regroupé dans le terme fourre-tout de "séquences non codantes" ou, pour les plus pessimistes, "junk DNA". Barbara McClintock, dans ses expériences fondatrices sur la cassure chromosomique du maïs [195], et dans son interprétation visionnaire et imaginative des résultats, a découvert l'existence au sein de cet ADN indésirable "d'éléments de contrôle" qui, par leur mouvement, influencent l'expression des gènes.

Le séquençage du génome entier nous a permis d'obtenir une image plus complète de la composition d'un génome, par la découverte d'une partie importante des génomes : les "gènes sauteurs" ou éléments transposables. Ces TE, composent plus de 50% du génome chez l'Homme [2]. Ces TE sont passés de déchets génomiques parasites à reconnus comme de puissants moteurs de l'évolution.

1.5.3. Classification des éléments répétés

Les séquences répétés sont l'un des principaux moteurs de la variation de la taille du génome chez les eucaryotes. Ils représentent un ensemble hétérogène comprenant des dizaines de familles, qui varient en longueur de motif, en nombre d'exemplaires ou en structure globale. Il existe plusieurs types de séquences répétitives présentes dans les génomes eucaryotes représentés par deux groupes principaux : les short tandem repeat et les éléments transposables (TE).

1.5.4. Short tandem repeats

Les short tandem repeat (STR) sont très abondants dans les génomes eucaryotes complexes. Ils sont constitués d'unités répétées organisées en tandem appelées monomères. La longueur du monomère peut varier de 2 à plusieurs centaines de nucléotides, ils forment généralement de longs réseaux contenant des milliers de copies. Ils en composent environ 8% du génome humain[77]. Les copies de monomères ne sont pas entièrement identiques et présentent un polymorphisme de séquence. En fonction de la longueur du monomère et de la taille du réseau, les STR sont classés en trois groupes : les microsatellites avec une longueur de monomère < 9 nucléotides et une taille de réseau < 1 kb, les minisatellites avec une longueur de monomère comprise entre 10 et 100 pb et les ADN satellite (satDNA) ayant des monomères de plus de 100 pb et forme souvent des réseaux de plus de 100 Mb. La représentation schématique d'un STR est illustrée à la Figure 1.27.

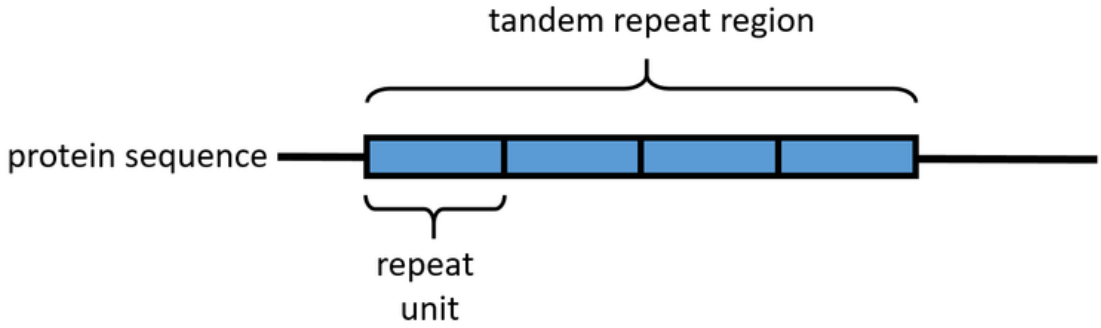


FIGURE 1.27. – ***Représentation schématique d'un STR.***[27].

Différentes familles de TR peuvent être présentes dans une espèce, il en existe 9 dans le génome humain. Bien que les TR aient été initialement considérés comme de l'ADN non fonctionnel, nous savons aujourd'hui qu'ils ont de nombreuses fonctions dans le génome. Les TR sont impliqués dans l'organisation chromosomique, le contrôle de l'allongement des télomères, la réponse transcriptionnelle au stress ou la modulation de l'expression des gènes [84].

27. https://www.wikiwand.com/en/Protein_tandem_repeats

1.5.5. Les éléments transposables

Les TE, également appelés gènes sauteurs, ont été découverts pour la première fois dans les années 1940 par la généticienne Barbara McClintock [193]. Les TE sont dispersés dans le génome à divers endroits. Ils ont la capacité de se déplacer ou même de se copier d'un emplacement génomique à un autre, ce qui peut entraîner leur amplification rapide dans le génome. Les TE peuvent se reproduire en centaines, voire en milliers d'exemplaires. En raison de leur nature répétitive et de leur variabilité, ils sont difficiles à analyser et restent un défi majeur dans le domaine de la bioinformatique.

Au début, ils étaient considérés comme de l'ADN indésirable sans aucune fonction et étaient ignorés par de nombreux chercheurs. De nos jours, on sait que les TE ont de nombreux rôles : ils affectent la taille du génome , jouent un rôle essentiel dans les réarrangements chromosomiques, ou ont été des acteurs cruciaux dans l'évolution du génome [34]. Les TE composent une partie importante des génomes eucaryotes et ont été trouvés dans presque tous les organismes étudiés jusqu'à présent. Par exemple, ils occupent 37% du génome de la souris [1], environ 50% du génome humain [2] et environ 80% du génome du maïs [256]. De nombreux types différents de TE ont été découverts depuis la découverte de McClintock. Les TE sont divisés en deux grandes classes selon qu'ils se transposent via un intermédiaire ARN, classe I (rétrotransposons), ou un intermédiaire ADN, classe II (transposons ADN) (Figure 1.28).

Ces classes sont en outre subdivisées en plusieurs sous-classes en ce qui concerne leur mécanisme d'intégration chromosomique. Une autre classification des éléments de classe I et de classe II est basée sur le fait que les TE codent ou non tous les domaines nécessaires à leur transposition et les divisent en éléments autonomes et non autonomes, respectivement. Les non-autonomes peuvent survenir de diverses manières, par exemple en dérivant de copies autonomes qui ont rassemblé des mutations, ainsi, ils n'encodent plus les domaines nécessaires à la transposition. Dans les parties suivantes, la structure et les caractéristiques des TE de classe I et de classe II seront décrites sur la base de Feschotte et al. (2007) [91], Muñoz-Lòpez et al. (2010) [206] et Wicker et al. (2007) [311].

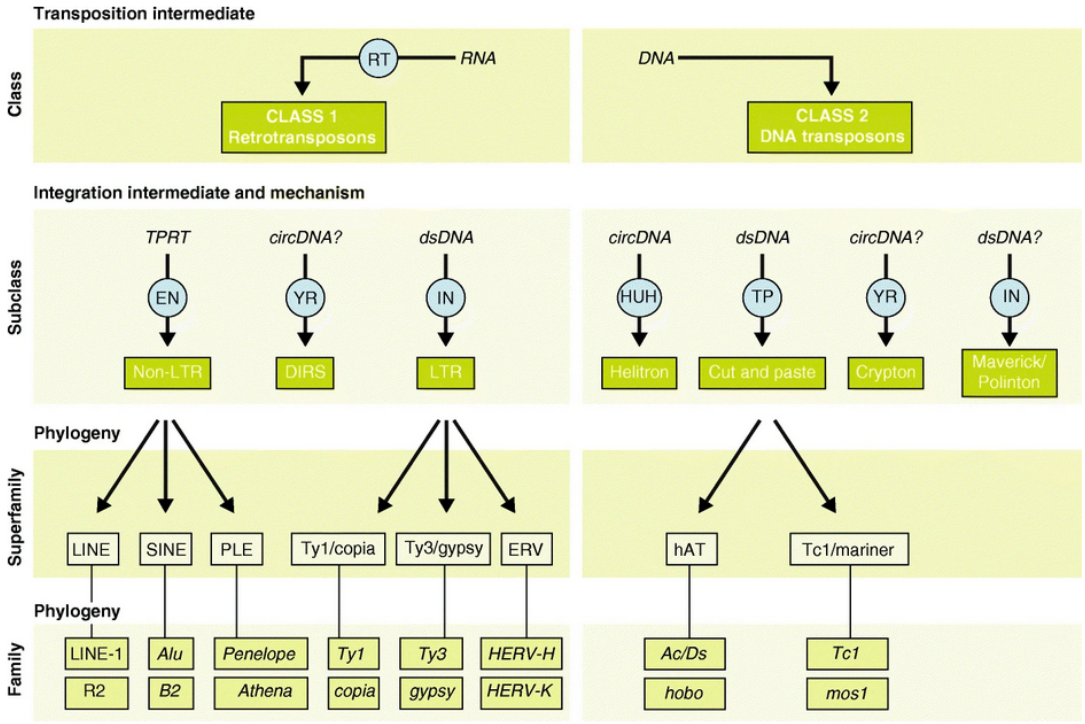


FIGURE 1.28. – ***Classification des éléments transposables eucaryotes.*** *Schéma et exemples montrant les principales caractéristiques et relations entre les classes TE, les sous-classes, les superfamilles et les familles. Les cercles bleus représentent les enzymes codées par TE. Intermédiaire d'ADN circulaire circDNA, DIRS Dictyostelium séquence répétitive, intermédiaire d'ADN double brin linéaire d'ADNdb, endonucléase EN, intégrase IN, PLEs Éléments de type pénélope, HUH, protéine Rep/Hélicase avec activité d'endonucléase HUH, transcriptase inverse RT, transposase TP, cible TPRT transcription inverse amorcée, YR tyrosine recombinase.*[28].

28. Bourque, Guillaume, et al. "Ten things you should know about transposable elements." Genome biology 19.1 (2018).

1.5.5.1. Class I - Retrotransposons

Les éléments inclus dans cette classe se transposent via un ARN intermédiaire lorsque l'ARN intermédiaire est transcrit à partir d'une copie génomique suivie d'une transcription inverse en ADN par la transcriptase inverse (RT) codée dans le TE. Ce mécanisme est généralement appelé "copier-coller" car chaque cycle de réplication complet génère une nouvelle copie du TE. La classification donnée par Wicker et al. [311] divise les rétrotransposons en cinq groupes en fonction de leurs caractéristiques, de leur organisation et de la phylogénie de la transcriptase inverse : les rétrotransposons LTR, les éléments de type DIRS, les éléments de type Penelope, les LINE et les SINE. La structure des éléments appartenant à ces groupes est illustrée à la Figure 1.29.

Les rétrotransposons LTR sont composés de longues répétitions terminales (LTR), qui renferment les domaines protéiques codant pour le corps du rétrotransposon. La longueur des LTR varie de quelques centaines de pb à 6 kb, et ils commencent par **5'-TA-3'** et se terminent par **5'-CA-3'**. Les rétrotransposons LTR contiennent généralement deux cadres de lecture ouverte (ORF) codant pour des protéines, **gag** et **pol**, mais à titre exceptionnel, des ORF supplémentaires de fonction inconnue peuvent être présents. La région pol code plusieurs domaines protéiques (RT, protéase, RNase H et intégrase), qui effectuent la transcription inverse et l'intégration dans un nouvel emplacement du génome. Après intégration, ils génèrent une duplication du site cible de longueur 4-6 pb. La longueur totale de ces éléments peut atteindre de manière surprenante 25 kb.

Les séquences répétées intermédiaires de Dictyostelium (DIRS) codent pour un domaine tyrosine recombinase au lieu de l'intégrase et ne produisent donc pas de duplication du site cible. Les éléments de ce groupe possèdent des séquences terminales qui ressemblent à des repeats inversés ou à des repeats directs fractionnés. Il existe également les éléments de type pénélope (PLE) qui ne codent que pour deux domaines protéiques, la reverse transcriptase et endonucléase. Leurs répétitions peuvent être en orientation directe ou inverse.

Les éléments nucléaires intercalés longs (LINE) peuvent être longs de plusieurs kb, contenir un polORF codant au moins la reverse transcriptase et une nucléase (endonucléase ou un endonucléase apurique ou apyrimidique). Un ORF de type bâillon et de fonction inconnue est parfois trouvé à l'extrémité 5 ' de la région pol. La région codante peut être flanquée de régions UTR de chaque côté de l'élément. Les LINE produisent des duplications du site cible, mais elles sont difficiles à trouver en raison des extrémités 5' tronquées. Leur extrémité 3 'peut contenir une queue poly (A), une TR ou une région riche en A.

Enfin, les éléments nucléaires intercalés courts (SINE) ne sont pas autonomes et proviennent de la rétrotransposition accidentelle de divers transcrits de polymérases III. Ils dépendent des LINE partenaires car ils utilisent le domaine reverse transcriptase des LINE pour leur transcription inverse. Ces éléments sont relativement courts, compris entre 80 et 500 pb, et génèrent des duplications du site cible de 5 à 15 pb. Les SINE se terminent par une région riche en A ou AT ou par une queue poly(T).

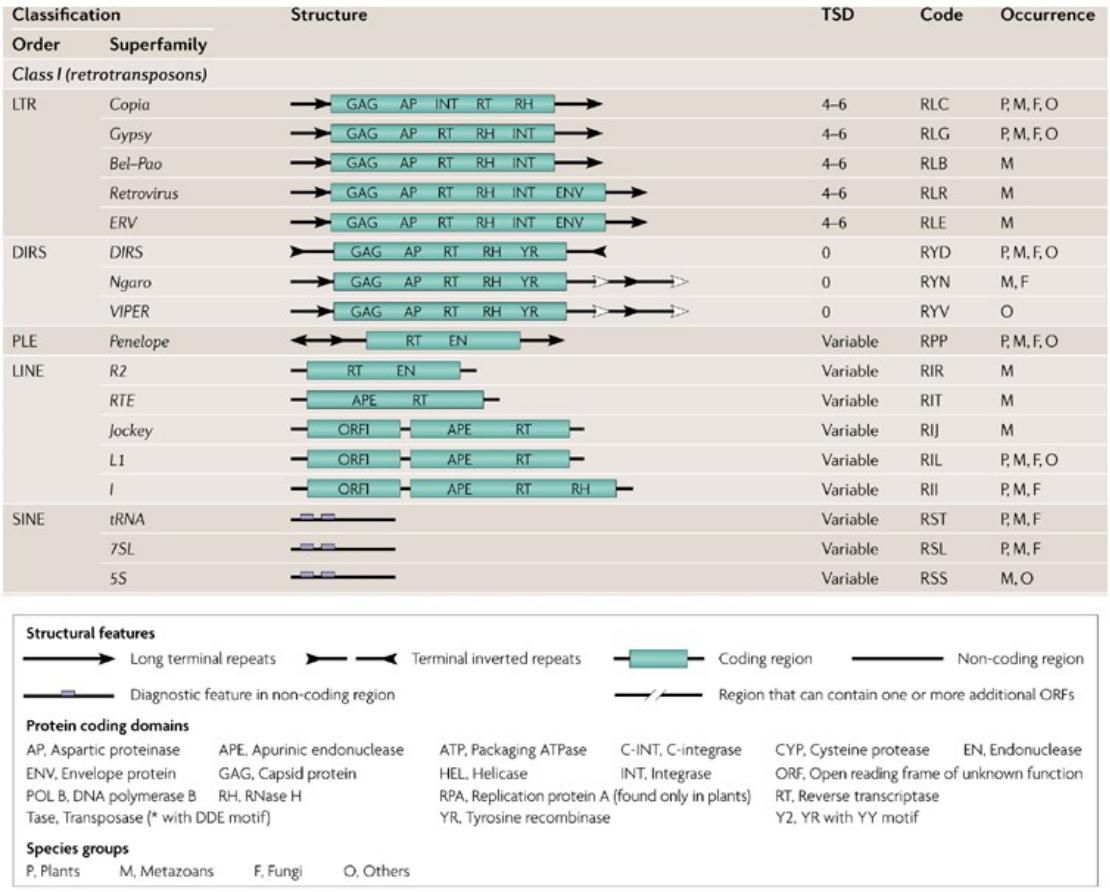


FIGURE 1.29. – ***Système de classification proposé pour les TE de Classe I.*** *Schéma représentant la structure de chaque superfamille de TE de la classe I.*[29].

29. Wicker, Thomas, et al. "A unified classification system for eukaryotic transposable elements." Nature Reviews Genetics 8.12 (2007).

1.5.5.2. Class II - Transposons d'ADN

Les transposons d'ADN se transposent généralement dans le génome par un mécanisme de "couper-coller" utilisant un intermédiaire d'ADN, mais il existe quelques exceptions. Le transposon d'ADN est découpé de l'emplacement chromosomique et est ensuite réinséré dans un nouvel emplacement chromosomique. Étant donné que la plupart des transposons d'ADN se déplacent par un mécanisme non réplicatif (ne génèrent pas de copies d'eux-mêmes), ils se produisent généralement en faible nombre de copies. Selon la classification proposée par Wicker et al. [311], les transposons d'ADN comprennent deux sous-classes basées sur le nombre de brins d'ADN coupés lors de la transposition : la sous-classe I et la sous-classe II. La sous-classe I implique deux ordres d'éléments, TIR et Crypton dont la structure est illustrée à la Figure 1.30.

Les transposons d'ADN TIR sont caractérisés par des répétitions inverses terminales (TIR) (de longueur variable) et la TPase entourée par ces TIR. Dans le processus de transposition, les éléments sont découpés à partir d'un emplacement actuel dans le génome et réintégrés dans un nouvel emplacement chromosomique sous forme d'ADN double brin. Ce processus est contrôlé par la TPase encodée dans l'élément. Les TIR sont capables d'augmenter leur nombre de copies en se déplaçant lors de la réplication chromosomique lorsqu'ils se transposent d'une position déjà répliquée à une autre position que la fourche de réplication n'a pas encore franchie [311]. Ils se distinguent en neuf superfamilles, Tc1/mariner, PIF/Harbinger, hAT, Mutator, Merlin, Transib, P, piggyBac et CACTA, selon les séquences TIR et la taille duplication du site cible. Les TIR varient en longueur entre 2 kb et 15 kb, et certaines superfamilles possèdent le deuxième ORF de fonction inconnue.

Les transposons d'ADN crypton n'ont été trouvés que chez les champignons et sont mal connus. Ils sont composés de tyrosine recombinase, manquent de TIR et semblent générer des duplications du site cible. Leur transposition nécessite également de couper les deux brins d'ADN. La sous-classe II se compose de transposons d'ADN appelés Helitrons et Mavericks qui utilisent un processus de transposition nécessitant une réplication sans clivage des deux brins d'ADN et une transposition par le mécanisme "copier-coller". Leur structure est illustrée à la Figure 1.30.

Les hélitrons se répliquent par un mécanisme de cercle roulant lorsqu'un seul brin d'ADN est coupé. Ils codent pour la tyrosine recombinase de type Y2 avec un domaine hélicaseet un motif initiateur de réplication. Les hélitrons ne génèrent pas de duplication du site cible et leurs extrémités peuvent être déterminées par des motifs TC ou CTRR (R est la purine) et une structure en épingle à cheveux courte avant l'extrémité 3 '.

Enfin les transposons d'ADN comprennent également des éléments non autonomes connus sous le nom d'éléments transposables miniatures à répétition inversée (MITE), qui ne codent pas pour les protéines et n'ont aucun potentiel de codage. Par conséquent, leur transposition dépend vraisemblablement de transposons autonomes. Ils sont largement distribués chez les eucaryotes en particulier les plantes [90]. Les MITE sont généralement courts et leur longueur varie entre 50 et 800 pb. Ils contiennent de courts TIR conservés flanqués de duplication du site cible, qui sont des caractéristiques communes des transposons d'ADN . Les MITE sont classés en superfamilles en fonction de la composition de leurs TIR et de la longueur des duplications du site cible. Ils sont généralement situés dans des régions riches en gènes et présentés en nombre élevé de copies [206].

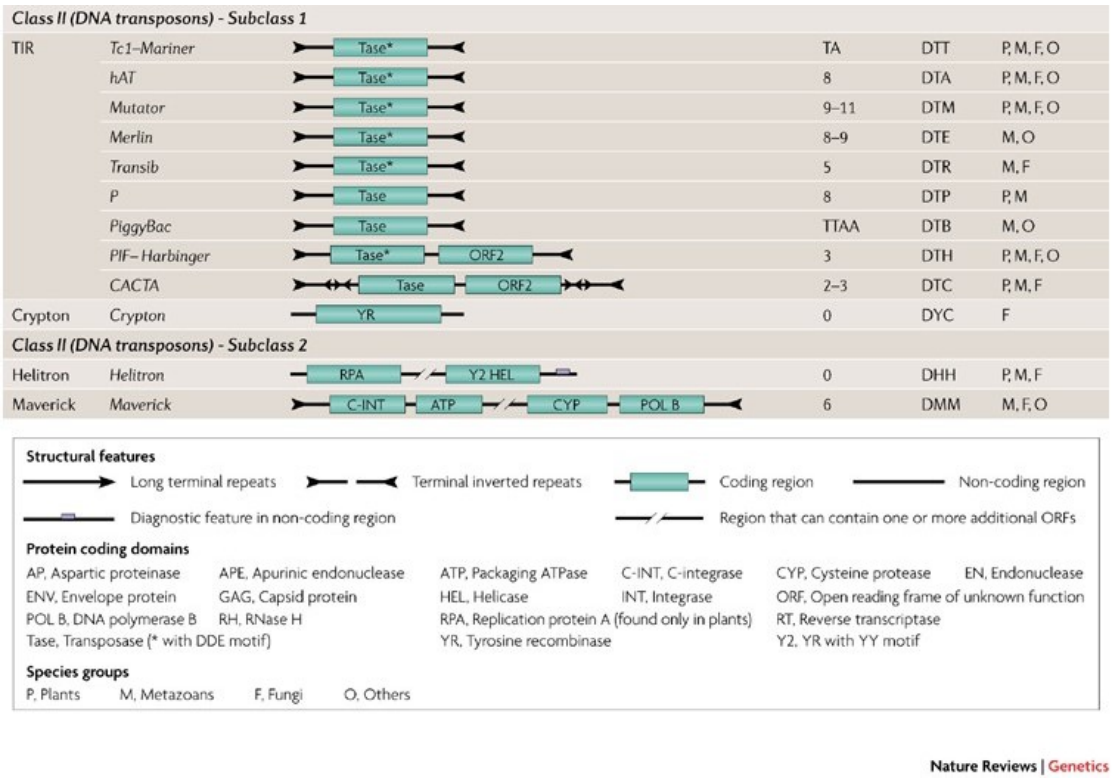


FIGURE 1.30. – ***Système de classification proposé pour les TE de Classe II.*** *Schéma représentant la structure de chaque superfamille de TE de la classe II.* [30].

30. Wicker, Thomas, et al. "A unified classification system for eukaryotic transposable elements." Nature Reviews Genetics 8.12 (2007).

1.5.6. Implication des TE dans la régulation

Les TE ont été initialement étudiés pour leur impact sur l'expression du gène de l'hôte en raison des effets destructeurs que les nouvelles insertions de TE ont sur les gènes de l'hôte ou les éléments régulateurs. Néanmoins, les TE fournissent en fait une source abondante d'éléments *cis*-régulateurs pour les génomes hôtes [89] comme des promoteurs [140] et des enhancers [50].

Il existe de nombreuses façons par lesquelles les TE peuvent influencer directement l'expression d'un gène voisin au niveau transcriptionnel. Pour donner quelques exemples, des travaux [89] révèlent qu'au moins 16% des éléments non codants conservés spécifiques euthériens sont dérivés des TE. Il a également été montré que des séquences TE étaient présentes dans 25% des promoteurs humains, ces résultats ont validé expérimentalement [140]. Conley et al. [52] ont révèlé que le génome humain contient plus de 50 000 séquences dérivées d'ERV qui ont été identifiées comme étant à l'origine de la transcription.

L'implication des TE est caractérisé par leur contribution dans l'insertion à de nombreux nouveaux sites de fixation aux facteurs de transcription chez les mammifères, y compris des sites importants pour le développement embryonnaire [35], la liaison à la protéine suppresseur de tumeur p53 [304] et des facteurs de remodelage génomique tels que le CTCF [253]. La dispersion des familles TE dans tout le génome permet aux motifs d'être engagés à de nombreux emplacements chromosomiques et, par conséquent, entraîne l'introduction de plusieurs gènes dans les mêmes réseaux de régulation. Par exemple, une étude sur le génome humain [304] suggère qu'un ensemble de familles étroitement liées de LTR ont dispersé plus de 1500 sites de fixation pour le facteur régulateur maître p53. Ces sites englobent 30% de tous les sites de fixation p53 qui ont été cartographiés. La classe ERV des TE est particulièrement riche en sites de fixation pour les TF importants dans la nature pluripotente des cellules souches [130]. Kunarso et al. montrent que les TE sont responsables de 25% des TFBS pour trois TF clés (POU5F1, NANOG et CTCF) (Figure 1.31) [162].

En 2009 Wang et al. ont également prouvé la présence d'un enrichissement en TFBS de c-Myc dans les TE [303]. Une étude plus grande a également été réalisée sur 26 TF et deux lignées cellulaires chez l'humain et la souris révèle un enrichissement significatif pour 710 couples TE-TF [283]. Bien que la contribution des TE à la variation des boucles de la chromatine ne soit pas entièrement comprise, une grande partie des sites de fixation du CTCF chez les mammifères relèvent des TE. Des études récentes montrent que les boucles variables peuvent s'expliquer par la divergence dans la fixation des CTCF dérivée de TE [253].

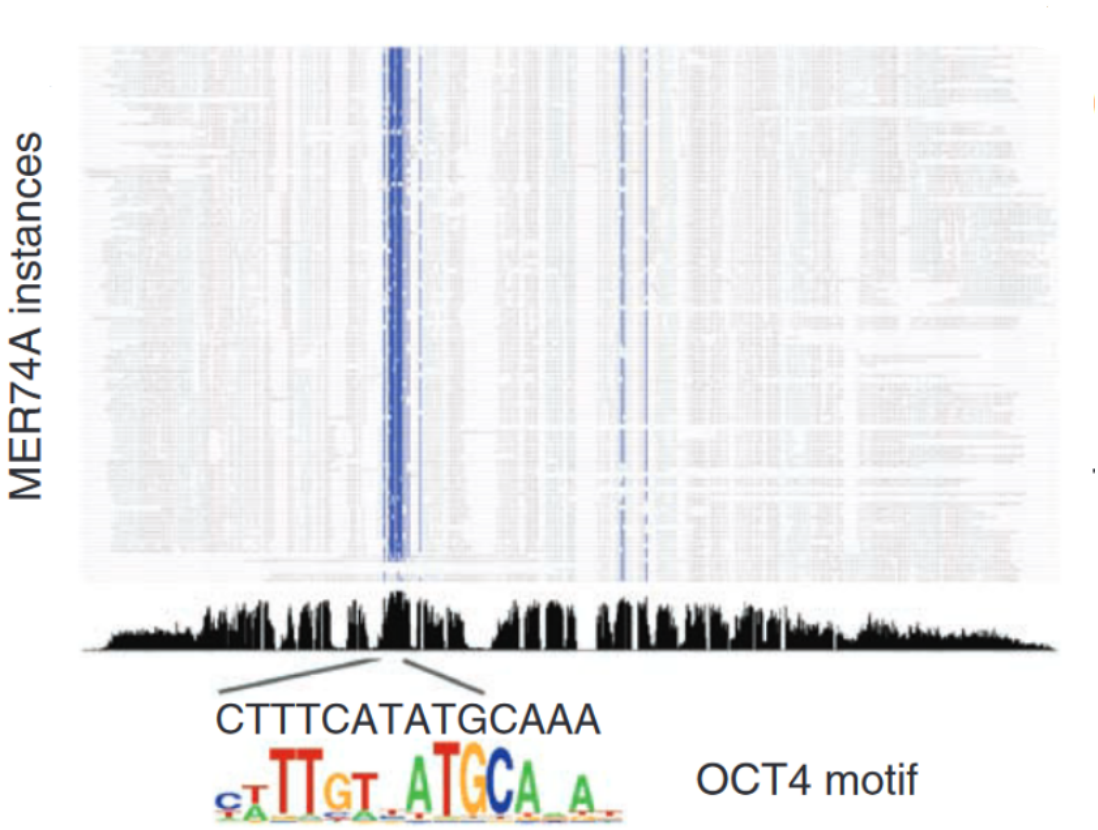


FIGURE 1.31. – ***Alignement de séquences multiples des instances de MER74A liées par OCT4.*** *Le graphique affiché sur l'axe des x montre le pourcentage d'identité de chaque colonne de l'alignement. Les colonnes avec plus de 70% d'identité sont en bleu et mettent en évidence une région de similarité de séquence plus élevée. Le consensus (séquence ancestrale) de la répétition à cette position correspond bien au motif de liaison OCT4*[31].

A la lumière de ces avancées sur la connaissance des TE au niveau de la régulation , je vous expose mes travaux qui s'articulent autour de la fixation des TF sur sur les séquences des TE dans le chapitre 3.

31. Kunarso, G., Chia, N. Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y. S., ... & Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells.

1.6. La régulation des ALDH pour le traitement des AML

1.6.1. Contexte

Ma thèse est co-financée Inserm/Région, impliquant un acteur académique et sociétal Advanced BioDesign (ABD). Ce partenariat a permis de réaliser un projet de recherche en collaboration avec ABD.

Advanced BioDesign est une société de biotechnologie de stade préclinique fondée en 2010 et axée sur le développement de nouveaux médicaments anticancéreux, en se concentrant spécifiquement sur l'inhibition de l'ALDH comme approche thérapeutique pour la leucémie myéloïde aiguë et d'autres cancers dont les besoins médicaux non satisfaits sont importants. La stratégie innovante de ABD vise à surmonter la chimiorésistance des tumeurs et à restaurer ou activer les mécanismes endogènes de mort cellulaire dans les cellules cancéreuses. Leur principal candidat-médicament : ABD-3001 est une petite molécule inhibitrice de la famille des enzymes ALDH qui est au essai clinique de phase I. Le composé est envisagé pour être évalué dans pour un large éventail de cancers, notamment la leucémie myéloïde aiguë (AML), le cancer du poumon à petites cellules et le cancer du sein.

Dans le cadre de l'éssai clinique de phase I "Première administration à l'homme" pour ABD3001 (ODYSSEY I) qui a commencé en 2023, le composé est actuellement administré à des patients adultes atteints d'AML. Des données pharmacocinétiques et pharmacodynamiques sont actuellement collectées, ce qui permettra de déterminer le dosage et le schéma d'administration pour l'évaluation de l'efficacité dans les prochains essais de phase II.

Dans ce contexte nous nous sommes efforcés d'identifier les régions régulatrices du gène ALDH1A1 (Chapitre 4). Je vous présente donc par la suite l'état de la recherche sur les ALDH dans les leucémies.

1.6.2. La leucémie et ses sous-types

La leucémie est une maladie maligne caractérisée par une croissance accélérée de cellules sanguines anormales ou blastes qui s'accumulent dans la moelle osseuse et entravent le développement normal des cellules sanguines. En général, la leucémie peut être classée en quatre groupes principaux : CML, AML, LLC et ALL (1.32 [49].

Table 1. Characteristics of Major Subtypes of Leukemia

Subtype	*Description*	*Typical group(s) affected*	*Common presenting features*	*Five-year relative survival rate**
Acute lymphoblastic leukemia	Blast cells on peripheral blood smear or bone marrow aspirate	Children and young adults (53% of new cases occur in persons < 20 years)	Symptoms: fever, lethargy, bleeding, musculoskeletal pain or dysfunction Signs: hepatosplenomegaly and lymphadenopathy	< 50 years: 75% ≥ 50 years: 25%
Acute myelogenous leukemia	Blast cells on peripheral blood smear or bone marrow aspirate; Auer rods on peripheral smear	Adults (accounts for 80% of acute leukemia in adults)	Symptoms: fever, fatigue, weight loss, bleeding or bruising Signs: hepatosplenomegaly and lymphadenopathy (rare)	< 50 years: 55% ≥ 50 years: 14%
Chronic lymphocytic leukemia	Clonal expansion of at least 5,000 B lymphocytes per μL (5.0×10^9 per L) in the peripheral blood	Older adults (85% of new cases occur in persons > 65 years)	Symptoms: 50% of patients are asymptomatic Signs: hepatosplenomegaly and lymphadenopathy	< 50 years: 94% ≥ 50 years: 83%
Chronic myelogenous leukemia	Philadelphia chromosome (*BCR-ABL1* fusion gene)	Adults	Symptoms: 20% of patients are asymptomatic Signs: splenomegaly	< 50 years: 84% ≥ 50 years: 48%

**—Relative survival compares a cohort of leukemia survivors (diagnosis made in 2005) to a similar cohort of cancer-free individuals.*

Information from references 1, and 9 through 18.

FIGURE 1.32. – *Charactéristiques des sous-types majeur de la leucémie.*[32].

32. Davis AS, Viera AJ, Mead MD. Leukemia : an overview for primary care. Am Fam Physician. 2014.

1.6.3. La leucémie aigüe myéloïde

La leucémies aiguës myéloïdes est une tumeur hématopoïétique qui résulte de l'expansion rapide de cellules progénitrices myéloïdes dont la différenciation est arrêtée. Elle a souvent un mauvais pronostic. Seuls deux patients adultes sur trois diagnostiqués avec une AML peuvent obtenir une rémission avec le meilleur traitement possible, généralement une chimiothérapie avec ou sans greffe de moelle osseuse. La survie à cinq ans pour la AML de novo diagnostiquée chez l'adulte est inférieure à 40% [217]. Les AML ont une incidence stable au cours du temps qui varie de 2,5 à 3,5/100 000 habitants/an dans les pays occidentaux. Non seulement nos options thérapeutiques sont insuffisantes, mais bon nombre de nos schémas thérapeutiques sont également dépassés. Le schéma d'induction le plus courant, consistant en sept jours de cytarabine suivis de trois jours d'une anthracycline, le plus souvent la daunorubicine, a été lancé il y a plus de trois décennies avec seulement des révisions mineures depuis [73] .

Des anomalies génétiques surviennent lors de la prolifération et de la différenciation des cellules souches hématopoïétiques et des progéniteurs telles que : mutations, altérations des copies chromosomiques et translocations. Ces anomalies peuvent coopérer pour modifier la différenciation, la mort cellulaire, la prolifération et l'autorenouvellement entraînant des quantités élevées de progéniteurs anormaux dans le sang périphérique, la moelle osseuse et l'infiltration des tissus caractéristiques de la AML. Les symptômes comprennent l'anémie, la fatigue, les saignements et la susceptibilité aux infections [182]. La AML est une maladie hétérogène, il est donc essentiel que les patients soient mieux classés avant le traitement [109].

À l'origine, le système franco-américain-britannique (FAB) définissait la leucémie comme des cas ayant plus de 30% de blastes dans la moelle osseuse, l'AML a ensuite été catégorisée en utilisant la morphologie et la cytochimie cellulaires [24].

TABLEAU 1.1. – *Number of patients according to the FAB classification*[1].

Subtype	Description	No. of patients	%
M0	Minimally differentiated AML	30	4.7
M1	AML without maturation	109	17.1
M2	AML with maturation	261	40.9
M4	Acute myelomonocytic leukemia (AMMoL)	148	23.2
M4Eo	AMMoL with eosinophils	23	3.6
M5a	Acute monoblastic leukemia	19	3.0
M5b	Acute monocytic leukemia	24	3.8
M6	Acute erythroleukemia	16	2.5
M7	Acute megakaryoblastic leukemia	5	0.8
Acute leukemia of ambiguous lineage		3	0.5
Total		638	100

Plus récemment, cette classification a été modifié par l'Organisation mondiale de la santé (OMS) [298]. Désormais, un minimum de 20% de blastes dans la moelle osseuse est suffisant pour le diagnostic de la AML et les méthodes de stratification utilisent en outre la cytogénétique et les mutations génétiques [298]. La AML peut également survenir en tant que tumeur maligne secondaire chez les patients atteints de néoplasmes non apparentés. Le traitement par agent alkylant d'autres cancers, comme le cancer du sein, peut stimuler la AML [297]. De plus, la AML peut évoluer à partir d'autres maladies du sang telles que le syndrome myélo-dysplasique (SMD), il a également été suggéré que la prédisposition à la AML peut être favorisée par des polymorphismes germinaux impliquant le gène CEBPα [270].

Heureusement, des progrès sont en cours. Bien que la recherche biomédicale récente sur la AML n'ait pas produit d'avancées révolutionnaires dans la pharmacothérapie dirigée contre la AML, elle a considérablement élargi notre compréhension des processus physiopathologiques qui entraînent l'établissement et la progression de la AML. En particulier, l'avènement de la génomique a révélé pour la première fois les événements génétiques qui sous-tendent la AML, identifiant les altérations récurrentes les plus courantes de la AML et façonnant notre compréhension des étapes génétiques nécessaires à la progression vers la maladie, chacune pouvant, en théorie représenter une cible thérapeutique [102]. L'avènement des technologies de séquençage a également ouvert le champ connexe de la génomique fonctionnelle par l'annotation de la fonction des gènes et des protéines par des approches omiques. Plus précisément, l'intégration des technologies de séquençage avec des expériences permettant d'étudier un gain fonctionnel et une perte de fonction. En principe, ces techniques ont permis une meilleure compréhension de la mécanique de la maladie.

1. Wakui, Moe, et al. "Diagnosis of acute myeloid leukemia according to the WHO classification in the Japan Adult Leukemia Study Group AML-97 protocol." International journal of hematology 87.2 (2008).

The European LeukemiaNet (ELN) a déjà publié des lignes directrices pour le diagnostic et la prise en charge de la leucémie myéloïde aiguë (LMA) en 2010 et 2017 [71]. En raison des progrès majeurs dans la classification de la LMA, les diagnostics génomiques et les marqueurs moléculaires de la maladie, l'ELN a récemment publié une mise à jour pour 2022, qui comprend une classification révisée du risque génétique [72]. Actuellement, les anomalies génétiques définissent la classification des AML selon l'ELN 2022, avec des caractéristiques supplémentaires comme critères de qualification du diagnostic primaire (Figure 1.33).

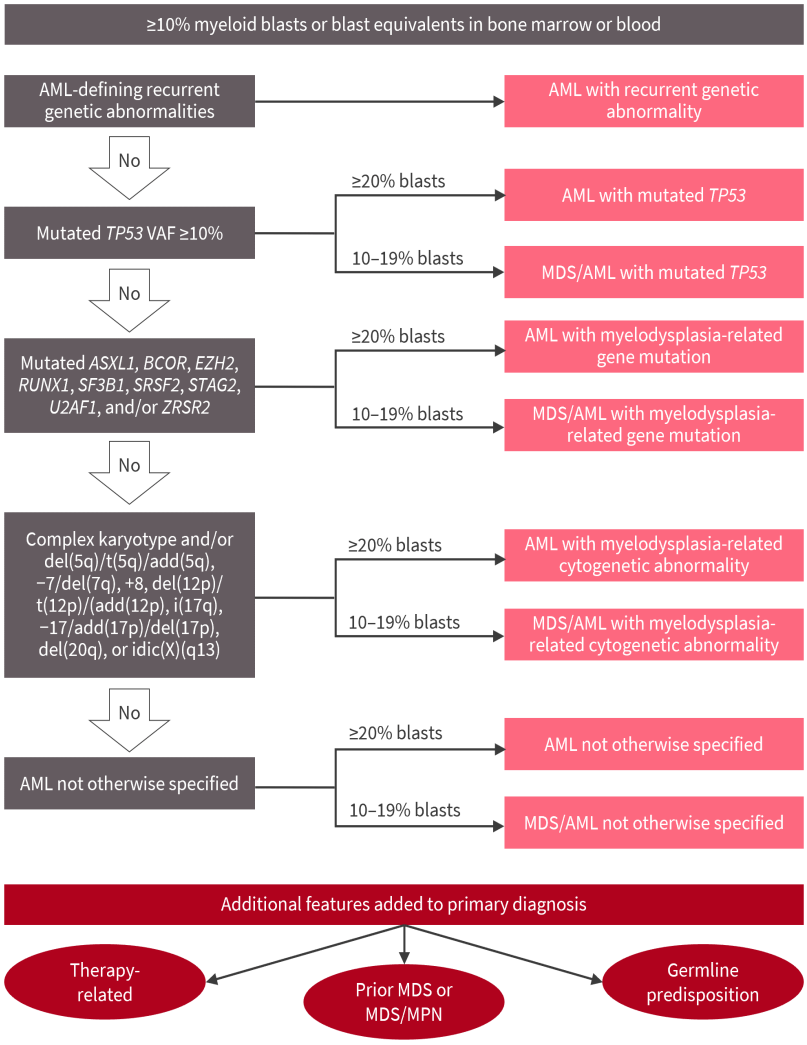


FIGURE 1.33. – *Classification hiérarchique de la l'International Consensus Classification of AML*[33].

33. 2022 ELN recommendations for the diagnosis of AML in adults By Marcia Rato, Becky Gribbell, Helen Croxall, Dylan Barrett

Risk Category	Genetic Abnormality
Favorable	t (8;21) (q22;q22.1); *RUNX1-RUNX1T1* inv (16) (p13.1q22) or t (16;16) (p13.1;q22); *CBFB-MYH11* Mutated *NPM1* without *FLT3*-ITD bZIP in-frame mutated *CEBPA*
Intermediate	Mutated *NPM1* with *FLT3*-ITD Wild-type *NPM1* with *FLT3*-ITD t (9;11) (p21.3;q23.3); *MLLT3-KMT2A* Cytogenetic abnormalities not classified as favorable or adverse
Adverse	t (6;9) (p23;q34.1); *DEK-NUP214* t (v;11q23.3); *KMT2A rearranged* t (9;22) (q34.1;q11.2); *BCR-ABL1* inv(3) (q21.3q26.2) or t (3;3) (q21.3;q26.2); *GATA2, MECOM(EVI1)* t (3q26.2;v); *MECOM (EVI1)*-rearranged −5 or del (5q); −7; −17/abn (17p) Complex karyotype, monosomal karyotype Mutated *ASXL1*, *BCOR*, *EZH2*, *RUNX1*, *SF3B1*, *SRSF2*, *STAG2*, *U2AF1*, or *ZRSR2* Mutated *TP53*

Reprinted with permission from Ref. [6]. 2022, American Society of Hematology.

FIGURE 1.34. – *Classification ELN 2022 des risques en fonction de la génétique au moment du diagnostic initial*[34].

Les travaux décrits ici se concentrent sur la AML. Notre objectif est de mieux comprendre les principes de régulation d'un gène causal dans la AML. Nous allons donc présenter les caractéristiques des ALDH et particulièrement leur implication dans les AML.

34. Döhner, Hartmut, et al. "Diagnosis and management of AML in adults : 2022 recommendations from an international expert panel on behalf of the ELN."

1.6.4. Les Aldéhydes Déshydrogénases

Dans les cellules humaines, 19 sous-types de gènes codant pour l'aldéhyde déshydrogénase (ALDH) ont été identifiés sur différents chromosomes, l'épissage alternatif fournit par ailleurs encore plus de variabilité lors de la traduction. Les protéines ALDH sont divisées en onze familles de protéines et quatre sous-familles qui sont localisées dans différents compartiments cellulaires : dans le cytoplasme, le noyau ou les mitochondries [29]. En plus de leur activité enzymatique, ils sont nécessaires lors de l'embryogenèse et du développement fœtal [118]. La famille ALDH se caractérise par son engagement dans un large éventail de processus biologiques essentiels à la survie cellulaire ainsi qu'aux mécanismes de protection cellulaire. Il joue un rôle crucial dans la conversion de la vitamine A en son métabolite actif, l'acide rétinoïque (RA). Les ALDH sont également régulés positivement chez les mammifères en réponse à la fois au stress oxydatif et à la peroxydation lipidique [267]. Ils détoxifient les aldéhydes potentiellement dangereux [299]. Ils ont des fonctions antioxydantes et osmorégulatrices et ils sont également engagés dans le métabolisme des médicaments et la différenciation cellulaire [299].

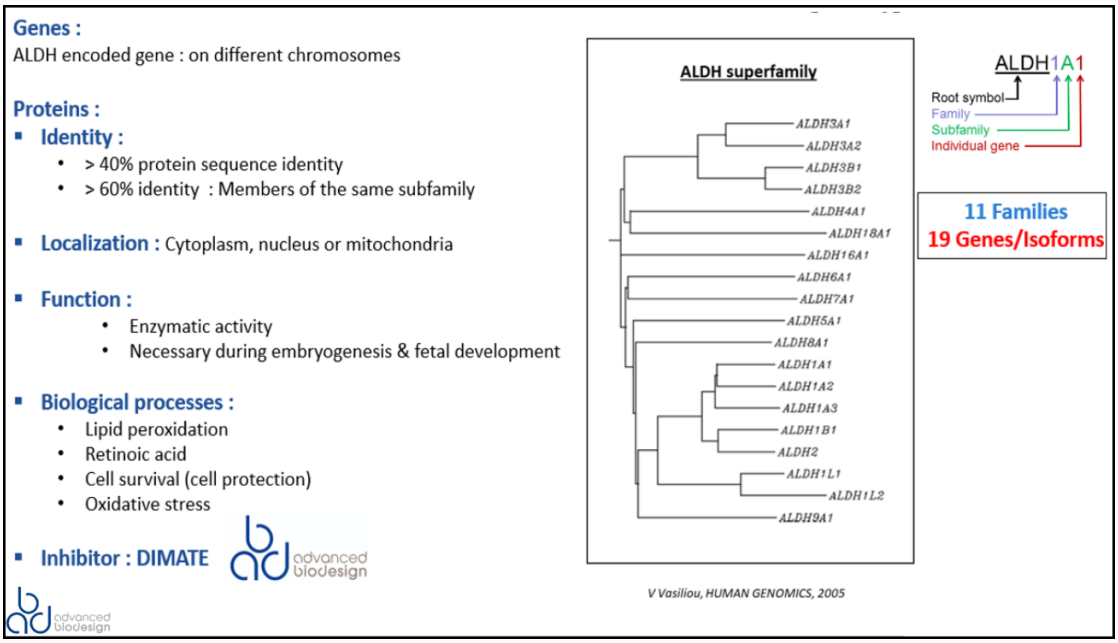


FIGURE 1.35. – *Carte d'identité de la super famille des ALDH.* [35].

35. Figure réalisé par notre collaboratrice Yasmine LABIAD.

1.6.5. Rôle des ALDH dans les AML

Bien que la majorité des patients atteints de AML atteignent une rémission complète après une chimiothérapie d'induction standard, la majorité rechute et meurt de la maladie. Ce manque de réponse complète est expliqué par le paradigme de la cellule souche de la leucémie (LSC), où les LSC sont hautement enrichies dans les cellules leucémiques CD34 + CD38- qui présentent une activité positive de l'aldéhyde déshydrogénase (ALDH +) sur la cytométrie en flux [300]. La mise à jour de l'ALDH en tant que biomarqueur potentiel et cible thérapeutique pour la AML par Xiangchou et al. [318] déclare que des études utilisant la transplantation de souris ont illustré le rôle de l'aldéhyde déshydrogénase (ALDH) dans la définition des cellules souches hématopoïétiques (HSC) et des cellules souches de la leucémie. En plus d'être un marqueur moléculaire, l'ALDH médie la résistance aux médicaments dans la AML, ce qui induit un mauvais pronostic pour les patients. D'autres travaux révélent que les AML les gènes qui codent pour les ALDH sont dérégulés chez 42 % des cas (amplifications ou surexpression) et cela impactent négativement la survie des patients [98, 47].

Ces LSC ALDH +, résistent aux traitements actuels de la AML tels que la cytarabine et l'anthracycline, qui sont hautement actifs contre la masse leucémique, mais épargnent les LSC responsables de la rechute [300]. Des études sur des modèles animaux ont montré que la suppression de l'expression d'ALDH1A1 ou l'inhibition de son activité enzymatique peut réduire la croissance tumorale et augmenter la sensibilité aux traitements chimiothérapeutiques [318]. L'inhibiteur d'ALDH, DIMATE, a été testé sur des sous-populations de patients atteints d'AML et de patients sains pour combattre les LSC de manière sélective. Il a été démontré que DIMATE est un puissant inhibiteur des ALDH 1 et 3 avec une activité cytotoxique majeure sur les lignées cellulaires de AML humaines. [300].

Ces travaux révèlent l'importance de l'étude des ALDH afin de trouver un traitement efficace pour les patients atteints de AML. Dans ce projet, en collaboration avec ABD, nous souhaitions comprendre la régulation des ALDH en vue de leur rôle important dans les cancer en général et dans la AML en particulier, et cela en identifiant des acteurs clés dans cette régulation tels que des facteurs de transcription ainsi que des régions promotrices ou de type enhancer. Cette compréhension de la régulation des ALDH pourra aider ABD à mieux réprimer les ALDH en inhibant (en plus du DIMATE) les facteurs de transcription importants dans leur régulation.

2. Catalogue de régions régulatrices dans quatre espèces

Sommaire

2.1. Introduction

2.1.1. Histoire du ChIP-seq

Le ChIP-seq (Chromatin Immunoprecipitation Sequencing) est une technique de biologie moléculaire utilisée pour identifier les régions de l'ADN qui sont liées à des protéines spécifiques dans les cellules vivantes. Cette technique a été développée dans les années 2000 et a rapidement gagné en popularité en raison de sa capacité à fournir des données de haute précision et de haute résolution sur l'interaction des protéines avec l'ADN.

L'histoire du ChIP-seq débute avec la technique de Chromatin Immunoprecipitation (ChIP), qui a été développée dans les années 1990 par Solomon et Al. [272]. Cette technique permet de purifier des régions spécifiques de l'ADN liées à des protéines spécifiques en utilisant un anticorps spécifique de la protéine cible. Cependant, cette technique ne permettait pas de localiser précisément les régions d'interaction entre la protéine et l'ADN. La première expérience de ChIP a utilisé des anticorps contre l'histone H4 pour étudier comment cette protéine est liée à l'ADN dans les cellules de D.melanogaster. Cette expérience a ouvert la voie à de nombreux autres travaux utilisant les expériences ChIP pour étudier les interactions protéine-ADN dans différents organismes et contextes biologiques.

En 2001, une nouvelle technique appelée ChIP-chip (Chromatin Immunoprecipitation on Chip) qui combinait la technique de ChIP avec l'hybridation sur puce pour permettre la localisation précise des régions d'interaction entre les protéines et l'ADN [137] a été introduite par Peggy Farnham et Michael Zhang [307]. Cependant, cette technique présentait des limites en terme de résolution et de précision. En 2007, un groupe de chercheurs dirigés par David S. Johnson et Al. à Stanford University a publié un article décrivant une nouvelle technique appelée ChIP-seq qui combinait la technique de ChIP avec le séquençage d'ADN (NGS) pour permettre une localisation encore plus précise des régions d'interaction entre les protéines et l'ADN [137, 20, 243, 202]. Cette technique a rapidement gagné en popularité en raison de sa capacité à fournir des données de haute précision et de haute résolution sur l'interaction des protéines avec l'ADN.

Depuis lors, le ChIP-seq est devenu un outil incontournable pour étudier les interactions entre les protéines et l'ADN dans les cellules vivantes, en particulier dans les études de génétique épigénétique et sur les gènes associés à des maladies [137, 243, 20]. La combinaison de la technique de ChIP avec le séquençage d'ADN a permis de découvrir de nouvelles régions d'interaction et de mieux comprendre comment les protéines régulent l'expression génique.

En plus de son utilisation dans l'étude de l'épigénétique et des maladies, la technique ChIP-seq a également été utilisée pour étudier l'architecture de la chromatine, les mécanismes de la régulation de l'expression des gènes, et la découverte de nouveaux gènes cibles de la régulation. Les données obtenues par ChIP-seq ont aussi été utilisées pour créer des modèles de réseaux de régulation, qui permettent de comprendre comment les différentes protéines interagissent pour réguler l'expression génique. De plus, des méthodes d'analyse de ChIP-seq se sont développées pour améliorer la précision et la sensibilité de la technique, permettant une meilleure compréhension des interactions protéine-ADN. Par exemple, la méthode de ChIP-exo (Chromatin Immunoprecipitation followed by exonuclease digestion) qui permet de distinguer les interactions protéine-ADN de courte durée des interactions de longue durée [241].

2.1.2. Stockage des données

Les données obtenues par ChIP-seq sont stockées dans des entrepôts de données tels que GEO. Cependant, il existe une grande diversité dans la manière dont les expériences ChIP-seq sont conçues, exécutées et annotées, ce qui peut entraîner des problèmes de variabilité et de qualité des données. Ces problèmes peuvent affecter non seulement les analyses de données ChIP-seq, mais également la capacité de comparer les données de plusieurs analyses ou d'effectuer des analyses intégratives sur plusieurs types de données.

En 2011, les consortiums ENCODE et modENCODE ont réalisé des expériences ChIP-seq de grande envergure pour étudier les interactions entre les protéines et l'ADN dans différents organismes modèles, et ainsi identifier la position des éléments régulateurs. En utilisant plusieurs pipelines indépendants pour la production et le traitement des données, ces consortiums ont réalisé plus d'un millier d'expériences ChIP-seq individuelles pour plus de 140 facteurs différents et modifications d'histones dans plus de 100 types de cellules dans quatre organismes différents : la drosophile (*D. melanogaster*), le ver de terre (*C. elegans*), la souris et l'Homme [85, 166].

Selon Feingold et Pachter (2004)[85], les consortiums ENCODE (Encyclopedia of DNA Elements) et modENCODE (Model Organism Encyclopedia of DNA Elements) ont mis en évidence les problématiques inhérentes à l'intégration de données provenant de sources hétérogènes, telles que l'uniformisation et l'homogénéisation de l'annotation et le processing des données. Landt et al. [166] ont également souligné les problèmes communs à tous les travaux portant sur les expériences ChIP-seq, tels que la spécificité et la qualité de l'immunoprécipitation, l'impact de la profondeur de séquençage de l'ADN, la qualité de l'annotation et de l'ensemble de données, les expériences de contrôle appropriées, les réplicats biologiques et la mise à disposition des données et métadonnées.

Il est important de noter que les techniques ChIP-seq ont continué à évoluer depuis les publications initiales, avec de nouvelles méthodologies et outils développés pour améliorer la qualité et la quantité des données obtenues. Cependant, les problèmes de qualité et de variabilité des données mentionnés dans les projets précédents [85, 166] restent encore d'actualité.

2.1.3. Le projet ReMap

C'est dans ce contexte que le projet ReMap (Regulatory Map) a été lancé en 2012 sous la direction de Dr. Benoit Ballester et Dr. Aurélien Griffon. L'objectif initial était de créer un catalogue de régions régulatrices de l'ADN en utilisant les données ChIP-seq disponibles dans les entrepôts de données publics tels que GEO et ENCODE. Pour répondre aux problèmes liés à l'intégration de données provenant de sources variables, l'objectif était de traiter les données de manière homogène et de les annoter manuellement.

Le premier article sur ReMap a été publié en 2015 [108], et dans sa première version, le catalogue comprenait 237 facteurs de transcription (TF) et 8 millions de pics ChIP-seq, étant l'un des premiers catalogues de ce type avec Cistrome [179]. Depuis, ReMap a continué à évoluer pour inclure de nouvelles espèces et de nouveaux facteurs de transcription, avec dans la dernière version de 2022, un catalogue comprenant 1210 facteurs de transcription et 182 millions de pics ChIP-seq.

Le projet ReMap vise à créer un catalogue de régions régulatrices de l'ADN pour différentes espèces, qui sont modèles en génomique. En utilisant des données issues d'expériences ChIP-seq hétérogènes et de différentes sources, ReMap permet de cartographier les régions de l'ADN qui sont liées à des protéines spécifiques dans les cellules vivantes, en fournissant des données de haute précision et de haute résolution sur l'interaction des protéines avec l'ADN.

2.1.4. Nouvelle version de ReMap

Au cours de mon doctorat, j'ai développé la quatrième version de ReMap. Cette version s'appuie sur le travail de mes prédécesseurs, mais apporte également des nouveautés au projet existant. En 2020, la troisième version de ReMap comprenait les catalogues de régions régulatrices de l'Homme ainsi que d'*Arabidopsis thaliana*. En 2022, j'ai ajouté deux nouvelles espèces modèles : la souris (*Mus musculus*) et la mouche (*Drosophila melanogaster*) [116]. Ces deux nouveaux catalogues répertorient les expériences ChIP-seq effectuées chez ces deux espèces entre 2009 et 2020, les données étant soumises dans GEO. J'ai également mis à jour le catalogue pour l'Homme et pour la plante, augmentant significativement les données disponibles. En 2020, le catalogue pour l'Homme contenait 165 millions de pics, alors qu'en 2022, avant les nouveaux filtres, il en contenait 279 millions.

Pour améliorer encore la qualité de l'annotation des régions régulatrices, en 2022, nous avons ajouté deux nouveaux filtres : l'un portant sur la longueur des pics, l'autre sur le nombre de pics par expérience [116]. Ces filtres ont permis d'obtenir un catalogue de meilleure qualité composé de 182 millions de pics. En 2022, j'ai mis en place un trackhub (piste) en local pour ReMap. Ce trackhub permet une navigation facile et efficace grâce au navigateur de génome UCSC (University of California Santa Cruz). Cette version de ReMap, qui est également disponible en natif sur UCSC depuis le printemps 2022, ajoute une courbe de densité et de nouvelles fonctionnalités de filtrage en fonction des biotypes et des facteurs de transcription. Un trackhub en natif de ReMap présente de nombreux avantages qui seront évoqués dans la section 2.9.

En outre, la quatrième version de ReMap, inclut une nouvelle version du script pour déterminer les modules *cis*-régulateurs (CRMs). Ces améliorations ont été développées par Pierre de Langen, doctorant dans notre groupe. Dans la section 2.10 nous parlerons également des analyses complémentaires que j'ai pu effectuer à l'aide des CRMs générés à partir du catalogue ReMap. Tel que l'analyse de segmentation (ChromHMM) du génome pour déterminer le biotype dominant des CRMs. Les résultats de cette version de ReMap ont été publiés dans Nucleic Acids Research en 2022, l'article est disponible dans la section suivante. [116].

D316–D325 *Nucleic Acids Research, 2022, Vol. 50, Database issue* *Published online 9 November 2021*
https://doi.org/10.1093/nar/gkab996

ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments

Fayrouz Hammal, Pierre de Langen, Aurélie Bergon, Fabrice Lopez and Benoit Ballester*

Aix Marseille Univ, INSERM, TAGC, Marseille, France

Received September 14, 2021; Revised October 07, 2021; Editorial Decision October 08, 2021; Accepted October 13, 2021

ABSTRACT

ReMap (https://remap.univ-amu.fr) aims to provide manually curated, high-quality catalogs of regulatory regions resulting from a large-scale integrative analysis of DNA-binding experiments in Human, Mouse, Fly and *Arabidopsis thaliana* for hundreds of transcription factors and regulators. In this 2022 update, we have uniformly processed >11 000 DNA-binding sequencing datasets from public sources across four species. The updated Human regulatory atlas includes 8103 datasets covering a total of 1210 transcriptional regulators (TRs) with a catalog of 182 million (M) peaks, while the updated Arabidopsis atlas reaches 4.8M peaks, 423 TRs across 694 datasets. Also, this ReMap release is enriched by two new regulatory catalogs for *Mus musculus* and *Drosophila melanogaster*. First, the Mouse regulatory catalog consists of 123M peaks across 648 TRs as a result of the integration and validation of 5503 ChIP-seq datasets. Second, the *Drosophila melanogaster* catalog contains 16.6M peaks across 550 TRs from the integration of 1205 datasets. The four regulatory catalogs are browsable through track hubs at UCSC, Ensembl and NCBI genome browsers. Finally, ReMap 2022 comes with a new Cis Regulatory Module identification method, improved quality controls, faster search results, and better user experience with an interactive tour and video tutorials on browsing and filtering ReMap catalogs.

INTRODUCTION

Transcriptional regulators (TRs) such as transcription factors (TFs), transcriptional coactivators (TCAs) and chromatin-remodeling factors (CRFs), drive gene transcription and chromatin organization through DNA binding. Since the advent of chromatin immunoprecipitation followed by sequencing (ChIP-seq (1)), it has become possible to study the genome-wide occupancy of DNA-binding proteins. The popularity of ChIP-seq and other ChIP-based sequencing techniques has increased research in genome occupancy maps for various TRs, in various conditions, cells, tissues, and species. This led to an accumulation of functional genomics datasets stored in data repositories such as NCBI GEO (2), ENA EBI (3) or DDBJ (4) providing a unique resource for thousands of DNA-binding sequencing studies. A large-scale integration of these studies would reveal the transcriptional regulatory repertoire as transcription of the Human genome is controlled by about 1600 transcription factors (5,6). The genomic architecture of the regulatory space has started to unfold thanks to large functional genomic consortia (ENCODE (7,8), Roadmap (9)) but more remains to be discovered. Such large-scale integration is challenged by the variety of bioinformatics methods and underlying data formats, the inconsistency in targets, cell types or tissue names, as well as experimental ChIP and sequencing quality. However, such integrative analysis would offer significant insights of the transcriptional regulatory repertoire in different cellular environments.

In 2015, the ReMap project initiated the first large-scale integrative analysis of heterogeneous ChIP-seq revealing the complex architecture of the Human regulatory landscape using dedicated curation and standardized data processing pipeline (10). The manual curation and annotation of DNA-binding sequencing studies are at the foundation of the ReMap project. Each dataset introduced in ReMap is assessed manually to ensure correct target and biotype annotation, as well as experimental metadata curation, making it distinct from other integrative projects (11–14). The 2015 version of ReMap (10) introduced a Human regulatory catalog of 13 million (M) DNA binding regions for 237 TRs across 83 biotypes (cell lines and tissue types) by

*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

Nucleic Acids Research, 2022, Vol. 50, Database issue **D317**

integrating 395 datasets from GEO and ENCODE. The 2018 ReMap version, followed by the 2020 version, released a Human regulatory catalog of 165M binding regions for 1135 TRs (15,16) by processing ∼5800 datasets. Also, the 2020 ReMap database introduced the first *Arabidopsis thaliana* regulatory atlas as a result of a large-scale data integration of public data and analysis efforts. Since 2018, the ReMap catalogs are used as one of the input sources for the computation of TF binding profiles for the JASPAR database (17,18).

Here, we present the fourth release of ReMap ('ReMap2022'), which comes with a major expansion of the Human and Arabidopsis regulatory atlases. Moreover, we introduce two new regulatory atlases for *Mus musculus* and *Drosophila melanogaster*. These new catalogs are the results of our continuing efforts in large-scale data integration for these two model species. Faced with large catalogs, this update includes new quality controls to improve the repertoire of binding locations as well as a new method for Cis Regulatory Modules (CRMs) identification. Finally, our database update is backed up by new web functionalities for better community access. The web portal displays an interactive tour and genome track filters are available through track hubs on genome browsers. Taken together the manual metadata curation and large-scale integration engaged in the ReMap project offers a unique and unprecedented collection of DNA-binding regions for four major species.

MATERIALS AND METHODS

Available datasets

New ChIP-seq experiments were retrieved from the NCBI Gene Expression Omnibus (GEO) and ENCODE databases (2,19). 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 studies. The same query was used with 'arabidopsis thaliana'[organism], 'mus musculus'[organism] and 'drosophila melanogaster'[organism] to get all the potential datasets for each study. The selected experiments metadata are then manually curated and annotated with official nomenclatures for target names and biotypes. For incomplete metadatas, the materials and methods of associated and published papers are often examined to complete the curation. For Human we used the HUGO Gene Nomenclature Committee (20) (www.genenames.org), the BRENDA Tissue Ontologies for cell lines (21) at the EBI Ontology Lookup Service (22) (www.ebi.ac.uk/ols/ontologies/bto) as well as the Cellosaurus database (23) to homogenize cell and tissue names (e.g. MCF-7 not MCF7, Hep-G2, not HepG2, Hepg2 etc.). For Arabidopsis (*A. thaliana*) we used gene names from the Ensembl Plant genome annotation (24). Ecotypes and biotypes descriptions were curated and homogenized when the information was available in the metadata or associated publication. For Mouse (*M. musculus*) we annotated gene names using the official MGI database (25) (http://www.informatics.jax.org/). For Drosophila (*D. melanogaster*) we used the Flybase database (26) (https://flybase.org/) for gene names. To improve automatic processing of files we removed parentheses and replaced them with hyphens to better handle these names in the pipeline (e.g. E(z) to E-z, See Supplementary Table S12). For Mouse and Drosophila the tissues or cell lines annotation were homogenized using BRENDA Tissue Ontologies as well as the Cellosaurus database. ChIP-seq studies involving RNA polymerases (RNA-Pol2 and RNA-Pol3) were filtered out. When multiple antibodies were pooled (e.g. RUNX1 and RUNX3, GSE17954) targets would be named as RUNX1-3. Also, when a 'global' antibody is used to pool a family of targets (eg: RAR, GSE35599) we would name the target as just RAR.

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, Larva, Leaf, Limb) in a given biological condition (e.g. 45min DMSO, 21d-wt-watered). Datasets are labeled with the concatenation of these informations (e.g. GSE37345.FOXA1.LNCAP_45min-DMSO). For the 2022 update a total of 12 976 new datasets were processed across all four species (Supplementary Table S1). Specifically, we analysed 4121 Human datasets deposited in public repositories from 11 November 2018 to 11 September 2020; 223 Arabidopsis datasets from 3 February 2018 to 9 March 2021; 7317 Mouse datasets from 1 January 2009 to 2 February 2020; and 1308 Drosophila datasets from 1 January 2008 to 17 September 2020 (full list of datasets in Supplementary Tables S3, S5, S7–S9 and S11).

In the 2022 update, as well as in previous updates, the new ENCODE ChIP-seq experiments for TFs, transcriptional and chromatin regulators were re-analysed starting from raw data (FASTQ files) following the same processing pipeline. For Human, we processed data between 6 February 2019 to 18 March 12021, for Mouse, all data until 16 November 2020 and for Drosophila all data until 28 April 2021. We retrieved the list of ENCODE ChIP-seq experiments as FASTQ files from the ENCODE portal (8,19) (https://www.encodeproject.org/) using the following filters: Assay: 'ChIP-seq', Organism: 'Homo sapiens', Target of assay: 'TF', Available data: 'fastq' on 6 February 2019. The same filters were used for the Organism: '*M. musculus*' and '*D. melanogaster*'. Metadata information in JSON format and FASTQ files were retrieved using the Python requests module. We processed 508 Human, 167 Mouse and 525 Drosophila ENCODE datasets, of whom 267 Human, 158 Mouse and 514 Drosophila passed our quality filters. We renamed TR ENCODE aliases using official HGNC and MGI identifiers (e.g. p65 into RELA, see Supplementary Tables S4 and S10) for Human and Mouse respectively, and renamed cell lines to official BRENDA and Cellosaurus conventions (e.g. K562 into K-562, for curated names see Supplementary Tables S4, S6, S10 and S12).

ChIP-seq processing

For each species, the GEO and ENCODE datasets were curated, processed and analysed in the same way. Bowtie 2 (version 2.2.9 (27)) with options -end-to-end -sensitive was used to align all reads on the human genome GRCh38/hg38 assembly, the *A. thaliana* TAIR10 assembly, the *D. melanogaster* BDGP6.32/dm6 assembly and the

D318 Nucleic Acids Research, 2022, Vol. 50, Database issue

M. musculus GRCm38/mm10 assembly. The GRCm39 assembly was released during the Mouse production. Trim Galore (https://github.com/FelixKrueger/TrimGalore) was used to remove adapters, 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. With samtools rmdup polymerase chain reaction duplicates were removed from the alignments. Following the ENCODE ChIP-seq guidelines (28) we used the MACS2 peak-calling tool (29) (version 2.1.1.2) with default thresholds (MACS2 default thresholds, Q-value: 5e–2, -g: with corresponding genome sizes) to identify the TR binding regions. For all the datasets, the corresponding bed file is available for download.

Quality assessment and filters

Because the analysed data comes from a variety of sources and is produced under varied experimental circumstances and platforms, the data quality differs from study to study. Since the first release of ReMap 2015, and unlike similar databases (Supplementary Tables S3, S5, S9 and S11), our pipeline has assessed the quality of each processed dataset. The same quality pipeline and cutoffs were used for ReMap 2022 as they were for ReMap 2020, we named these quality assessments as QC1 in Figure 1. Briefly, we derived a score based on the ENCODE consortium's cross-correlation and FRiP (percent of reads in peaks) metrics (Supplementary Figures S1–S4, ENCODE quality coefficients http://genome.ucsc.edu/ENCODE/qualityMetrics.html) for all species and ChIP-seq datasets for this release. The normalized strand cross-correlation coefficient (NSC), which is a ratio of the maximal fragment-length cross-correlation value to the background cross-correlation value, and the relative strand cross-correlation coefficient (RSC), which is a ratio of the fragment-length cross-correlation to the read-length cross-correlation, are then computed by our pipeline. Datasets that failed QC1 were not included in the catalogs or the BED files available for download. Rejected datasets are listed in (Supplementary Tables S3, S5, S9 and S11).

In 2022, two new sets of filters were added in our assessment steps, named QC2 in Figure 1. These filters were applied to the new catalogs, and retroactively to previous data for Human and Arabidopsis. We discarded datasets having less than 100 peaks or more than two times the number of annotated genes according to the Ensembl gene annotation statistics. As of late 2021, about 20 000 coding and non-coding genes are identified for Drosophila giving a cutoff of a maximum of 40 000 peaks, about 30 000 annotated genes for Arabidopsis giving a cutoff of 60 000 peaks, and about 40 000 annotated genes for Human or Mouse giving a cutoff of 80 000 peaks (Supplementary Figure S5). For the second filter, within each dataset we discarded peaks that fall outside a base pair length range. The range is defined as 50bp minimum and an upper cutoff in which we have 99% of catalogue (Supplementary Figure S6A). These upper cutoffs are rounded to 1.5 kb for Human and Mouse, 2kb for Drosophila and Arabidopsis (see Supplementary Figure S6B). Full data with rejected peaks are made available in the download table as 'Permissive peaks' in the Supplementary tab.

Open ReMap pipeline

The code of the ReMap pipeline is available on Github in the ReMap Github organisation (https://github.com/remap-cisreg). The pipeline uses Snakemake (30) in a Conda (https://conda.io) or Singularity (https://sylabs.io/) environment, depending on the High-Performance Computing (HPC) resources available, and both Torque and Slurm managers are supported. You can also find multiple python and shell scripts used to format the datasets. The repository contains information in the Github wiki on the utilisation of the pipeline.

Non-redundant peak sets

ReMap provides non-redundant binding regions (NR peaks) for each target, a unique feature not seen in other databases (Supplementary Table S2). This gives an accurate genomic location of peaks regardless of biotypes, in a multicell manner. All peak lengths for a TR were trimmed to the median size of all peaks in that TR. Then, we used BedTools to intersect overlapping truncated peaks across multiple datasets to discover clusters of duplicate peaks (with at least 25% overlap, both ways). After the clusters of overlapping peaks have been identified, non-redundant peaks are computed by averaging start, end and summits coordinates of all peaks in a cluster using original ReMap peaks lengths. The non-redundant peak set across all experiments for a particular factor consists of calculated non-redundant peaks plus singletons and is available for download from the ReMap website.

Cis regulatory modules

To accurately delineate CRMs in the 2022 release, we have developed a new methodology that relies on detecting peak density along the genome. Briefly, we delineate CRMs at each local minimum of NR peak density function (See Supplementary Figure S7B). A NR peak is assigned to a CRM if its peak summit falls in between the identified flanking local minimas. Finally, the CRM boundaries are reduced using the 5′ and 3′ coordinates of the flanking NR peaks. The 'peakMerge' code is available at https://github.com/remap-cisreg/peakMerge.

UPDATE OF THE HUMAN AND ARABIDOPSIS CATALOGS

Human transcriptional regulatory expansion

This fourth release of ReMap comes with a significant update of the Human regulatory catalog by adding a large number of new datasets, new transcriptional regulators and biotypes. We curated, processed and analysed 4121 ChIP-seq datasets targeted against TRs collected from GEO and ENCODE databases since the last release. Since 2015 we provide consistency and comparability between datasets by processing raw data through our standardized ReMap pipeline, which includes read filtering, read mapping, peak calling, and quality controls (See 'Materials and methods'). In contrast to other databases, the manual curation and annotation of studies we process are at the foundation of the

Nucleic Acids Research, 2022, Vol. 50, Database issue **D319**

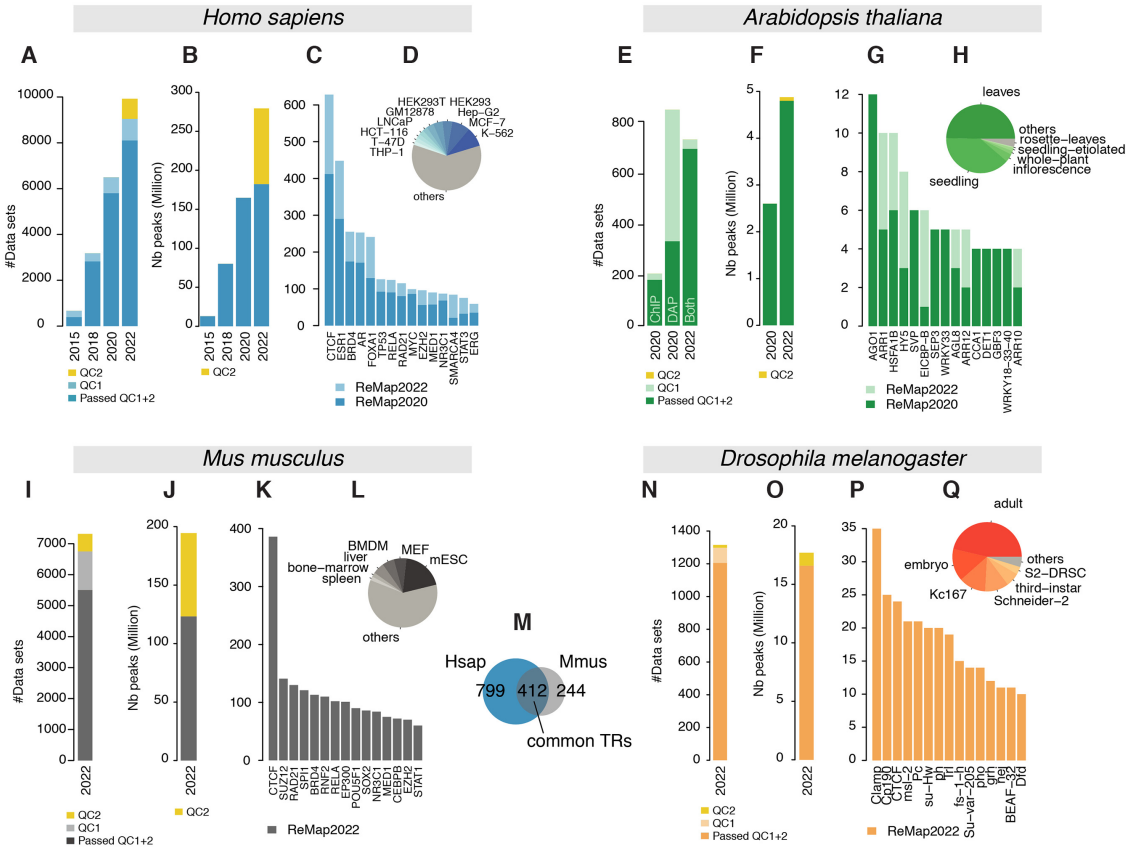


Figure 1. Overview of the ReMap 2022 database growth. (**A**) Analysed Human datasets growth in ReMap 2022 compared to 2020, 2018 and 2015, removed datasets from the new quality control (QC2) in yellow. (**B**) Human ChIP-seq peaks growth in ReMap 2022 compared to previous releases, yellow corresponds to removed peaks. (**C**, **D**) Evolution of the number of datasets across the top 15 transcriptional regulators (TRs) and biotypes between ReMap 2022 and 2020. (**E**) Analysed *Arabidopsis thaliana* datasets in 2022 compared to 2020. (**F**) Arabidopsis regulatory peaks growth in ReMap 2022 compared to 2020, yellow corresponds to removed peaks (**G, H**) Evolution of the number of datasets across the top 15 transcriptional regulators and biotypes between the two Arabidopsis ReMap catalogs. (**I**) Analysed ChIP-seq datasets for the ReMap 2022 *Mus musculus* catalog, removed datasets from QC2 in yellow. (**J**) Size of the *Mus musculus* regulatory catalog, before and after QC2 (yellow). (**K, L**) Number of datasets for the top 15 TRs and top 6 biotypes. (**M**) Transcriptional regulators shared between the Human and Mouse regulatory catalogs. (**N**) Analysed ChIP-seq datasets for the ReMap 2022 *Drosophila melanogaster* catalog, removed datasets from QC2 in yellow. (**O**) Size of the *Drosophila melanogaster* regulatory catalog, before and after QC2 (yellow). (**P,Q**) Number of datasets for the top 15 TRs and top 6 biotypes.

ReMap project. It involves analysing the warehouse study design descriptions and reading submitted materials and methods from GEO or in the manuscripts to curate heterogeneous experimental information. Furthermore, to correct the diverse quality of DNA binding experiments the pipeline contains quality controls and filtering steps (See 'Materials and methods'). After applying our quality filters, we retained 2828 datasets (65%) from the 4121 new deposited ChIP-seq datasets (Supplementary Figures S1, S5–S6). As a result, the updated Human regulatory atlas contains 182 416 820 peaks, derived from 8,103 datasets (Figure 1A), which includes 1210 TRs (Figure 1B). More precisely 181 426 344 peaks spread over 1002 TRs come from ChIP-seq studies, while 990 476 peaks spread over 208 TRs come from ChIP-exo studies. A large ChIP-exo study was processed (GSE151287), but the resulting peaks were inconsistent with published regions, and not added in this release. When compared to ReMap 2020, the large data gain is dispersed over practically all TRs (Figure 1C, light blue bars). We observe that the most studied transcription factors (e.g. ESR1, AR, FOXA1, TP53), transcriptional repressors (e.g. CTCF), and CRFs (e.g. BRD4) display, as expected, more datasets than other DNA-binding proteins. Nonetheless, all of the top 15 TRs show additional datasets integrated in ReMap 2022 (Figure 1C, light blue bars). The top 10 biotypes present in the human catalog correspond to the most common cell lines used in genomics (e.g. MCF-7, K-562; Figure 1D). Our uniform data processing contributes to an updated ReMap 2022 human regulatory atlas of 182M binding regions revealing an unprecedented view of the regulatory landscape and complexity. To illustrate this complexity with dense co-localizations of peaks creating tight clusters (CRMs), we have been tracking the Human ELAC1 promoter since 2015. This genomic region

D320 *Nucleic Acids Research, 2022, Vol. 50, Database issue*

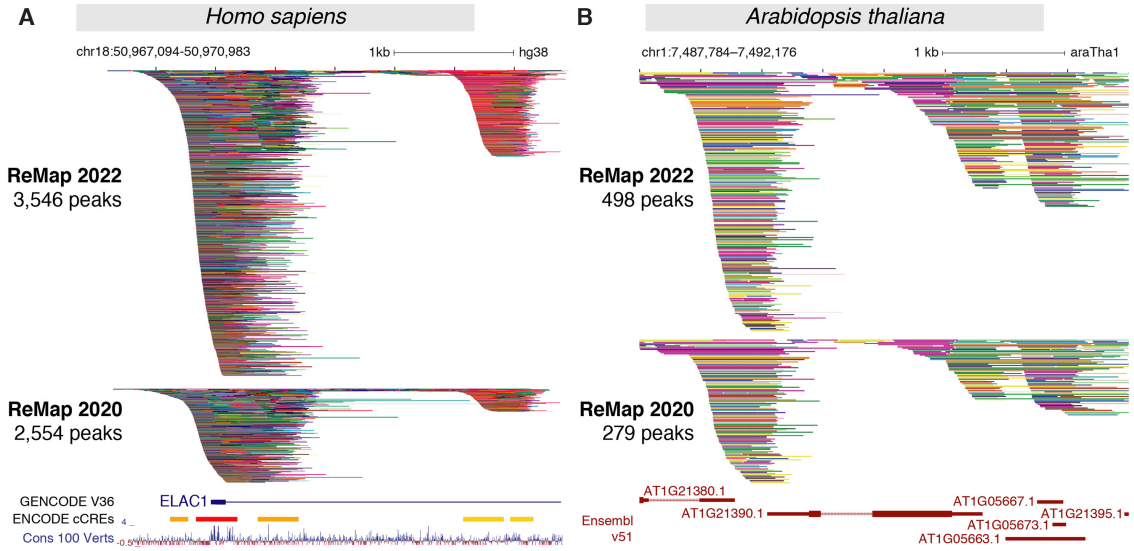


Figure 2. Updated ReMap 2022 regulatory atlas for Human and Plant. (**A**) ReMap 2022 Human DNA-protein binding pattern of 8103 valid datasets. This genome browser example of the DNA-binding peak depth of ReMap 2022 atlas is compared to ReMap 2020 at the vicinity of the ELAC1 promoter (chr18:50 967 094–50 970 983). The tracks displayed are compacted to thin lines so the depth of ReMap 2022 bindings can be compared to 2020. Around this ELAC1 location ReMap 2022 displays 3546 peaks, while the 2020 version contains 2554 peaks. The following genome tracks correspond to the GENCODE v36 annotation, the ENCODE candidate Cis Regulatory Elements (cCREs, red promoters, orange proximal enhancer-like, yellow distal enhancer-like) and the 100 vertebrates base-wise conservation showing regions predicted to be conserved (positive scores in blue). (**B**) A genome browser view of the ReMap 2022 Arabidopsis TF a atlas compared to the ReMap 2020 version at the vicinity of the AT1G21390.1 gene model (chr1:7 487 784–7 492 176). The annotation genome track corresponds to the latest Ensembl Plants v51 TAIR10 gene annotation. All peaks have been compacted for rendering.

(chr18:50 967 094–50 970 983) highlights the 2022 catalog expansion compared with 2020 (Figure 2A), and across all four catalogs with 229, 1037, 2554 and 3546 binding regions respectively (Supplementary Figure S8). Three clusters of peaks can be observed, one large at the promoter embedding a second cluster after the transcription start site (TSS) at about +500 bp and +2 kb from the Gencode TSS (31). Interestingly these clusters concords with locations of candidate Cis-Regulatory Elements (cCREs) derived from ENCODE data (8). The third cluster located further up the first intron has been described in depth in precedent ReMap papers (10,15,16) to illustrate how combining data from various sources enhances genome annotations. Indeed, this third cluster contains 179 FOXA1 peaks in the 2022 catalog (93 in 2020, 60 in 2018, 15 in 2015) including one peak from ENCODE (7) (Supplementary Figure S9). The ReMap database provides three majors atlases, the main catalog containing all binding regions, a non-redundant set and a CRMs atlas. Indeed, to show a discrete repertoire of binding regions in the genome, the redundant binding regions are merged for identical TRs resulting in a multi-cell, multi-tissue regulatory map of 68M non-redundant binding regions (see 'Materials and Methods' for details). The CRM atlas (3.4M regions) has been improved with a new methodology that relies on detecting peak density along the genome, reinforcing the identification of regulatory hotspots by the integration of ChIP-seq from various cells, antibodies, and laboratories. Overall, this 2022 update of the human catalog expands the genome regulatory space revealing complex regulatory architectures strengthening the identification of DNA bound regions across thousands of experimental evidences.

Arabidopsis thaliana regulatory update

The *A. thaliana* 2022 release focuses on updating the regulatory catalog as illustrated in Figure 1E–H and in Figure 2. In this update, we curated, processed and analysed 223 new ChIP-seq datasets against TRs submitted in GEO since our previous release (Figure 1E). These datasets were processed uniformly using the ReMap pipeline. After applying quality controls and filters, 185 datasets (79%) were retained then integrated with the current catalogue leading to a total of 694 datasets (Figure 1E, Supplementary Figure S2). The 2022 Arabidopsis regulatory catalog contains 4.8M binding regions for 423 TRs in 23 biotypes and 14 ecotypes (Figure 1F–H). This update shows a growth in both the number of peaks and the number of TRs. In fact, the number of peaks is almost 2-fold superior to the 2020 Arabidopsis regulatory catalog (Figure 1F). The top three most represented immunoprecipitated DNA-binding proteins are Argonaute protein AGO1 (mRNA and chromatin binding), the two-component response regulator ARR1 (Transcription activator), the Heat stress transcription factor HSFA1B (Figure 1G), while the two most represented biotypes are leaves and seedlings (Figure 1H). This ReMap Arabidopsis regulatory catalog reveals an unprecedented view of the landscape and complexity of a Plant transcription factors occupancy map. This complex architecture in a plant genome is illustrated in the vicinity of

Nucleic Acids Research, 2022, Vol. 50, Database issue **D321**

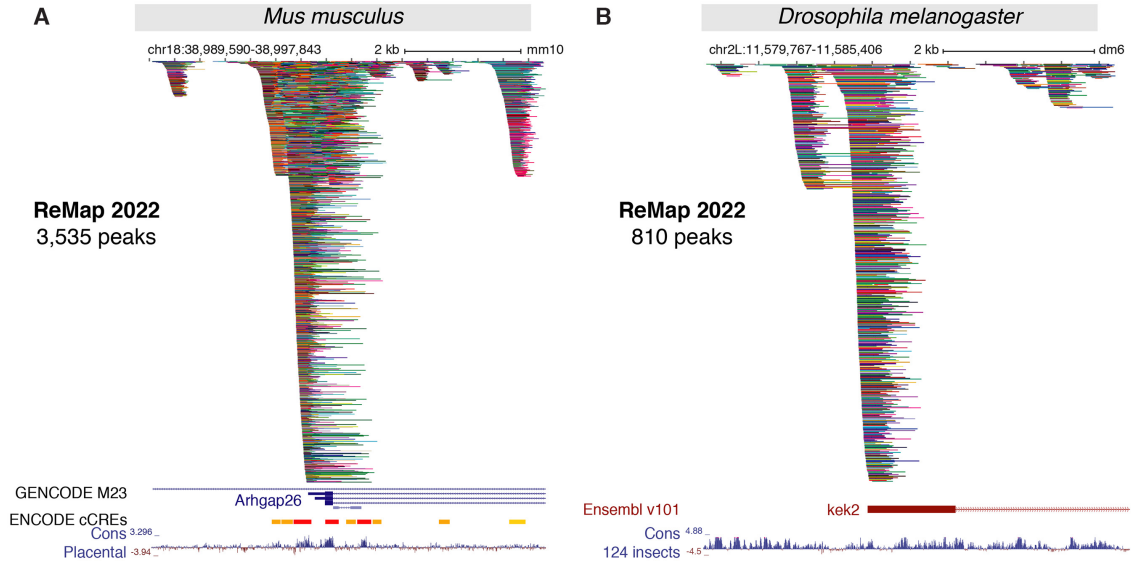


Figure 3. New ReMap 2022 regulatory atlas for *Mus musculus* and *Drosophila melanogaster*. (**A**) A genome browser view of the first Mouse ReMap 2020 atlas for transcriptional regulators at the vicinity of the Arhgap26 gene (chr18:38 989 590–38 997 843). Genome tracks correspond to the GENCODE M23 gene annotation, the ENCODE candidate Cis Regulatory Elements for mouse (cCREs, red promoters, orange proximal enhancer-like, yellow distal enhancer-like) and the Placental Mammal Basewise Conservation showing regions predicted to be conserved (positive scores in blue). (**B**) The Drosophila ReMap 2022 regulatory atlas at the vicinity of the kekkon 2 promoter (kek2, FBgn0265689, chr2L:11 579 767–11 585 406). The genome tracks correspond to the Ensembl v101 *Drosophila melanogaster* gene annotation and the 124 insects Basewise Conservation showing regions predicted to be conserved (positive scores in blue). All peaks have been compacted for rendering.

the AT1G21390.1 gene model (emb2170; chr1:7 487 784–7 492 176) revealing the data growth between 2022 and 2020 catalogs (498 and 279 peaks respectively, Figure 2B). The genomic view highlights how the regulatory repertoire can be complemented by uniform processing of public studies. The 2022 Arabidopsis release comes with the usual three catalogs, all peaks, non-redundant peaks and CRMs identifications. To date, this Arabidopsis ReMap catalog is still the first to provide a global view of all detected TRs binding in a wide variety of biological contexts and a variety of experiments.

NEW REGULATORY CATALOGS FOR MOUSE AND DROSOPHILA

Mouse transcriptional regulatory catalog

A new regulatory catalog for *M. musculus* is included in this fourth release of ReMap, as a result of a one-of large-scale integration of public murine DNA-binding assays. While few Mouse regulatory atlases from consortiums (8,32–34) are available for viewing in genome browsers, none offer significant details and depth of transcription factors occupancy map. To create this unique Mouse regulatory atlas we have collected, curated, uniformly processed, and analysed 7317 ChIP-seq datasets against TRs, precisely 7050 from GEO and 167 from Mouse ENCODE (Figure 1I). After applying quality controls and filters 5503 datasets (76%) were retained (Supplementary Figure S3). Our analyses lead to a final Mouse regulatory atlas of 123 207 170 binding regions for 648 TRs in 373 different murine biotypes (Figure 1J–L). The top three most represented TRs are the transcriptional repressor *Ctcf*, the polycomb protein *Suz12* and the double-strand-break repair protein *Rad21* homolog Rad21 (also a member of the cohesin complex) (Figure 1K). The most commonly represented biotypes (cells or tissues) are the mouse Embryonic Stem Cells (mESC), Mouse Embryonic Fibroblast (MEF) and Bone Marrow Derived Macrophage (BMDM) (Figure 1L). A recent study (35) has experimentally quantitatively identified over 60% of Mouse TFs ($n = 941$) out of the approximated 1500 TFs encoded in the mammalian genome (6). The complex regulatory architecture of the Mouse ReMap catalog is illustrated in the vicinity of the Rho GTPase-activating protein 26 (Arhgap26) gene (Figure 3A), combined with the Mouse ENCODE cCREs annotation (8). We observe a good correlation between cCREs regions and ReMap peak clusters (99.6% overlap), with some clusters (e.g. first cluster) yet undescribed by ENCODE. While our catalog contains 648 transcriptional regulators (TRs), which may represent ∼44% of the current TF census (6), it provides a unique collection of manually curated and uniformly processed ChIP-seq datasets from heterogeneous sources. When comparing with the Human atlas, 412 transcriptional regulators are found in common with the Mouse catalogs (Figure 1M), allowing exploration of evolutionary conservation of cis-regulatory modules (36,37). We present a unique collection of regulatory regions in Mouse as a result of a large-scale integrative analysis of ChIP-seq experiments for hundreds of transcriptional regulators.

D322 *Nucleic Acids Research, 2022, Vol. 50, Database issue*

Drosophila transcriptional regulatory catalog

Finally, this new ReMap release allows the Fruit Fly community to browse and study the regulatory landscape of the *D. melanogaster* genome as we present a regulatory catalog of Drosophila ChIP-seq studies in diverse experimental conditions and various tissues and cells. A total of 1308 datasets were processed with 790 datasets from GEO and 525 from ENCODE (461 modERN, 69 modENCODE) (38,39). These datasets were manually curated and annotated using the Flybase database (26) for the official gene name convention. After applying our quality controls and filters 1,205 datasets (92%) were retained (692 from GEO and 514 from ENCODE, Supplementary Figure S4). Our analyses lead to a final Drosophila regulatory atlas of 16 634 486 binding regions for 550 TRs in 17 different fly biotypes (Figure 1N-O). These biotypes correspond to either cell lines (e.g. Schneider-2) or developmental stages (e.g. embryo, first-instar, second-instar, third-instar and adult), where names were standardized across studies. The top three most represented biotypes in our atlas are the adult fly, the embryo and the cell line Kc167 (Figure 1Q). Regarding TRs the top three TRs are the chromatin-linked adaptor for MSL proteins Clamp, a component of the gypsy chromatin insulator complex Cp190 and the transcription factor CTCF (Figure 1P). Like all ReMap catalogs from previous or current releases, the Drosophila regulatory atlas can be browsed with major Genome Browsers (40–42). We illustrate the complexity of the Fly regulatory landscape around the *kek2* gene with 810 peaks forming various clusters at the *kek2* promoter and up/downstream of the TSS (Figure 3B). We present here a unique Fly regulatory occupancy map forming complex architecture revealed by a large-scale integration of public *D. melanogaster* DNA-binding experiments.

CATALOG IMPROVEMENTS

Less is more

In this release we focused on improving the content of the ReMap catalogs by extending our filtering procedures. With updates adding thousands of published datasets, the ReMap catalogs have expanded dramatically for Human, but also for the new Mouse catalog. With a range of 100–200M peaks, a small fraction of spurious peaks or datasets may influence or bias the characterisation of the regulatory landscape. Indeed, ReMap has reached a point where the identification of regulatory occupancy regions by adding large quantities of datasets may have been achieved. High quality redundant ChIP-seq evidence, illustrated by the FOXA1 peaks in ELAC1 gene (Supplementary Figure S9), allows to improve the identification of ReMap occupancy regions. To remove spurious peaks, two sets of controls were added in our quality control steps, named QC2 in Figure 1. These filters were applied to new catalogs, and retroactively for Human and Arabidopsis. We discarded datasets with less than 100 peaks or with more than two times the number of annotated genes, according to the Ensembl gene annotation statistics (Supplementary Figure S5). A few datasets passing our initial quality controls would generate an unusual amount of peaks, potentially affecting the occupancy repertoire. The second filter removes peaks whose length are outside set cutoffs. Ranges were defined as a minimum of 50bp and a maximum upper cutoff for which we have 99% of catalogue, either 1.5kb or 2kb (See Material and Method, Supplementary Figure S6). This discards large peaks spanning multiple regulatory regions, those peaks not being informative for the transcription factor repertoire identification, or definition of CRMs. In Human, these filters (QC2) remove 97M peaks as they are applied retroactively to the entire catalog (279M without QC2), 71M peaks removed in Mouse, 1.1M in Drosophila and 0.1M in Arabidopsis. However, following the open science principles, the peaks filtered out are available in the download section as 'filtered out' peaks. We believe a conservative approach will benefit the ReMap catalogs by removing uninformative peaks and datasets, leading to clearer regulatory catalogs.

Improved CRM identification

With the Human ReMap catalog reaching ∼182M peaks, the identification of clusters of peaks located often in close proximity, or tightly grouped around a TSS, can be problematic (Figure 2A). In this update, we applied a newly developed method to better identify Cis-Regulatory Modules (CRMs). The initial method consisted in merging overlapping non-redundant peaks (NR peaks), but this approach is only applicable when the number of peaks remains small. When dealing with catalogs of >100 millions of peaks, the genome coverage becomes too high, making the initial but simple approach unable to distinguish CRMs properly in high density regions such as in the ELAC1 example (Supplementary Figure S7A). Thus, to accurately identify CRMs, we have developed a new methodology that detects individual regions in tight clusters of peaks by defining CRMs at each local minimum of NR peak density function (Supplementary Figure S7B). An NR peak is assigned to a CRM if its peak summit falls in between the identified flanking local minimas. The CRM boundaries are reduced using the 5′ and 3′ coordinates of the flanking NR peaks. This newly developed method better defines the genomic organization of our atlas by better identifying dense co-localizations of non-redundant peaks forming tight clusters of transcriptional regulators.

AN IMPROVED USER EXPERIENCE

Interactive tour and fast search

The ReMap 2022 release comes with an improved user experience. On the ReMap homepage, as well as on the header of the site, we provide an interactive tour walking users through the main features of the website. The tour is activated by clicking on the 'Tour' button right in the middle of the homepage or header bar. The tour dynamically shows the different types of catalogs available on the website (all peaks, non redundant and CRMs), how to browse the ReMap catalogs, download and search the database. This newly introduced interactive tour is a useful feature to better understand the ReMap database, the data content and its functionalities. The ReMap database can be browsed by using the navigation links on the left sidebar, or searched for individual TRs, specific biotypes (cell lines or tissues)

using the simple or advanced search box. With an increased number of datasets and species in 2022 the search engine has been rewritten allowing for faster search queries. Search results are returned in a paginated table along with TRs aliases, TF classification, experiment IDs and data source. Also this responsive search results table can be dynamically searched to refine results. Finally, the ReMap database can be queried programmatically using the RESTful API, which contains added functions in 2022.

Genomic tracks and videos

Since ReMap 2015 an annotation track is available on the UCSC Genome browser website (41,43) within public sessions or public hubs (Figures 2 and 3). These genome track hubs are an essential tool for the visualization of the expanding ReMap catalogs. Track hubs are a convenient, efficient, mechanism for importing the large ReMap collections of regulatory features, providing standards for data tracks across genome-browsing platforms. For each ReMap species, and in different assemblies (hg38, hg19, mm39, mm10, dm6, TAIR10), track hubs have been added in the public hubs listing of the UCSC Genome browser as well as deposited to the EMBL-EBI Track Hub Registry (https://trackhubregistry.org/, Supplementary Figure S10). This ReMap release comes along with the new track hub filtering options available from the UCSC Genome browser. ReMap users can now filter which peaks are displayed by selecting specific biotypes and/or by specific TRs (e.g. FOXA1 in MCF-7). This new feature enables better flexibility especially as the ReMap database increases in size. Additionally, the peak names displayed on the browser can be adapted for a better visualisation. To illustrate and explain these browsing options we added multiple videos. These will guide users to make the best of the ReMap catalogs. Furthermore, to represent the depth of regulatory regions, we added a density track on top of the ReMap catalog track. By adding these browsing features, our objective is to make it easier for researchers to explore the variety and the complexity of ReMap catalogs when combined with other biological tracks.

CONCLUSION AND FUTURE DIRECTIONS

This fourth release of the ReMap database pursues its commitment to provide manually curated datasets and high-quality regulatory catalogs for the research community. For this 2022 update, we processed 13 000 ChIP-seq datasets starting from raw data. With four species, this version of ReMap continues the long-term goal of maintaining accessible, browseable, high-quality regulatory catalogs. The ReMap 2022 database comes with many updates, (i) a substantially expanded human regulatory catalog followed by (ii) an update of the *Arabidopsis thaliana* regulatory catalog; (iii) a new regulatory atlas for *Mus musculus* with more than 7,000 datasets curated and processed; (iv) the first *Drosophila melanogaster* regulatory atlas for the Fly scientific community; (v) an improved CRM identification method, (vi) a conservative filtering approach to improve our catalogs; (vii) an improved user experience with an interactive tour and (viii) updated genomic track hubs in different assemblies for better visualization in genome browser. Finally, since 2018 each ReMap update is incorporated into the JASPAR pipeline to infer new and updated TF binding profiles (17,18).

Overall, the ReMap database reaches 327M binding regions across four regulatory atlases. We anticipate that upcoming functional studies will provide additional experimental regulatory evidence giving rise to even larger catalogs. While manual curation and annotation are the foundation of ReMap, we intend to evolve to more conservative approaches in future releases, such as directing our efforts towards redundant peaks to build robust regulatory regions. Our long term goal of providing qualitative high quality catalogs will focus on redundant occupancy evidence for new releases, rather than releasing quantitative catalogs consisting in incremental inventories. Depending on outcomes, future ReMap releases may be crossed with other regulatory catalogs (DNAse, Histone marks) to strengthen and filter the regulatory space.

FEEDBACK

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

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Nathalie Arquier and Laurent Perrin for their expertises and scientific discussions regarding the curation and annotation of *Drosophila melanogaster* ChIP-seq data. We would like to thank the JASPAR Team led by Anthony Mathelier from NCMM Norway for constant scientific feedback on the ReMap catalogues. Finally, we would like to thank Maximilian Haeussler, Gerardo Perez and the UCSC Genome informatics groups for their help with public track hubs and their latest hub development, also the Ensembl and Ensembl Plant group for their help with the Human, Mouse, Fly and Arabidopsis track hubs.

FUNDING

PhD Fellowship to F.H. from the Provence-Alpes-Côte d'Azur Regional Council (Région SUD); Institut National de la Santé et de la Recherche Médicale (INSERM); PhD Fellowship to P.D.L. from the French Ministry of Higher Education and Research (MESR); 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 for Open Access charge: INSERM, MarMaRa, this project has received funding from the Excellence Initiative of Aix-Marseille University - A*Midex a French "Investissements d'Avenir programme"- Institute MarMaRa AMX-19-IET- 007.

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. 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. Sarkans,U., Füllgrabe,A., Ali,A., Athar,A., Behrangi,E., Diaz,N., Fexova,S., George,N., Iqbal,H., Kurri,S. *et al.* (2021) From ArrayExpress to BioStudies. *Nucleic Acids Res.*, **49**, D1502–D1506.
4. Fukuda,A., Kodama,Y., Mashima,J., Fujisawa,T. and Ogasawara,O. (2021) DDBJ update: streamlining submission and access of human data. *Nucleic Acids Res.*, **49**, D71–D75.
5. 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.
6. 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.
7. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
8. ENCODE Project Consortium, Moore,J.E., Purcaro,M.J., Pratt,H.E., Epstein,C.B., Shoresh,N., Adrian,J., Kawli,T., Davis,C.A., Dobin,A. *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.
9. 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.
10. 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.
11. Kolmykov,S., Yevshin,I., Kulyashov,M., Sharipov,R., Kondrakhin,Y., Makeev,V.J., Kulakovskiy,I.V., Kel,A. and Kolpakov,F. (2021) GTRD: an integrated view of transcription regulation. *Nucleic Acids Res.*, **49**, D104–D111.
12. 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.
13. 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.
14. 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.
15. 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.
16. Chèneby,J., Ménétrier,Z., Mestdagh,M., Rosnet,T., Douida,A., Rhalloussi,W., Bergon,A., Lopez,F. and Ballester,B. (2020) ReMap 2020: a database of regulatory regions from an integrative analysis of human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res.*, **48**, D180–D188.
17. Castro-Mondragon,J.A., Riudavets-Puig,R., Rauluseviciute,I., Berhanu Lemma,R., Turchi,L., Blanc-Mathieu,R., Lucas,J., Boddie,P., Khan,A. and Manosalva Pérez,N. (2021) JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, https://doi.org/10.1093/nar/gkab1113.
18. Gheorghe,M., Sandve,G.K., Khan,A., Cheneby,J., Ballester,B. and Mathelier,A. (2019) A map of direct TF-DNA interactions in the human genome. *Nucleic Acids Res.*, **47**, 7715.
19. 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.
20. Tweedie,S., Braschi,B., Gray,K., Jones,T.E.M., Seal,R.L., Yates,B. and Bruford,E.A. (2021) Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res.*, **49**, D939–D946.
21. 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.
22. Côté,R., Reisinger,F., Martens,L., Barsnes,H., Vizcaino,J.A. and Hermjakob,H. (2010) The Ontology Lookup Service: bigger and better. *Nucleic Acids Res.*, **38**, W155–W160.
23. Bairoch,A. (2018) The Cellosaurus, a Cell-Line Knowledge Resource. *J. Biomol. Tech. JBT*, **29**, 25–38.
24. Howe,K.L., Contreras-Moreira,B., De Silva,N., Maslen,G., Akanni,W., Allen,J., Alvarez-Jarreta,J., Barba,M., Bolser,D.M., Cambell,L. *et al.* (2020) Ensembl Genomes 2020-enabling non-vertebrate genomic research. *Nucleic Acids Res.*, **48**, D689–D695.
25. Law,M. and Shaw,D.R. (2018) Mouse Genome Informatics (MGI) is the international resource for information on the laboratory mouse. *Methods Mol. Biol.*, **1757**, 141–161.
26. Larkin,A., Marygold,S.J., Antonazzo,G., Attrill,H., Dos Santos,G., Garapati,P.V., Goodman,J.L., Gramates,L.S., Millburn,G., Strelets,V.B. *et al.* (2021) FlyBase: updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Res.*, **49**, D899–D907.
27. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
28. 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.
29. 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.
30. Köster,J. and Rahmann,S. (2012) Snakemake–a scalable bioinformatics workflow engine. *Bioinforma. Oxf. Engl.*, **28**, 2520–2522.
31. Frankish,A., Diekhans,M., Jungreis,I., Lagarde,J., Loveland,J.E., Mudge,J.M., Sisu,C., Wright,J.C., Armstrong,J., Barnes,I. *et al.* (2021) GENCODE 2021. *Nucleic Acids Res.*, **49**, D916–D923.
32. Stamatoyannopoulos,J.A., Snyder,M., Hardison,R., Ren,B., Gingeras,T., Gilbert,D.M., Groudine,M., Bender,M., Kaul,R., Canfield,T. *et al.* (2012) An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol.*, **13**, 418.
33. Lesurf,R., Cotto,K.C., Wang,G., Griffith,M., Kasaian,K., Jones,S.J.M., Montgomery,S.B. and Griffith,O.L. (2016) ORegAnno 3.0: a community-driven resource for curated regulatory annotation. *Nucleic Acids Res.*, **44**, D126–D132.
34. Zerbino,D.R., Wilder,S.P., Johnson,N., Juettemann,T. and Flicek,P.R. (2015) The ensembl regulatory build. *Genome Biol.*, **16**, 56.
35. Zhou,Q., Liu,M., Xia,X., Gong,T., Feng,J., Liu,W., Liu,Y., Zhen,B., Wang,Y., Ding,C. *et al.* (2017) A mouse tissue transcription factor atlas. *Nat. Commun.*, **8**, 15089.
36. Schmidt,D., Wilson,M.D., Ballester,B., Schwalie,P.C., Brown,G.D., Marshall,A., Kutter,C., Watt,S., Martinez-Jimenez,C.P., Mackay,S. *et al.* (2010) Five-vertebrate ChIP-seq reveals transcription factor binding. *Science*, **328**, 1036–1040.
37. Ballester,B., Medina-Rivera,A., Schmidt,D., Gonzàlez-Porta,M., Carlucci,M., Chen,X., Chessman,K., Faure,A.J., Funnell,A.P., Goncalves,A. *et al.* (2014) Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *Elife*, **3**, e02626.
38. 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.
39. modENCODE Consortium, Roy,S., Ernst,J., Kharchenko,P.V., Kheradpour,P., Negre,N., Eaton,M.L., Landolin,J.M., Bristow,C.A., Ma,L. *et al.* (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, **330**, 1787–1797.

40. Howe,K.L., Achuthan,P., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R., Bhai,J. *et al.* (2021) Ensembl 2021. *Nucleic Acids Res.*, **49**, D884–D891.
41. Navarro Gonzalez,J., Zweig,A.S., Speir,M.L., Schmelter,D., Rosenbloom,K.R., Raney,B.J., Powell,C.C., Nassar,L.R., Maulding,N.D., Lee,C.M. *et al.* (2021) The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.*, **49**, D1046–D1057.
42. Rangwala,S.H., Kuznetsov,A., Ananiev,V., Asztalos,A., Borodin,E., Evgeniev,V., Joukov,V., Lotov,V., Pannu,R., Rudnev,D. *et al.* (2021) Accessing NCBI data using the NCBI Sequence Viewer and Genome Data Viewer (GDV). *Genome Res.*, **31**, 159–169.
43. 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.

2.2. ReMap versus les autres ressources

ReMap est un atlas de régions régulatrices répertoriant des données ChIP-seq issues d'entrepôts de données publics. Il existe d'autres bases de données qui ont un objectif similaire, chacune ayant des caractéristiques uniques. Dans cette section nous ferons une liste exhaustive de ces bases de données afin de mettre en valeur le catalogue ReMap.

ChIP-Atlas [326] est une base de données qui utilise également des données ChIP-seq pour cartographier les régions régulatrices de l'ADN. Il se concentre principalement sur les facteurs de transcription chez 7 espèces modèles en particulier l'Homme, la drosophile et la souris, et utilise des données provenant d'entrepôts de données publics. Il permet également d'accéder à d'autres types de données de modification d'histone, l'ATAC-seq et le bisulfite-seq. Il fournit une annotation manuelle des métadonnées mais ne fait pas de contrôle qualité sur ces données.

Cistrome [324] est également une base de données similaire à ReMap. Il a été lancé en 2010 et utilise également des données ChIP-seq, DNase-seq et ATAC-seq pour cartographier les régions régulatrices de l'ADN chez l'Homme et la souris. L'annotation des métadonnées est automatique.

GTRD (Gene TRanscription Regulation Database) [156] est une base de données qui utilise des données ChIP-seq, ChIP-exo, DNase-seq, MNase-seq, ATAC-seq et RNA-seq identifier les domaines régulateurs de l'ADN chez l'Homme, la souris et la drosophile et 4 autres espèces modèles. L'annotation des données est automatique.

ReMap est une base de données qui propose un catalogue de données de région régulatrice pour quatres espèces modèles : l'Homme, la souris, la drosophile et *A. thaliana*. Il se distingue des autres bases de données comme Cistrome, GTRD et ChIP-atlas de plusieurs manières. Tout d'abord, ReMap propose une annotation et une curation manuelle des métadonnées, contrairement à GTRD et Cistrome qui utilisent des méthodes automatisées. Ensuite, ReMap traite les données du FASTQ au fichier BED de manière uniforme. Enfin, ReMap applique plusieurs contrôles qualités sur les données avant de les inclure dans la base de données, contrairement à ChIP-Atlas. Les caractéristiques et avantages de l'atlas ReMap seront présentés dans les sections suivantes.

2.3. Annotation et curation manuelle

ReMap se distingue des autres bases de données en proposant à la fois une annotation manuelle et un contrôle qualité des données, ce qui garantit l'intégrité et la qualité des données. Dans cette partie, nous allons décrire les étapes de l'annotation manuelle que j'ai réalisé pour les quatre catalogues.

L'annotation manuelle permet d'éviter les erreurs d'annotation automatique. Elle assure une homogénéité des données grâce à l'utilisation de protocoles et critères d'annotation normalisés. Elle permet une meilleure compréhension des résultats obtenus car nous pouvons également apporter des informations contextuelles et des connaissances expérimentales pour interpréter les résultats. Cela garantit que les données soient fiables, précises et pertinentes pour les chercheurs qui utilisent ces données pour leur propre recherche.

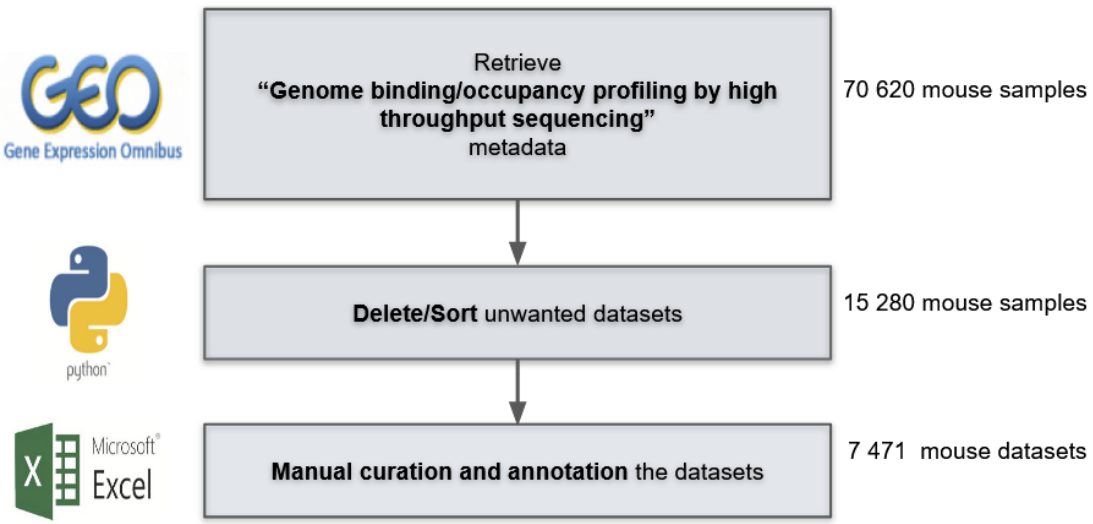


FIGURE 2.1. – *Étape de l'annotation des données ChIP-seq intégré au catalogue ReMap.*

L'annotation comprend plusieurs étapes (Figure 2.1). La première consiste à collecter les données disponibles sur GEO par un script python développé par un de nos collaborateurs. La requête utilisée pour la recherche GEO des données est "Genome binding/occupancy profiling by high throughput sequencing". Ce script génère un fichier contenant les métadonnées des expériences ChIP-seq correspondant à la requête (figure 2.2). Il contenait 70,620 lignes pour la souris, listant les expériences ChIP-seq de souris de 2009 à 2020.

	A	B	G	H	I	N	O	P	Q	R	S
1	accession	sample_title	source_name	descripti	Experiment_	rp_acc	bioproject_ac	Study_unimap_	experiment_id	instrument_pla	name
6107	GSE82144	aB_wt_Brg1	activated B cell		GSM2184295	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184295	ILLUMINA	Genome binding/occupancy profiling by high throughp
6108	GSE82144	aB_wt_Chd4	activated B cell		GSM2184296	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184296	ILLUMINA	Genome binding/occupancy profiling by high throughp
6109	GSE82144	aB_wt_GCN5	activated B cell		GSM2184297	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184297	ILLUMINA	Genome binding/occupancy profiling by high throughp
6110	GSE82144	aB_wt_HDAC1	activated B cell		GSM2184298	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184298	ILLUMINA	Genome binding/occupancy profiling by high throughp
6111	GSE82144	aB_wt_HDAC1_rep	activated B cell	replicate	GSM2184299	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184299	ILLUMINA	Genome binding/occupancy profiling by high throughp
6112	GSE82144	aB_wt_HDAC2	activated B cell		GSM2184300	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184300	ILLUMINA	Genome binding/occupancy profiling by high throughp
6113	GSE82144	aB_wt_Mll1	activated B cell		GSM2184301	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184301	ILLUMINA	Genome binding/occupancy profiling by high throughp
6114	GSE82144	aB_wt_Wdr5	activated B cell		GSM2184302	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184302	ILLUMINA	Genome binding/occupancy profiling by high throughp
6115	GSE82144	aB_wt_p300	activated B cell		GSM2184303	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184303	ILLUMINA	Genome binding/occupancy profiling by high throughp
6116	GSE82144	rB_wt_Brg1	resting B cell		GSM2184309	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184309	ILLUMINA	Genome binding/occupancy profiling by high throughp
6117	GSE82144	rB_wt_Chd4	resting B cell		GSM2184310	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184310	ILLUMINA	Genome binding/occupancy profiling by high throughp
6118	GSE82144	rB_wt_GCN5	resting B cell		GSM2184311	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184311	ILLUMINA	Genome binding/occupancy profiling by high throughp
6119	GSE82144	rB_wt_HDAC1	resting B cell		GSM2184312	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184312	ILLUMINA	Genome binding/occupancy profiling by high throughp
6120	GSE82144	rB_wt_HDAC1_rep	resting B cell	replicate	GSM2184313	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184313	ILLUMINA	Genome binding/occupancy profiling by high throughp
6121	GSE82144	rB_wt_HDAC2	resting B cell		GSM2184314	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184314	ILLUMINA	Genome binding/occupancy profiling by high throughp
6122	GSE82144	rB_wt_Mll1	resting B cell		GSM2184315	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184315	ILLUMINA	Genome binding/occupancy profiling by high throughp
6123	GSE82144	rB_wt_Wdr5	resting B cell		GSM2184316	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184316	ILLUMINA	Genome binding/occupancy profiling by high throughp
6124	GSE82144	rB_wt_p300	resting B cell		GSM2184317	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184317	ILLUMINA	Genome binding/occupancy profiling by high throughp
6125	GSE82144	aB_wt_CTCF_rep1	activated B cell	replicate1	GSM2184318	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184318	ILLUMINA	Genome binding/occupancy profiling by high throughp
6126	GSE82144	aB_wt_CTCF_rep2	activated B cell	replicate2	GSM2184319	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184319	ILLUMINA	Genome binding/occupancy profiling by high throughp
6127	GSE82144	aB_wt_CTCF_rep3	activated B cell	replicate3	GSM2184320	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184320	ILLUMINA	Genome binding/occupancy profiling by high throughp
6128	GSE82144	rB_wt_CTCF_rep1	resting B cell	replicate1	GSM2184321	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184321	ILLUMINA	Genome binding/occupancy profiling by high throughp
6129	GSE82144	rB_wt_CTCF_rep2	resting B cell	replicate2	GSM2184322	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184322	ILLUMINA	Genome binding/occupancy profiling by high throughp
6130	GSE82144	rB_wt_CTCF_rep3	resting B cell	replicate3	GSM2184323	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184323	ILLUMINA	Genome binding/occupancy profiling by high throughp
6131	GSE82144	aB_wt_Rad21_rep1	activated B cell	replicate1	GSM2184324	SRP075985	PRJNA324130	UNI-GSE82144	UNI-GSM2184324	ILLUMINA	Genome binding/occupancy profiling by high throughp

FIGURE 2.2. – *Fichier tabulé après extraction des métadonnées provennant de la base données GEO.*

Nous intégrons uniquement les données de ChIP-seq de facteurs de transcriptions dans le catalogue ReMap. J'ai donc développé un script python afin de filtrer les données ATAC-seq, FAIRE-seq, DNase-seq et les données d'histones. Après avoir supprimé ces données, il ne restait plus que 15,280 lignes. J'ai ensuite parcouru manuellement chacune de ces lignes afin d'annoter les métadonnées, c'est à dire identifier le facteur de transcription cible et le nommer selon la nomenclature officielle (HGNC [232] pour l'Homme, MGI [78] pour la souris, Flybase [55] pour la mouche, Ensembl plants [32] et TAIR [99] database pour *A. thaliana*). Ensuite, pour homogénéiser les noms des lignées cellulaires et des biotypes chez l'Homme et la souris, nous avons utilisé les nomenclatures de BRENDA [258] et Cellosaurus [16]. Pour la drosophile, nous nous sommes appuyés sur les stades de développement de la mouche (Figure 2.3 et tableau 2.4) ainsi que sur Cellosaurus pour les lignées cellulaires, enfin pour *A. thaliana* nous nous sommes basés sur les annotations précédentes.

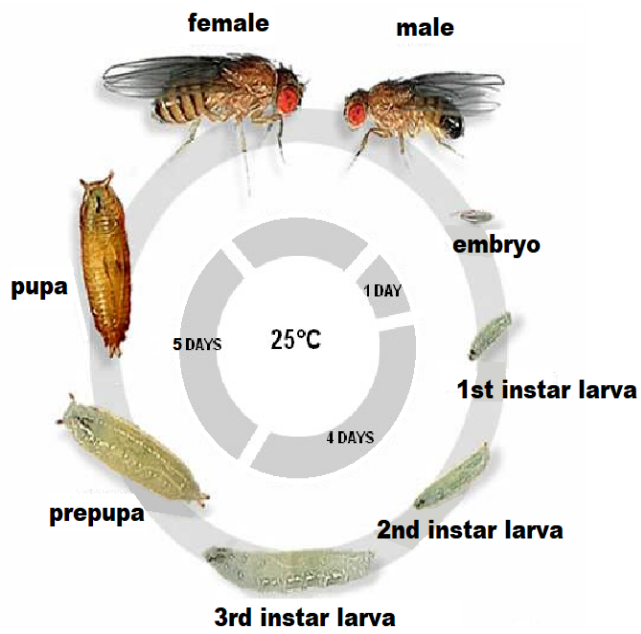


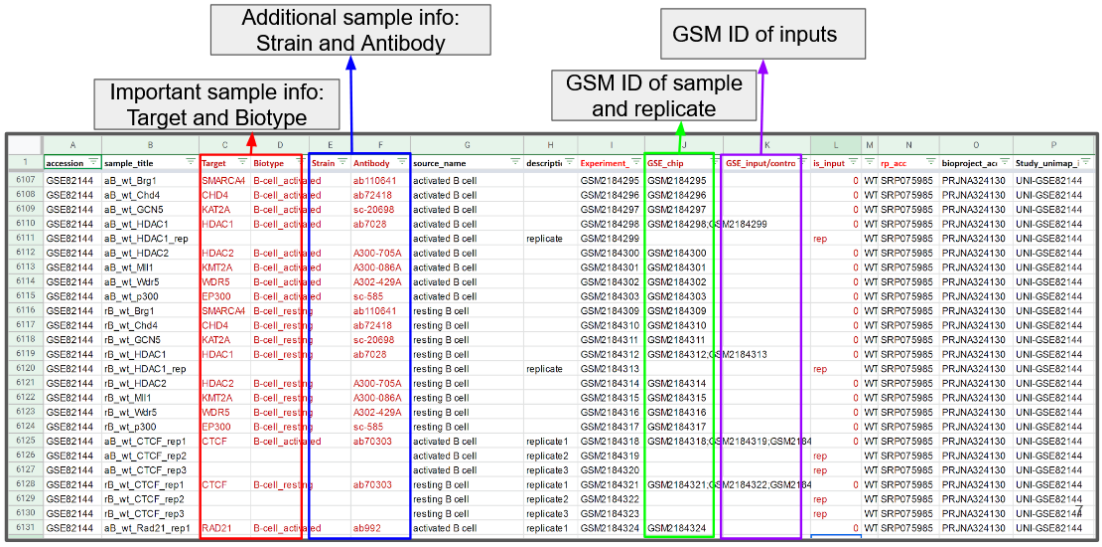
FIGURE 2.3. – *Cycle de vie de la drosophile*[1].

Biotype	Catégorie
adult	stade cellulaire
embryo	stade cellulaire
larva	stade cellulaire
pharate	stade cellulaire
prepupa	stade cellulaire
pupa	stade cellulaire
second-instar	stade cellulaire
third-instar	stade cellulaire
Kc	lignée cellulaire
Kc167	lignée cellulaire
ML-DmBG3-c2	lignée cellulaire
S2-DRSC	lignée cellulaire
S2R-plus	lignée cellulaire
Schneider-2	lignée cellulaire
Schneider-3	lignée cellulaire
mushroom-body	tissue
ovarian-somatic-cell	tissue

FIGURE 2.4. – *Liste des biotype de drosophile utilisé lors de l'annotation des données ReMap.*

1. https://www.cherrybiotech.com/scientific-note/drosophila-life-cycle-and-fly-anatomy

L'annotation manuelle des métadonnées est un processus très détaillé qui nécessite une grande attention aux détails. Parfois, pour compléter l'annotation, il est nécessaire de se référer aux articles scientifiques décrivant l'expérience originale. Cela peut inclure des informations sur les protocoles expérimentaux utilisés, les conditions de culture des cellules, les caractéristiques des échantillons utilisés, etc. Ces informations sont cruciales pour comprendre les données et les interpréter correctement. En utilisant des sources d'information complémentaires comme les articles scientifiques, nous pouvons garantir que les métadonnées soient complètes et précises. L'annotation implique également de reconstituer les datasets des expériences. Nous définissons un dataset comme étant une collection organisée de données provenant d'expériences scientifiques qui comprennent les réplicats et les inputs testés (Figure 2.5). Cette étape permet ensuite de traiter les données comme des datasets complets plutôt que des expériences indépendantes.



	accession	sample_title	Target	Biotype	Strain	Antibody	source_name	descripti	Experiment_	GSE_chip	GSE_input/contro	is_input	rp_acc	bioproject_ac	Study_unimap_j
6107	GSE82144	aB_wt_Brg1	SMARCA4	B-cell_activated		ab110641	activated B cell		GSM2184295	GSM2184295		0	WT SRP075985	PRJNA324130	UNI-GSE82144
6108	GSE82144	aB_wt_Chd4	CHD4	B-cell_activated		ab72418	activated B cell		GSM2184296	GSM2184296		0	WT SRP075985	PRJNA324130	UNI-GSE82144
6109	GSE82144	aB_wt_GCN5	KAT2A	B-cell_activated		sc-20698	activated B cell		GSM2184297	GSM2184297		0	WT SRP075985	PRJNA324130	UNI-GSE82144
6110	GSE82144	aB_wt_HDAC1	HDAC1	B-cell_activated		ab7028	activated B cell		GSM2184298	GSM2184298;GSM2184299		0	WT SRP075985	PRJNA324130	UNI-GSE82144
6111	GSE82144	aB_wt_HDAC1_rep					activated B cell	replicate	GSM2184299			rep	WT SRP075985	PRJNA324130	UNI-GSE82144
6112	GSE82144	aB_wt_HDAC2	HDAC2	B-cell_activated		A300-705A	activated B cell		GSM2184300	GSM2184300		0	WT SRP075985	PRJNA324130	UNI-GSE82144
6113	GSE82144	aB_wt_MI1	KMT2A	B-cell_activated		A300-086A	activated B cell		GSM2184301	GSM2184301		0	WT SRP075985	PRJNA324130	UNI-GSE82144
6114	GSE82144	aB_wt_Wdr5	WDR5	B-cell_activated		A302-429A	activated B cell		GSM2184302	GSM2184302		0	WT SRP075985	PRJNA324130	UNI-GSE82144
6115	GSE82144	aB_wt_p300	EP300	B-cell_activated		sc-585	activated B cell		GSM2184303	GSM2184303		0	WT SRP075985	PRJNA324130	UNI-GSE82144
6116	GSE82144	rB_wt_Brg1	SMARCA4	B-cell_resting		ab110641	resting B cell		GSM2184309	GSM2184309		0	WT SRP075985	PRJNA324130	UNI-GSE82144
6117	GSE82144	rB_wt_Chd4	CHD4	B-cell_resting		ab72418	resting B cell		GSM2184310	GSM2184310		0	WT SRP075985	PRJNA324130	UNI-GSE82144
6118	GSE82144	rB_wt_GCN5	KAT2A	B-cell_resting		sc-20698	resting B cell		GSM2184311	GSM2184311		0	WT SRP075985	PRJNA324130	UNI-GSE82144
6119	GSE82144	rB_wt_HDAC1	HDAC1	B-cell_resting		ab7028	resting B cell		GSM2184312	GSM2184312;GSM2184313		0	WT SRP075985	PRJNA324130	UNI-GSE82144
6120	GSE82144	rB_wt_HDAC1_rep					resting B cell	replicate	GSM2184313			rep	WT SRP075985	PRJNA324130	UNI-GSE82144
6121	GSE82144	rB_wt_HDAC2	HDAC2	B-cell_resting		A300-705A	resting B cell		GSM2184314	GSM2184314		0	WT SRP075985	PRJNA324130	UNI-GSE82144
6122	GSE82144	rB_wt_MI1	KMT2A	B-cell_resting		A300-086A	resting B cell		GSM2184315	GSM2184315		0	WT SRP075985	PRJNA324130	UNI-GSE82144
6123	GSE82144	rB_wt_Wdr5	WDR5	B-cell_resting		A302-429A	resting B cell		GSM2184316	GSM2184316		0	WT SRP075985	PRJNA324130	UNI-GSE82144
6124	GSE82144	rB_wt_p300	EP300	B-cell_resting		sc-585	resting B cell		GSM2184317	GSM2184317		0	WT SRP075985	PRJNA324130	UNI-GSE82144
6125	GSE82144	aB_wt_CTCF_rep1	CTCF	B-cell_activated		ab70303	activated B cell	replicate1	GSM2184318	GSM2184318;GSM2184319;GSM2184		0	WT SRP075985	PRJNA324130	UNI-GSE82144
6126	GSE82144	aB_wt_CTCF_rep2					activated B cell	replicate2	GSM2184319			rep	WT SRP075985	PRJNA324130	UNI-GSE82144
6127	GSE82144	aB_wt_CTCF_rep3					activated B cell	replicate3	GSM2184320			rep	WT SRP075985	PRJNA324130	UNI-GSE82144
6128	GSE82144	rB_wt_CTCF_rep1	CTCF	B-cell_resting		ab70303	resting B cell	replicate1	GSM2184321	GSM2184321;GSM2184322;GSM2184		0	WT SRP075985	PRJNA324130	UNI-GSE82144
6129	GSE82144	rB_wt_CTCF_rep2					resting B cell	replicate2	GSM2184322			rep	WT SRP075985	PRJNA324130	UNI-GSE82144
6130	GSE82144	rB_wt_CTCF_rep3					resting B cell	replicate3	GSM2184323			rep	WT SRP075985	PRJNA324130	UNI-GSE82144
6131	GSE82144	aB_wt_Rad21_rep1	RAD21	B-cell_activated		ab992	activated B cell	replicate1	GSM2184324	GSM2184324		0	WT SRP075985	PRJNA324130	UNI-GSE82144

FIGURE 2.5. – ***Fichier contenant l'annotation des métadonnées de datasets de souris.****Chaque colonne encadrée correspond à une information précise et annoté manuellement. Les colonnes rouges correspondent aux TF et biotypes. Les colonnes bleu correspondent aux informations complémentaires, la souche pour les souris et l'identifiant de l'anticorps utilisé lors de l'expérience ChIP-seq. La colonne verte contient les identifiants GSM provenant de GEO, plusieurs identifiants correspondent à des replicates. Enfin, la colonne violette correspond aux identifiants GSM des inputs.*

À titre d'exemple, après l'annotation et la curation des données de Souris, nous sommes passés de 70,620 lignes à 7,471 datasets. Ces données sont ensuite utilisées en entrée du pipeline Snakemake de ReMap.

2.4. Evolution du catalogue ReMap chez l'Homme et *A. thaliana*

Mettre à jour les catalogues ReMap pour l'Homme et *Arabidopsis thaliana* est un processus continu qui permet de maintenir l'exactitude et la pertinence des données disponibles pour la communauté scientifique. En 2020, le catalogue pour l'Homme était composé de 1,135 facteurs de régulation (TR) pour 602 biotypes, il contenait 5,798 datasets avec 165 millions de pics ChIP-seq. En 2022, le catalogue pour l'Homme est composé de 1,210 TR, 737 biotypes, ainsi que 8103 datasets, formant un catalogue de 182 millions de pics (Figure 2.6). Cette augmentation de 6.6% en termes de facteurs de transcription et 10% en termes de pics montre une évolution du catalogue de ReMap pour l'Homme entre 2020 et 2022 (après nouveaux QC, qui n'avait pas été effectué en 2020, 70% sans). Le catalogue *A. thaliana* a également été mis à jour en utilisant les dernières données ChIP-seq disponibles sur GEO. En 2020, il était composé de 372 TR et 2,6 millions de pics (Figure 2.6). En 2022, le catalogue est composé de 423 TR et 4,8 millions de pics. Le catalogue ReMap d'*A. thaliana* a donc connu une augmentation de 14% pour les TR et de 85% pour les pics de 2020 à 2022. L'augmentation de la quantité de données disponibles pour ces deux espèces, reflète les efforts continus pour améliorer la qualité des régions régulatrices chez l'Homme et *A. thaliana*.

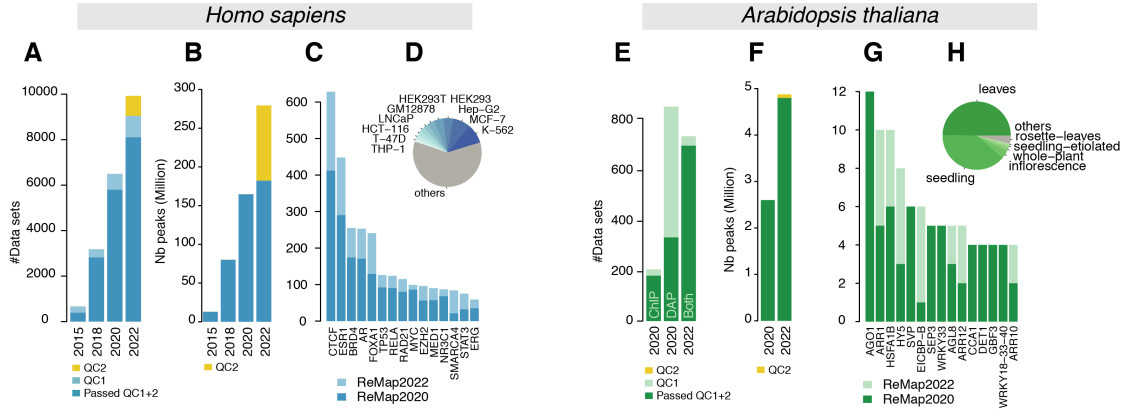


FIGURE 2.6. – ***Vue d'ensemble de la croissance de la base de données ReMap 2022 pour l'Homme et A. thaliana.*** *(A) Augmentation des datasets de l'Homme analysées dans ReMap 2022 par rapport à 2020, 2018 et 2015, les données filtrées par le nouveau contrôle qualité (QC2) sont en jaune. (B) Croissance des pics ChIP-seq de l'Homme dans ReMap 2022 par rapport aux versions précédentes, les pics filtrés sont en jaune. (C, D) Évolution du nombre de données pour les 15 principaux régulateurs de la transcription (TR) et biotypes entre ReMap 2022 et 2020. (E) Datasets A. thaliana analysées en 2022 par rapport à 2020. (F) Croissance des pics de régulation d'A. thaliana dans ReMap 2022 par rapport à 2020, les pics supprimés sont en jaune. (G, H) Évolution du nombre de datasets pour les 15 principaux TR et biotypes entre les deux catalogues ReMap A. thaliana.*

2.5. Régions régulatrices chez la souris

En 2020, ReMap était composé de deux catalogues de régions régulatrices pour l'Homme et la plante. En 2022, nous avons intégré et traité les données de ChIP-seq de souris publiquement disponibles dans NCBI-GEO, déposées entre 2009 et 2020. Le catalogue final est composé de 5503 datasets de 123 millions de pics ChIP-seq après avoir appliqué les contrôles qualités et les filtres. Les pics portent sur 648 TR dans 373 biotypes différents. Nous avons également proposé deux catalogues pour identifier les régions non-redondantes (43.9 millions pics) et CRM avec 2.4 millions de régions. La Figure 2.7 représente les statistiques portant sur le catalogue de souris. On constate que 412 régulateurs transcriptionnels sont en commun entre les catalogues de l'Homme et la souris. Tout comme pour les données ReMap chez l'Homme, CTCF, SUZ12, RAD21 et BRD4 font partie des TF les plus fréquemment étudiés.

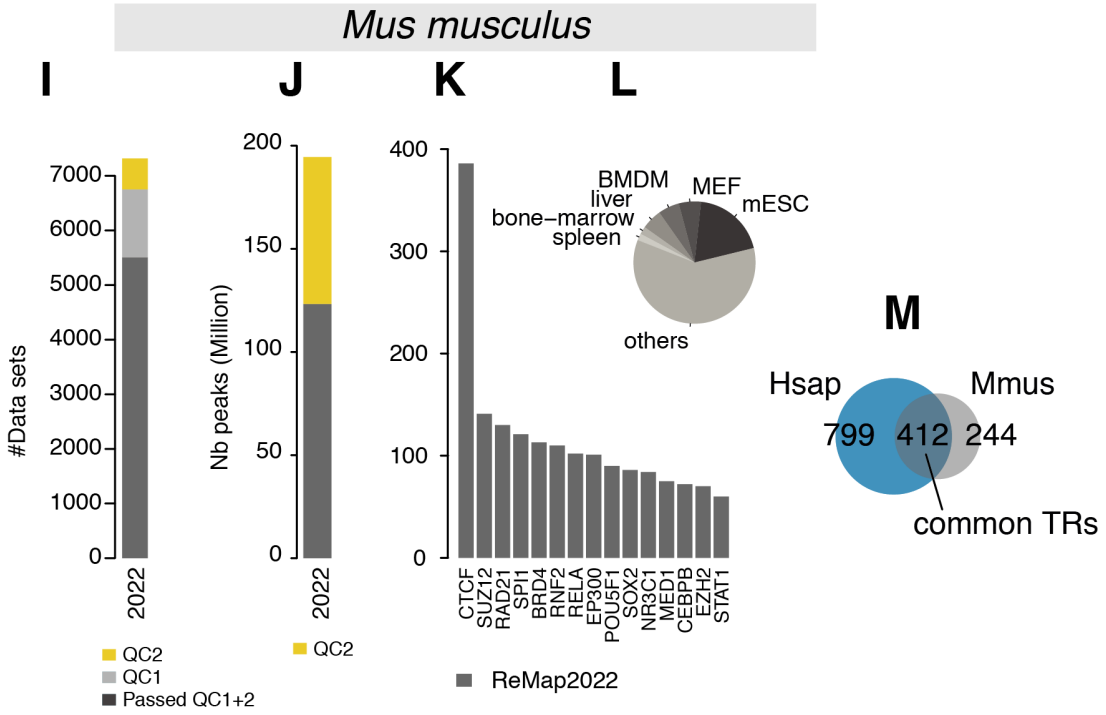


FIGURE 2.7. – ***Aperçu de la base de données ReMap 2022 de souris.*** *(I) Datasets ChIP-seq analysés pour le catalogue Mus musculus de ReMap 2022, les données retirés lors de l'étape QC2 sont représentés en jaune. (J) Nombre de pics du catalogue de Mus musculus, avant et après l'étape QC2 (en jaune). (K, L) Nombre de datasets pour les 15 principaux TR et les 6 principaux biotypes. (M) TR partagés entre les catalogues de l'Homme et la souris.*

2.6. Régions régulatrice chez la drosophile

L'ajout de données de Drosophile dans le catalogue de régions régulatrices de ReMap est une étape importante dans l'amélioration de la qualité et de la pertinence des données disponibles pour la communauté des drosophilistes. Nous avons collecté toutes les données de ChIP-seq de Drosophile disponibles publiquement entre 2008 et 2020, en utilisant des entrepôts de données tels que GEO et ENCODE. Parmi les données intégrées au catalogue ReMap de drosophile, nous avons collecté 59 expériences provenant du projet modENCODE [205] et 499 expériences du projet modERN [161]. Ces données étaient très hétérogènes, provenant de différents laboratoires et utilisant différentes méthodologies d'expérimentation.

Le catalogue définitif de la drosophile contient 1205 datasets et 16,6 millions de pics ChIP-seq. Ces derniers couvrent 550 TR dans 17 biotypes différents (Figure 2.8). En outre, le catalogue comprend 12,9 millions de pics non redondants et 617 567 pics de CRMs. La drosophile est une espèce de modèle de choix pour l'étude de la régulation de la transcription, en raison de sa simplicité génétique et de sa biologie bien caractérisée [242]. Les données de ChIP-seq ajoutées dans ReMap permettent donc d'étudier de manière plus approfondie les régions régulatrices chez cette espèce, ainsi que de les comparer aux régions régulatrices identifiées chez d'autres espèces, comme l'Homme, la souris et *A. thaliana*.

Les fichiers d'alignements au format BAM ont été partagés au Dr Elleen Furlong travaillant dans le laboratoire EMBL à Heidelberg, où ils étudient la régulation dans le génome de drosophile. Le Dr. Furlong utilisera les BAM de drosophile de ReMap afin d'effectuer des analyses de machine learning. Ce travail mènera à la publication d'un article sur lequel nous serons co-auteur.

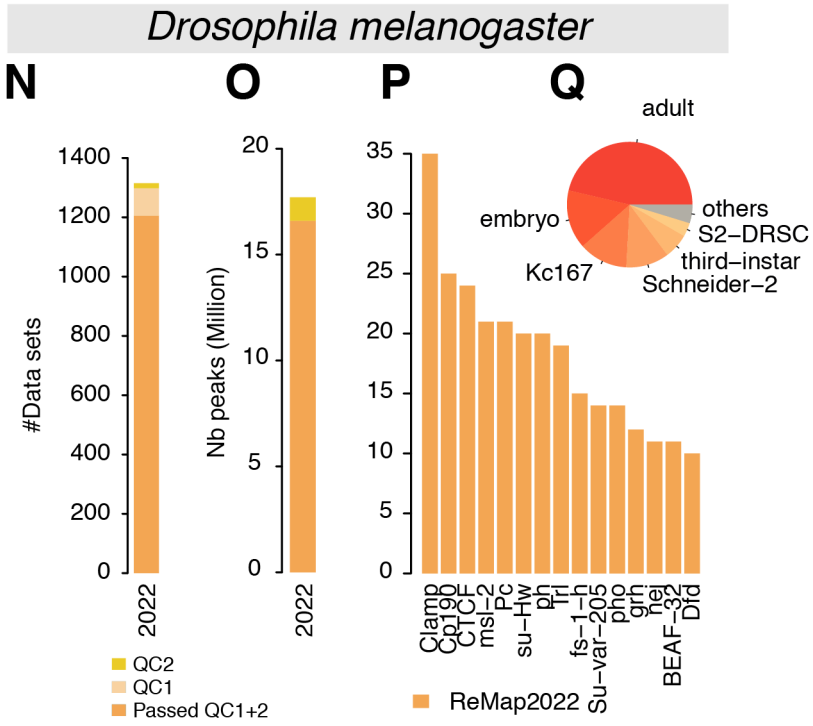


FIGURE 2.8. – ***Statistiques sur les données de drosophile de ReMap.*** *(N) Datasets ChIP-seq analysés pour le catalogue de Drosophila melanogaster de ReMap 2022, les données retirés lors de l'étape QC2 sont représentés en jaune. (O) Nombre de pics ChIP-seq de Drosophila melanogaster, avant et après l'étape QC2 (en jaune). (P, Q) Nombre de datasets pour les 15 principaux TR et les 6 principaux biotypes.*

2.7. Evolution du pipeline ReMap

Le pipeline de ReMap a été développé en 2018/2019 par Jeanne Cheneby, la deuxième étudiante ayant travaillé sur le projet. Pour la mise à jour de 2022, j'ai été la première étudiante a utilisé le pipeline. J'ai été amené à effectuer des modifications sur les scripts afin de faciliter l'utilisation et corriger les erreurs. J'ai aussi intégré au Github les scripts complémentaires à ReMap afin de créer le fichier CRM, et appliquer les nouveaux contrôles qualités (voire section 2.8). De plus, j'ai réorganisé le Github afin de faciliter la navigation et l'utilisation du pipeline. Il est disponible publiquement sur Github [1]. Le pipeline de ReMap est basé sur Snakemake, un outil de gestion de workflow pour les pipelines de bioinformatique. Il comprend plusieurs étapes, notamment la préparation des données, leur analyse, la curation manuelle des métadonnées et la génération de catalogues de régions régulatrices.

La première étape consiste à télécharger les métadonnées des expériences ChIP-seq depuis GEO ou ENCODE via un script python. Ensuite, nous procédons à la curation manuelle des métadonnées. Cela permet de réunir les réplicats et les inputs en datasets, d'annoter et de curater les données de manière homogène. L'annotation des métadonnées ChIP-seq permet de générer un fichier contenant toutes les informations nécessaires au traitement pour chaque datasets. Ce fichier est utilisé en entrée du pipeline ReMap. Ce pipeline est un ensemble d'outils développés pour l'analyse de données ChIP-seq et permet de traiter les données uniformément.

1. https://github.com/remap-cisreg/remap-pipeline

Les étapes générales du processing des données ChIP-seq sont les suivantes :

- Téléchargement des données brutes de ChIP-seq au format FASTQ depuis l'entrepôt de données de l'EBI/ENA via l'outil aria2. Cette étape permet de collecter les données expérimentales de manière automatisée et fiable. En effet, les fichiers FASTQ ne sont pas toujours disponibles sur GEO et ne sont pas nommées de manière standardisée.

- Couper les adaptateurs avec l'outil trim galore. Cela permet d'éliminer les séquences d'adaptateur qui ont été ajoutées lors de la préparation de la librairie et qui peuvent causer des erreurs dans l'alignement ultérieur.

- Alignement des fragments de séquençage sur le génome cible en utilisant Bowtie2. Cette étape permet de localiser les séquences d'ADN ChIP-seq sur le génome de référence et d'identifier les régions d'intérêt.

- Conversion du fichier SAM en fichier BAM avec l'outil SAMtools. Cette étape permet de convertir les données d'alignement en un format binaire, moins lourd pour les étapes ultérieures.

- Élimination des mismatchs, tri et filtrage des duplicats PCR afin de nettoyer les données en éliminant les erreurs d'alignement et les reads en double pour obtenir une meilleure précision des résultats.

- Élimination des reads non spécifiques en utilisant l'outil MACS2 pour identifier les pics ChIP-seq. MACS2 effectue un prétraitement des pics pour éliminer les artefacts tels que les régions de bruit élevé ou les régions de basse qualité et élimine les pics non significatifs. Cette étape permet de sélectionner les régions d'intérêt sur lesquelles la protéine cible s'est fixée.

Des mises à jour ont été effectuées pour la version 2022 de ReMap, qui incluent l'ajout d'exemples des fichiers d'input du pipeline, la correction des erreurs existantes dans le format des fichiers de configuration de snakemake, et l'ajout d'un nouveau script permettant de générer les fichiers CRM (*Cis* rehulatory Module). Ces améliorations ont pour but de faciliter la gestion et l'analyse des données ChIP-seq pour les utilisateurs de ReMap.

Le stockage des données (fichiers) est effectué sur le site de ReMap [1]. Les données peuvent être visualisées en utilisant des outils tels que UCSC Genome Browser et ENSEMBL genome browser (disponibles uniquement pour les versions antérieures à ReMap 2022).

1. https://remap2022.univ-amu.fr/

2.8. Nouveaux contrôles qualités

2.8.1. L'ajout de données affine les modules de régulation

En observant la Figure 2.9, qui compare les pics de 2020 aux pics de 2022, il est évident que l'augmentation de la quantité de données collectées n'a pas entraîné une répartition différente des pics de fixation à l'ADN. Les pics s'allignent sur les mêmes régions. En effet, l'ajout de pics permet d'affiner les clusters de pics et précise davantage la position des CRM.

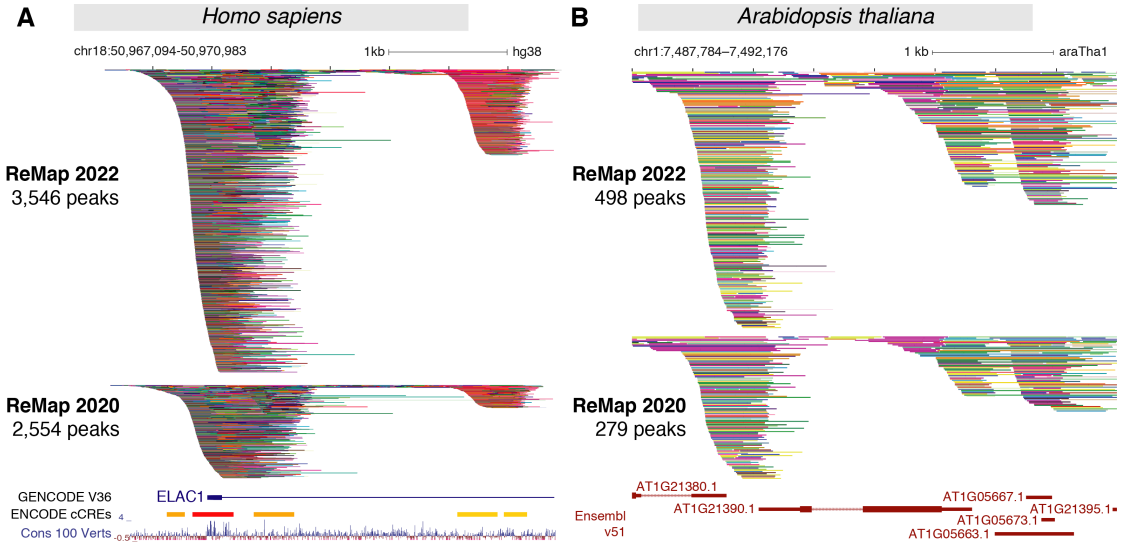


FIGURE 2.9. – ***Atlas ReMap 2022 pour l'Homme et la plante.*** *(A) Pics ReMap 2022 à partir des 8103 datasets humains. Cet exemple de navigateur de génome illustre la profondeur des pics de l'atlas ReMap 2022 comparé à ReMap 2020 au voisinage du promoteur ELAC1 (chr18 :50 967 094-50 970 983). Les pistes affichées sont condensées en lignes fines pour permettre la comparaison de la profondeur des liaisons ReMap 2022 et 2020. Autour de cette région ELAC1, ReMap 2022 affiche 3,546 pics, tandis que la version 2020 en contient 2,554. Les pistes du génome suivantes correspondent à l'annotation GENCODE v36, aux éléments cis régulateurs candidats ENCODE (cCRE, promoteurs en rouges, enhancer proximaux en orange, enhancer distaux en jaune) et à la conservation de 100 bases entre vertébrés, montrant les régions prédites comme conservées (scores positifs en bleu). (B) Vue du navigateur de génome de l'atlas de TF d'Arabidopsis de ReMap 2022 par rapport à la version 2020 au voisinage du gène modèle AT1G21390.1 (chr1 :7 487 784-7 492 176). La piste d'annotation du génome correspond à la dernière annotation de gène Ensembl Plants v51 TAIR10. Tous les pics ont été compactés pour le rendu visuel.*

La Figure 2.10 illustre les résultats du calcul du nombre de régions régulatrices (CRM) en fonction d'un nombre croissant de régulateurs transcriptionels pris au hasard (110 répétitions à chaque étape). On peut constater qu'à partir d'un seuil d'environ 600 TR, le nombre de CRMs atteint un plateau. Cela corrobore l'hypothèse selon laquelle l'ajout de nouveaux TR ne contribue pas significativement à l'augmentation du nombre de CRMs.

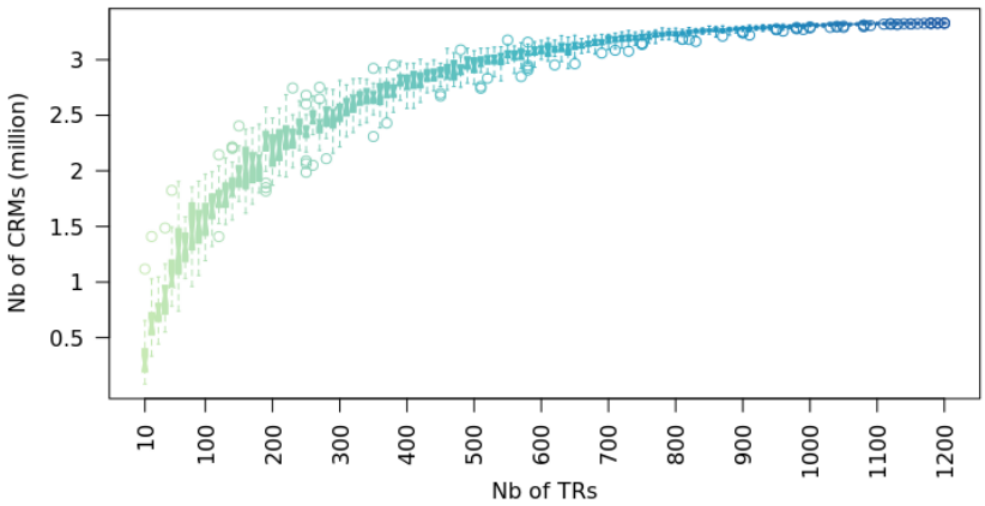


FIGURE 2.10. – ***Analyse de saturation des données ReMap avec l'augmentation du nombre de TR.*** *Ce graphique illustre la saturation des CRM identifiés par ChIP-seq des TR, à mesure que des facteurs supplémentaires sont analysés à travers l'analyse intégrative multicellulaire. Nous calculons le nombre de CRM dans l'ensemble du génome à partir d'un nombre croissant de TR sélectionnés au hasard. La distribution du nombre de CRMs pour une sélection de 100 TF est représentée sur l'axe des x sous forme de boîtes à moustaches. Nous continuons à le faire pour toutes les étapes incrémentales, y compris l'ensemble des 1200 TR.*

2.8.2. Les contrôles qualité standard

Depuis le commencement du projet, ReMap met un point d'honneur à inclure des données de haute qualité. Les contrôles qualité (QC) sont donc une étape cruciale dans l'analyse des données ChIP-seq, car ils permettent de vérifier la qualité et la fiabilité des données obtenues. Les deux principaux contrôles qualité utilisés pour les données ChIP-seq sont le FriP (False discovery rate of peaks) et le NSC/RSC (normalized strand cross-correlation/relative strand cross-correlation).

FRiP est un indicateur de la qualité de la détection des pics (peak calling) dans les données ChIP-seq. Il mesure le taux de faux positifs parmi les pics détectés, c'est-à-dire la proportion de pics détectés qui ne correspondent pas à des régions réelles de fixation de la protéine ciblée. Un taux élevé de faux positifs peut indiquer des problèmes de qualité de la librairie, de la détection de pics ou de l'analyse des données. Il est donc important d'avoir un score FriP bas pour avoir des données de qualité.
NSC (non-specific control) est un contrôle qui mesure le niveau de bruit non spécifique dans les données de séquençage. Ce bruit peut être dû à des contaminants tels que les produits d'amplification non ciblés, les artefacts de l'instrument de séquençage ou les séquences non spécifiques dans les sondes. L'analyse NSC permet de déterminer le niveau de bruit général dans les données et de déterminer si ce bruit est acceptable pour l'analyse.

RSC (RNA spike-in control) est un contrôle qui mesure la qualité de l'analyse et la reproductibilité des résultats. Ce contrôle consiste à ajouter des échantillons de référence connus, appelés "spike-in", à chaque échantillon d'analyse. Les échantillons de spike-in sont utilisés pour normaliser les données et corriger les biais d'analyse. L'analyse RSC permet de déterminer la qualité de l'analyse et de vérifier que les résultats sont fiables et reproductibles.

Un rapport NSC/RSC élevé signifie que le niveau de bruit non spécifique (NSC) est plus élevé que le niveau du contrôle de référence (RSC). Cela peut indiquer un problème avec les données de séquençage, tels que des contaminants ou des artefacts, qui peuvent affecter la qualité des résultats. En résumé, les contrôles qualités NSC et RSC sont des outils importants pour évaluer la qualité des données de séquençage et garantir la fiabilité et la reproductibilité des résultats obtenus.

2.8.3. Vers un catalogue de meilleur qualité

Depuis 2022 après avoir obtenu ce catalogue nous avons décidé de miser sur la qualité des données plutôt que sur la quantité. Nous avons donc décidé de mettre en place de nouveaux filtres qualité. Le premier filtre (appelé QC2, Figure 2.11) porte sur la longueur des pics. Le deuxième filtre (appelé QC1, Figure 2.12) porte sur le nombre de pics par datasets. Ces filtres ont été ajoutés aux filtres déjà existant décrits en amont. Ils ont été appliqués aux nouveaux catalogues (Souris et Drosophile), et rétroactivement pour l'Homme et *A. thaliana*.

Dans un premier temps, nous avons observé la fraction cumulée des longueurs de pics ainsi que la distribution des longeurs de pics à travers les datasets (Figure 2.11). En observant ces courbes, nous avons décidé de retirer les pics ayant des longueurs qui se situent au delà d'un seuil déterminé en fonction des données observées. Le seuil commun à toutes les espèces a été défini comme un minimum de 50pb et un maximum supérieur pour lequel nous avons 98% du catalogue, soit 1,5kb (pour l'Homme et la souris) ou 2kb (pour *A. thaliana* et la drosophile). Ces seuils ont été définis pour éliminer les pics aberrants tout en préservant la qualité du catalogue.

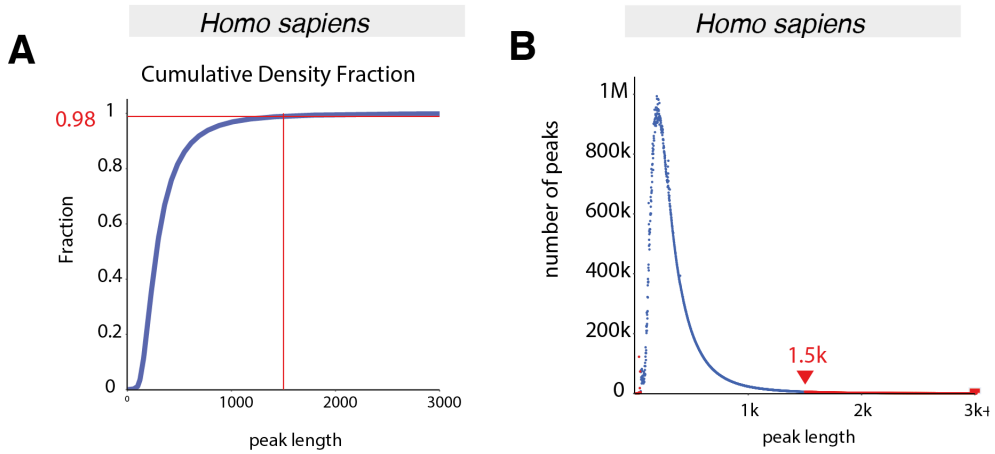


FIGURE 2.11. – ***Nouveaux filtres sur la longueur des pics.*** *A) Fonction de densité cumulative de la longueur des pics dans notre catalogue de l'Homme. Pour chaque datasets, nous avons éliminé les pics se situant en dehors d'une plage de longueurs en paires de bases. Cette plage est définie comme ayant une longueur minimale de 50 pb et une longueur maximale pour laquelle nous avons 99% de notre catalogue (ligne rouge horizontale). Ces longueurs maximales sont arrondies à 1,5 kb pour l'Homme (1 526 pb) et la souris (1 503 pb), 2 kb pour la drosophile (2 147 pb) et A. thaliana (2 347 pb). (B) Distribution de la longueur des pics, avec les pics situés en dehors des seuils représentés en rouge.*

Le deuxième contrôle qualité porte sur le nombre de pics par dataset. Pour observer la tendance des données nous avons construit un barplot représentant le nombre de datasets par nombres de pics par datasets (Figure 2.12 A) ainsi qu'un boxplot représentant les nombre de pics par datasets (Figure 2.12 B). On observe sur ces deux graphiques qu'il existe certaines expériences avec un nombre de pics très élevé, qui dépasse pour certaines expériences les 100 000 pics. Le seuil choisi pour ce contrôle qualité dépend du nombre de gènes codant de l'espèce concernée. Nous avons décidé d'enlever les datasets ayant un nombre de pics supérieur à deux fois le nombre de gènes codant de l'espèce (80k pour l'Homme et la souris, 40k pour la *A. thaliana* et la drosophile).

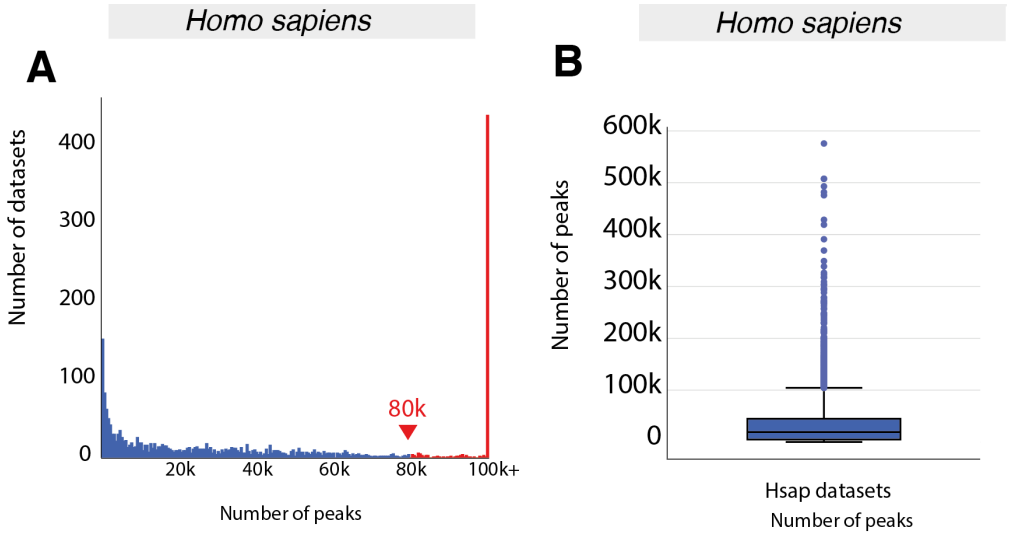


FIGURE 2.12. – ***Graphiques représentant le filtre qualité portant sur le nombre de pics par datasets.*** *A) En 2022, nous avons écarté les datasets contenant moins de 100 pics ou plus de deux fois le nombre de gènes annotés selon les statistiques d'annotation des gènes Ensembl. La flèche rouge représente la limite appliquée, et les barres rouges représentent les ensembles de données supprimées. (B) Distribution en boxplot du nombre de pics par datasets avant filtrage.*

En appliquant ces nouveaux filtres sur les données ReMap, nous avons pu éliminer les pics superflus et les expériences ChIP-seq, probablement, de mauvaise qualité. Cela a permis d'obtenir un catalogue plus précis et plus fiable pour la définition des régions régulatrices. La Figure 2.13 montre l'amélioration significative de la qualité des données ReMap en comparant les données de 2020 (à gauche) avec celles de 2022 (à droite) après l'application des nouveaux filtres. On peut constater que les pics de longueur excessive ont été supprimés, ce qui rend le catalogue plus clair en 2022.

FIGURE 2.13. – ***Avant/après nouveaux filtres qualités.*** *Capture d'écran des trackhub ReMap avant et après les nouveaux filtres. En rouge sont entourés les pics éliminés lors du filtre dû à une longueur de pics supérieur au seuil choisi lors du QC.*

2.9. Mise à disposition de ReMap : Trackhub UCSC

Depuis 2020 la visualisation des données ReMap est possible par l'intermédiaire d'un trackhub construit sur le navigateur de génomes UCSC. Ce trackhub était partagé via une session UCSC et était également disponible dans le navigateur de génomes d'Ensembl. Cependant même si le navigateur de génomes d'Ensembl est un navigateur puissant pour exploration des données génomique, il possède des limites en termes de quantité de "features" (pics) qu'il peut afficher dans une région. Cette limitation empêche l'utilisation effective d'un trackhub Ensembl en 2022.

En 2022 j'ai donc créé un trackhub UCSC sur le serveur local dans notre data-center de Marseille-Luminy (permet de ne pas dépendre de la session). J'ai également ajouté une piste avec la densité des pics de ReMap, car le nombre de pics étant conséquent à certaines positions du génome, il était difficile de visualiser la profondeur des pics entre eux.

La nouveauté la plus importante pour le trackhub est notamment l'ajout d'une fonctionnalité, le filtre des données. En utilisant ces fonctions de filtrage, les utilisateurs peuvent cibler les données qu'ils souhaitent afficher et analyser dans le navigateur de génomes UCSC, ce qui facilite la recherche de données génomiques spécifiques. La Figure 2.14 montre une capture d'écran du trackhub de ReMap affichant la fenêtre de paramétrage permettant de filtrer les données. Les filtres peuvent être appliqués sur les TF ou sur les biotypes de deux manières différentes, soit avec une recherche manuelle soit avec une liste déroulante.

Pour se faire nous avons ajouté dans le fichier trackDb disponible en local, les fonctions `filter.fieldName`, `filterText.fieldName` et `filterValues.fieldName`. Ces fonctions sont utilisées pour filtrer les données dans le navigateur de génomes UCSC. La fonction `filter.fieldName` permet de filtrer les pics en fonction d'un champ spécifique, tels que le nom de gène ou le biotype. La fonction `filterText.fieldName` est similaire à `filter.fieldName`, mais elle permet de rechercher des chaînes de caractères spécifiques dans le champ spécifié, tel que par exemple GATA* qui affiche tous les pics de fixation des TF de la famille GATA comme GAT1, GATA2, GATA3... etc. (Figure 2.14).

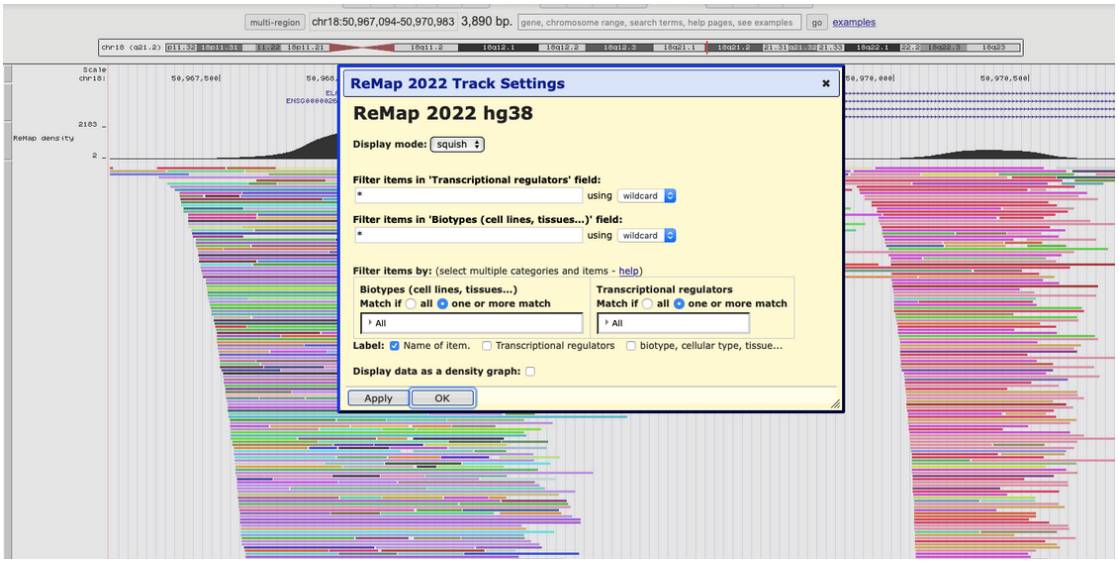


FIGURE 2.14. – ***Capture d'écrans des paramètres de filtrage du trackhub UCSC.*** *Plusieurs options sont disponibles. Les données peuvent être filtrées en fonction des TF ou des biotypes. Les filtres peuvent être appliqués à l'aide d'une barre de recherche ou une liste déroulante.*

Le résultat des filtres est montré à titre d'exemple dans la Figure 2.15. Dans cette figure nous avons filtré le catalogue pour n'avoir que les pics du complexe CTCF/cohesine tel que CTCF, RAD21 et les TF SMC pour toutes les lignées/biotypes.

FIGURE 2.15. – ***Capture d'écran du trackhub de l'homme avec comme filtres les TF du complexe CTCF/cohesin.*** *Ces filtres ont été appliqués sur les TF suivants : CTCF, CTCFL, NIPB, RAD21, RAD51, SMC1A, SMC1B et SMC3.*

Ces filtres permettent de faciliter l'utilisation du catalogue ReMap et se concentrer sur les TF ou biotypes d'intérêt pour les travaux de l'utilisateur. Dans la Figure 2.16 nous avons filtrés ReMap pour afficher uniquement les pics provenant de lignées cellulaires du cancer du sein. Pour se faire, nous avons sélectionné une liste de biotypes tels que MCF-7 et MCF-10A.

FIGURE 2.16. – ***Capture d'écran du trackhub de l'homme avec comme filtres les biotypes breast cancer.*** *Ces filtres ont été appliqués sur les biotypes suivants : MCF-7 et MCF-10A.*

Depuis Avril 2022, les données de ReMap sont également disponible sur le navigateur de génome UCSC "en natif", c'est à dire intégré aux données de régulation de UCSC, ce qui représente de nombreux avantages. Tout d'abord, il permet un accès direct et immédiat aux données, sans avoir besoin de se connecter à une session UCSC partagée. Cela facilite l'accès à ReMap et améliore la rapidité et la flexibilité de l'exploration des données.

2.10. Analyses complementaires de ReMap

2.10.1. Distribution des CRMs dans les biotypes Gtex

Pour approfondir l'identification des régions régulatrices, nous avons utilisé ReMap et effectué des analyses complémentaires. Une première analyse très simple consistait à identifier les tissus les plus utilisés dans les expériences ChIP-seq. Pour cela, nous avons utilisé l'annotation des tissus GTEx pour uniformiser les lignées/tissus de ReMap en 23 biotypes avec vocabulaire contrôlé. GTEx (Genotype-Tissue Expression) [180] est un projet de recherche en génomique qui a pour but de cartographier la variation de l'expression génique dans les différents tissus et cellules humaines. GTEx a collecté des échantillons de tissus de différents donneurs et a mesuré l'expression de milliers de gènes dans chaque échantillon. Les données sont disponibles au grand public et peuvent être utilisées pour étudier la manière dont les gènes sont activés ou inhibés dans les différents types de tissus, ce qui peut aider à comprendre les mécanismes sous-jacents aux maladies et à développer des stratégies thérapeutiques ciblées. En utilisant cette annotation, nous avons pu obtenir une vue plus complète des biotypes les plus utilisés dans les expériences ChIP-seq (Figure 2.17) et ainsi continuer notre analyse de manière plus efficace. Cette annotation a été obtenue en regroupant les lignées/tissus de ReMap en biotypes d'où proviennent les différents types cellulaires. Grâce à ce regroupement nous passons de plus de 600 biotypes à seulement une trentaine, ce qui simplifie les analyses.

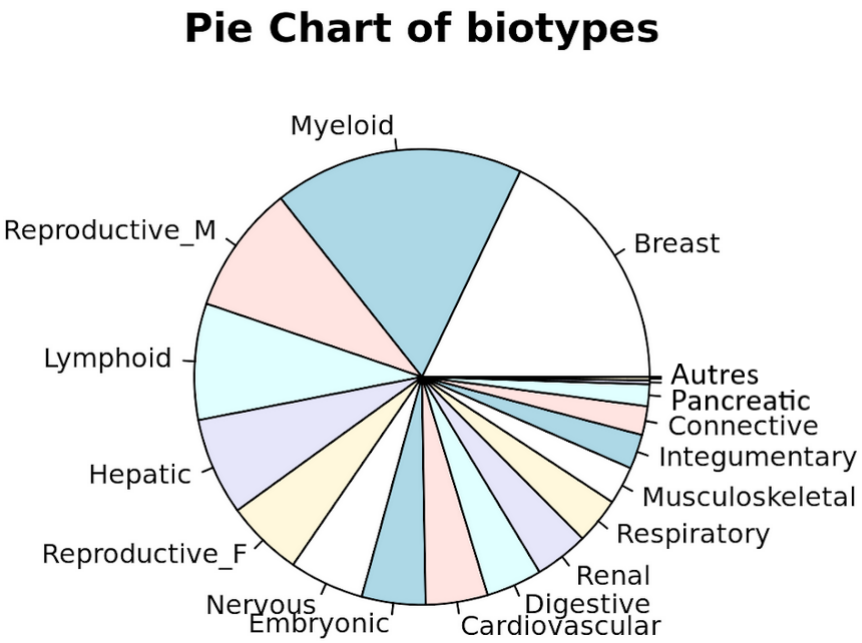


FIGURE 2.17. – ***Pie chart des biotypes GTEx de ReMap.*** *Ce graphique a été construit grâce au package R ChIPPeakAnno.*

La pie chart de la Figure 2.17 représente les différents biotypes dans lesquels ont été réalisées des expériences de ChIP-seq de ReMap. Il est possible de voir sur ce graphique que les tissus les plus représentés dans les expériences de ChIP-seq sont les tissus du sein et les tissus myéloïdes, qui représentent chacun environ 15% des expériences. Les tissus suivants sont les tissus lymphoïdes et hépatiques, qui représentent environ 10% des expériences. Les autres tissus représentés sont les tissus reproducteurs féminin, le système nerveux, les cellules embryonnaires, le cœur, l'appareil digestif et la peau. Il est possible de voir que les autres tissus ont des pourcentages moins importants, moins de 5% chacun.

Dans de nombreux projets de recherche utilisant la technique ChIP-seq, différents types de tissus sont étudiés pour identifier les régions de l'ADN qui sont spécifiques à chaque tissu. Les tissus les plus étudiés en ChIP-seq sont généralement les tissus les plus accessibles et les plus abondants, ainsi que les tissus qui sont impliqués dans des maladies spécifiques.

2.10.2. Distribution des CRMs ReMap dans les régions génomiques

L'analyse que nous allons décrire dans cette partie, porte sur la distribution génomique des modules cis-régulateurs (CRMs) identifiés par ReMap. La Figure 2.18 présente une illustration de la distribution des CRMs de ReMap à l'aide de BEDtools et les annotations disponibles sur UCSC Genome Browser (CCDS, CpG islands). Les résultats montrent que les régions contenant le plus de CRMs sont les régions intergéniques et les introns. On remarque également que plus d'1 millions de nos CRMs chevauchent les éléments *cis*-régulateurs candidats (cCREs) d'ENCODE.

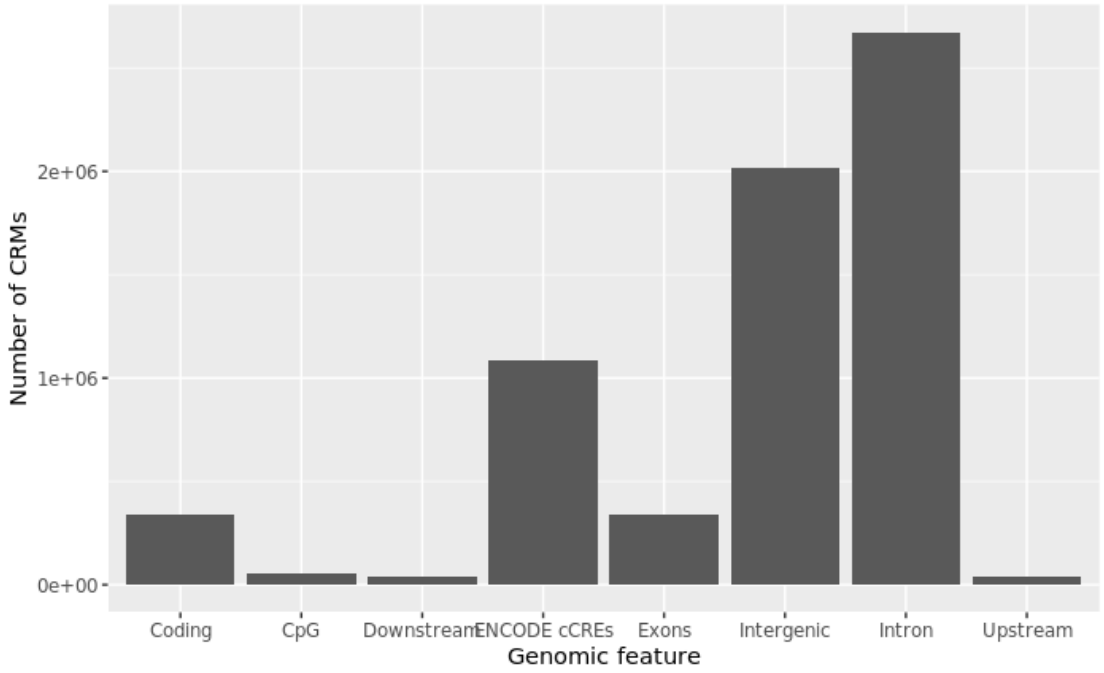


FIGURE 2.18. – ***Distribution genomique des CRMs.*** *Cette figure représente un barplot qui affiche le nombre de CRMs ReMap pour chaque type de régions génomiques dans le génome humain. Les régions ont été éxtraites de l'outil table Browser de UCSC, puis nous avons compté le nombre de CRM qui overlappe chaque feature à l'aide de BEDtools intersect.*

Les CRMs sont des modules de régulation obtenus à partir des données ReMap, ils représentent les régions une forte densité de pics de fixation des TF. Ils peuvent inclure des éléments tels que des promoteurs, des régions régulatrices *cis*- et *trans*-médiées, des enhancers ou encore des insulateurs. La distribution de ces modules de régulation dans le génome sont distribués de manière non uniforme dans le génome, avec certaines régions génomiques qui en contiennent plus que d'autres (Figure 2.18).

Ces résultats sont cohérents car les régions génomiques qui ont un rôle important dans la régulation de l'expression génique, sont généralement localisés à proximité des gènes qu'elles régulent et sont contenus dans les introns et les régions intergéniques. Les régions intergéniques sont des régions génomiques qui se trouvent entre les gênes, ils ne codent pas pour des protéines, mais peuvent contenir des éléments régulateurs importants tels que des enhancers et promoteurs. Les introns sont les régions que l'on identifie entre les exons et qui sont également non codants pour des protéines. Nous pouvons y trouver des éléments régulateurs tels que les promoteurs alternatifs.

Les travaux de B. Borsari et al. [33] examine la corrélation entre l'emplacement génomique des enhancers et leur rôle dans l'expression génique spécifique au tissu. Les auteurs révèlent que les enhancers spécifiques au tissu sont souvent situés dans des régions introniques et régulent l'expression de gènes impliqués dans des fonctions spécifiques au tissu, tandis que les gènes ubiquitaires sont plus souvent contrôlés par des enhancers intergéniques communs à de nombreux tissus. Les résultats montrent une transition d'enhancers actifs intergéniques à introniques au cours du développement, où les tissus les plus différenciés présentent des taux plus élevés d'enhancers introniques. Ces résultats suggèrent que l'emplacement génomique des enhancers actifs est crucial pour le contrôle spécifique au tissu de l'expression génique.

Nous pouvons nous poser la question de savoir si les régions introniques de ReMap sont également tissu-spécifiques et si les régions intergéniques régulent des gènes ubiquitaires. L'analyse suivante porte précisément sur la spécificité tissulaire des régions régulatrices. Nous avons également observé que certains CRMs chevauchent les cCREs d'ENCODE. Nous nous demandons donc si les CRMs qui ne chevauchent pas les cCREs pouraient permettre de caractériser de nouveaux éléments *cis*-régulateur.

2.10.3. Segmentation du génome

La segmentation du génome par ChromHMM [80] est une méthode utilisée pour décrire les régions génomiques en termes d'états épigénétiques tels que les modifications d'histone, la méthylation de l'ADN, et les structures chromatine. Cette méthode utilise des modèles statistiques pour décrire les états. Les modèles sont entraînés sur des données épigénétiques de référence pour identifier différents états les plus couramment observés dans les régions génomiques. Une fois entraînés, ces modèles sont utilisés pour segmenter les régions génomiques.

Lors de cette analyse nous cherchions à savoir s'il était possible d'identifier des régions régulatrices spécifiques à certains biotypes. Cette spécificité pourrait expliquer les différences entre les mécanismes de régulation des différents biotypes. Pour cela, nous avons voulu utiliser la méthode ChromHMM afin de segmenter le génome sur la base des régions régulatrices identifiées avec ReMap. Le but étant d'identifier des états biotypes spécifiques sur la base de données et ainsi de déterminer si certains CRM sont en fait spécifiques au biotypes.

Cette analyse peut être particulièrement utile pour les travaux de recherche sur régulation de la transcription impliquant des maladies, car de nombreuses maladies ont des perturbations spécifiques à un tissu [269, 222, 21]. Le modèle de segmentation que nous avons choisi est le modèle de 25 états car c'est celui qui a permis d'identifier le plus de biotypes différents. En utilisant ce nombre d'états, 11 biotypes différents ont pu être distingués (Figure 2.19).

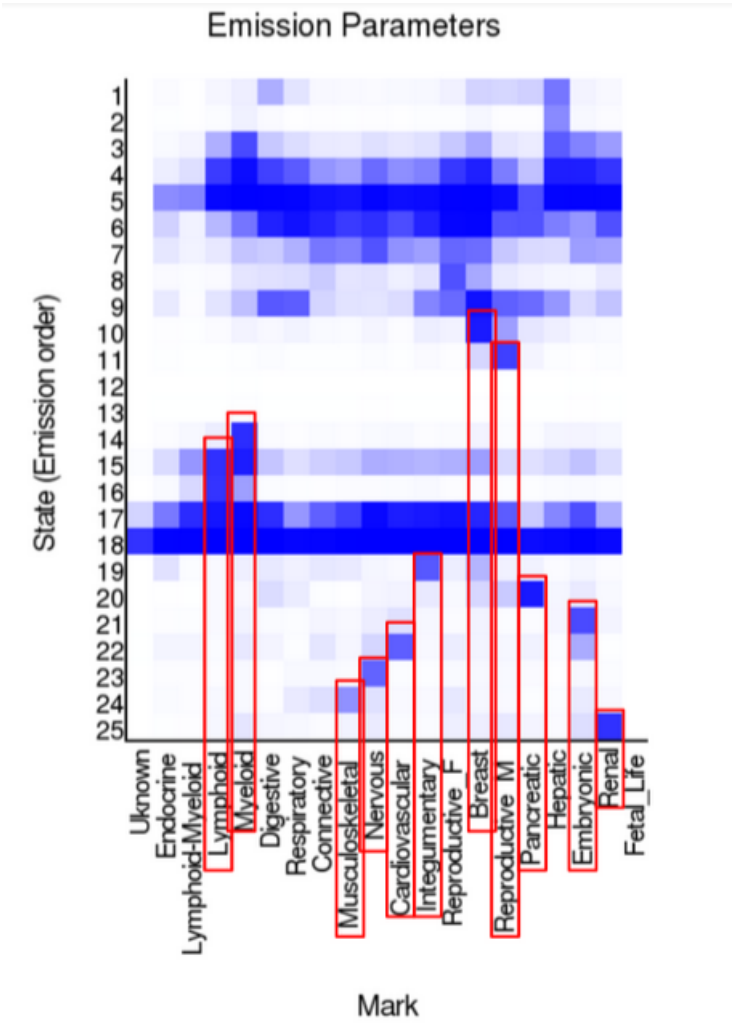


FIGURE 2.19. – *segmentation ChromHMM avec 25 état, permettant de distinguer 11 biotypes.*

Pour visualiser ces résultats, un trackhub a été mis en place, où la première piste représente l'état du CRM. Par exemple, sur la Figure 2.20 nous pouvons voir que l'état 10 du CRM est associé au tissu mammaire (Sein). En analysant les biotypes des pics fixés par ReMap pour ce CRM, on observe que la plupart correspondent effectivement à la lignée cellulaire modèle du cancer du sein MCF-7.

FIGURE 2.20. – ***Exemple d'un résultat de l'analyse chromHMM : le tissu du sein.*** *Capture d'écran de l'état 10 du chromHMM qui correspond au tissu du sein ("Breast") on peut remarquer que les pics correspondant à cet état sont des pics de biotypes de Breast.*

Attribuer des biotypes aux CRM permettrait d'identifier, parmi tous les CRM (n=3,4 millions), les régions régulatrices dont la fixation des TF seraient spécifiques à un tissu. Ce travail pourrait aider à identifier les différences entre les biotypes et les mécanismes de régulation qui en sont responsables.

Les régions régulatrices sont connues pour être tissu-spécifiques, comme le démontrent des travaux antérieurs [33]. D'autres études, datant de 2011, ont utilisé ChromHMM dans neuf tissus différents pour modéliser des enhancers et des promoteurs spécifiques à chaque tissu [81]. Cela prouve que notre analyse est pertinente. Cependant, il est important de noter que l'analyse ChromHMM n'est qu'à son commencement et nécessite encore beaucoup de travail que je n'ai pas eu le temps de continuer lors de ma thèse. J'ai tout de même tenu à présenter cette analyse car elle promet, j'en suis sûre, des résultats très utiles et intéressants pour la communauté scientifique si elle est menée jusqu'au bout.

2.11. Conclusion

Depuis sa première publication, ReMap a été largement utilisé par la communauté scientifique pour des travaux concernant la régulation de la transcription [46] et a été validé par de nombreux travaux complémentaires [261]. ReMap, depuis 2015, a été cité 580 fois.

Le catalogue ReMap est régulièrement mis à jour, et il est accessible à la communauté depuis 2015. Nous avons créé un atlas de régions régulatrices pour l'Homme, la souris, la drosophile et *A. thaliana* en réalisant une analyse intégrative à grande échelle des expériences ChIP-seq pour des centaines de régulateurs transcriptionnels. Il apporte à la fois une annotation et une curation manuelle de données provenant de sources différentes ainsi que des contrôles qualité.

La mise à jour du catalogue en 2022 a permis d'ajouter deux catalogues, ceux de la souris et de la drosophile permettant ainsi de toucher une plus grande communauté de chercheurs. La quantité de données étant conséquente nous avons décidé dans cette version de privilégier la qualité des données en ajoutant deux nouveaux filtres. Pour les prochaines mises à jour du catalogue, nous prévoyons d'ajouter de nouvelles espèces au catalogue afin de rendre disponible ce travail à une plus grande communauté de chercheurs.

Les données de ChIP-seq disponibles publiquement ne cessent d'augmenter. Entre 2020 et 2022, on observe une augmentation drastique de la quantité de données intégrées au catalogue ReMap. Pour les prochaines versions nous nous attendons à ce que cette augmentation continue. Il devient donc essentiel pour la pérennité du projet d'établir des règles afin de gérer ces données. La première limite que nous pouvons rencontrer réside dans l'annotation manuelle des données. Cette étape nécessite beaucoup de temps et risque d'être de plus en plus complexe à mesure que ReMap se développe. Une aide automatique permettrait de réduire le temps de traitement et de faciliter le travail des prochains annotateurs.

En 2022, nous avons amélioré la qualité des données ChIP-seq ReMap en ajoutant deux nouveaux filtres. Le premier filtre que nous avons ajouté porte sur la longueur des pics et le deuxième sur le nombre de pics par datasets. Environ 36% des pics du catalogue de l'homme ont été filtrés. Pour les prochaines mises à jour, il serait intéressant d'ajouter d'autres filtres pour améliorer encore plus la qualité des données ChIP-seq ReMap. Par exemple, nous pourrions ajouter un filtre pour éliminer les données qui ont été collectées dans les régions blacklisté d'ENCODE [10], qui peuvent introduire des artefacts dans les données.

Il serait également utile de valider les données ChIP-seq ReMap avec d'autres données génomiques. Les données de séquençage d'ADN méthylé ou de transcriptome pourraient être utilisées pour valider les régions liées aux gènes identifiées par ChIP-seq. De plus, des données de ChIP-seq supplémentaires pourraient être utilisées pour valider les résultats de ReMap (ATAC-seq, DNase-seq, etc.).

J'ai également mis en place l'intégration des données ReMap dans le navigateur de génome UCSC permettant une visualisation intuitive et interactive des données, facilitant la compréhension et l'analyse des résultats. Le trackhub permet maintenant une meilleure personnalisation des fonctionnalités, comme la possibilité de filtrer les pics en fonction des biotypes et/ou des facteurs de transcription. Ainsi les utilisateurs peuvent cibler le type de données d'intérêt.

Il semble que les régions génomiques les plus enrichies en CRM identifiées par ReMap, sont les régions intergéniques et les introns, ce qui suggère que ces régions jouent un rôle important dans la régulation. La segmentation ChromHMM a été utilisée sur les données ReMap pour essayer de déterminer la spécificité tissulaire des régions régulatrices. Les régions régulatrices peuvent être classées en différentes régions ou la fixation des TF semble tissus spécifiques, et des régions ou la fixation des TF apparaît "ubiquitaire". Le modèle retenu pour l'instant est celui à 25 états et il permet de distinguer 11 tissus différents. Ce travail est encore en cours et nécessite d'être approfondi. Il permettra aux chercheurs de se focaliser sur des régions génomiques qui seraient tissus-spécifiques, ces régions pourraient ensuite être validées expérimentalement.

3. TE dans le contexte de la régulation

Sommaire

3.1. Etat de l'art

Plusieurs travaux ont révélé que les éléments transposables peuvent jouer un rôle important dans la régulation en servant de sites de fixation pour les facteurs de transcription (TF). Dans cette partie nous tenterons de faire une liste exhaustive des différents articles traitant du sujet afin de donner un contexte à nos travaux.

Les travaux de 2006 dirigée par Paz Polak [231] ont examiné le rôle des éléments transposables Alu dans la régulation des processus de développement. Les résultats suggèrent que les éléments Alu peuvent jouer un rôle important dans la régulation de l'expression des gènes en agissant comme des sites de fixation pour les TF. Les auteurs concluent que ces travaux fournissent une nouvelle perspective sur le rôle des éléments transposables dans la régulation de la transcription et ouvrent la voie à des recherches futures dans ce domaine.

Dans la revue "Transposable elements and the evolution of regulatory networks" de Cédric Feschotte (2008) [89], l'auteur dépeint le rôle des éléments transposables dans l'évolution de la régulation. Les TEs peuvent altérer la fonction des gènes, ce qui peut entraîner des avantages évolutifs. Cette revue résume également des travaux précédents montrant que les TE peuvent servir de sites de fixations pour les facteurs de transcription dans divers organismes eucaryotes. En conclusion les éléments transposables jouent un rôle dans la complexité et l'évolution des réseaux de régulations.

Dans cette même année G.Bourque et al. publie un article sur l'évolution des TFBS de mammifères via les éléments transposables [35]. Cet article étudie les régions de fixation de sept facteurs de transcription (ESR1, TP53, MYC, RELA, POU5F1, SOX2 et CTCF) chez les mammifères et comment ils contribuent à la régulation de l'expression génique. Les résultats révèlent qu'une petite fraction des sites de fixation sont conservés au niveau des séquences, mais un grand nombre d'entre eux sont associés à différentes familles d'éléments transposables. Cette association est ainsi appelée "sites de fixation associés aux répétitions génomiques (RABS)". Les auteurs ont également identifié la signification fonctionnelle de ces RABS en démontrant qu'ils sont sur-représentés à proximité de gènes régulés et que les motifs de liaison à l'intérieur de ces répétitions ont subi une sélection évolutive. Leurs résultats prouvent que les réseaux de régulation transcriptionnelle sont très dynamiques dans les génomes des eucaryotes et que les éléments transposables jouent un rôle important dans l'expansion du répertoire des sites de fixation aux TF (TFBS).

En 2010, G. Kunarso et al. [162] ont publié un article intitulé "Transposable elements have rewired the core regulatory network of human embryonic stem cells" dans la revue Nature Genetics. Dans cet article, les auteurs ont étudié l'impact des TE sur l'insertion/création de sites de fixation des facteurs de transcription dans les génomes humains et de souris. Les travaux de Kunarso et al. ont révélé que les TE ont contribué à l'insertion de près de 25% des TFBS dans les génomes d'humain et souris pour les lignées cellulaires embryonnaires (mESC chez la souris, hESC chez l'Homme). Les auteurs ont utilisé des techniques de ChIP-seq pour identifier les TFBS liés à trois TF (POU5F1, NANOG, CTCF) au sein de séquences de TE. Les résultats de ces travaux ont établi que l'insertion des TFBS dans les séquences de TE n'est pas due au hasard, elle est d'ailleurs parfois spécifique aux familles de TE et/ou aux lignées cellulaires.

En 2013 Jacques PE et al. publie un article sur ce sujet [130]. D'après les résultats de ces travaux, les TE auraient contribué à près de la moitié des éléments actifs du génome humain. Les chercheurs ont révélé que 44% des régions d'ADN ouvertes se trouvaient dans des TE et que ce pourcentage atteignait 63% pour les régions spécifiques aux primates. Ils ont également décelé que certaines sous-familles de rétrovirus endogènes (ERV) ont contribué à plus de régions accessibles que ce que l'on pouvait attendre par hasard, avec jusqu'à 80% de leurs instances dans l'ADN ouvert. Les auteurs ont ainsi pu caractériser 2 150 paires de sous-famille de TE/TF qui étaient liées in vivo ou enrichies de motifs de liaison spécifiques, et ont observé que les TE contribuant à l'ADN ouvert avaient des niveaux plus élevés de conservation de séquence. D'après les auteurs, des milliers de séquences dérivées d'ERV étaient activées de manière spécifique à un type de cellule, en particulier dans les cellules embryonnaires et cancéreuses, influant ainsi sur l'expression des gènes voisins à ces TE.

L'article "Transposable elements as a potent source of diverse *cis*-regulatory sequences in mammalian genome" de Vasavi Sundaram, publié en 2020 [284], décrit la contribution des TE à la régulation chez les eucaryotes. Les TE peuvent jouer un rôle important en fournissant des séquences *cis*-régulatrices pour les TF. Des travaux du même auteur réalisés à plus grande échelle indiquent que jusqu'à 40% des sites de fixation des TF sont dérivés des TE [283]. Les TE auraient contribué à la modification des réseaux de régulation existants et introduit de nouveaux éléments régulateurs spécifiques à une espèce ou à un type cellulaire. Les mécanismes tels que la modification épigénétique et les mutations peuvent également influencer la capacité des TF à se fixer sur les TE. Ces découvertes récentes ont renforcé l'idée que les TE jouent un rôle important dans la diversité de la régulation de l'expression.

Les travaux de 2021 publiés par Jiayue-Clara Jiang et al. révèlent que les éléments transposables LINE-1 serviraient de sites de fixation pour les TF dans les cellules du cancer du sein [134]. Les résultats de leur analyse différentielle entre tissu sain et tissu cancéreux suggèrent que l'exaptation des L1 pour la fixation des TF peut jouer un rôle important dans le développement et la progression du cancer du sein.

La revue de C.Hermant, publié en 2021 [121], explore la manière dont les TF interagissent avec les TE dans les génomes de mammifères. Dans cet article, les auteurs passent en revue les TF connus pour être impliqués dans la reconnaissance de séquence spécifique et l'activation transcriptionnelle de familles ou sous-familles spécifiques de TE. Ils abordent le sujet de la diversité des éléments régulateurs de TE dans les génomes mammifères et mettent en évidence l'importance de la mobilité des TE dans la dispersion des TFBS au cours de l'évolution.

3.2. Objectifs des travaux

Dans le chapitre précédent nous avons vu que les TE sont très probablement impliqués dans la régulation. Ce mécanisme se ferait entre autres par l'insertion de TFBS par les TE au cours de l'évolution. Ces différents travaux nous ont confortés dans l'existence d'un lien entre TE et TF.
En visualisant les régions régulatrices identifiées par ReMap nous avons observé la présence de plusieurs régions régulatrices dans les séquences de TE (Figure 3.1). Nous nous sommes donc demandé s'il était possible d'identifier, à la lumière de nos données unique (ReMap) l'étendu de l'impact des éléments transposables sur les réseaux de régulation.

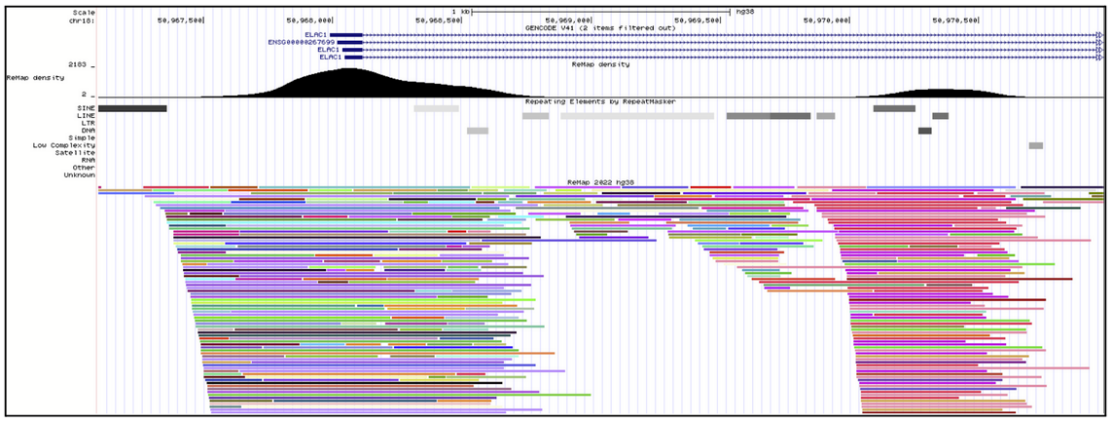


FIGURE 3.1. – ***Capture d'écran des données de ReMap chez l'humain chevauchant des TE.*** *Track 1 : Genes annotation GENCODE; Track 2 : ReMap Density Track; Track 3 : Elements transposables annotation Repeat Masker; Track 4 : catalogue ReMap.*

Les articles précédemment référencés dans la partie 3.1 mettent en évidence des travaux qui se sont concentrées sur de petits ensembles de données génomiques, avec un maximum de 87 TF [130]. Nous avons donc décidé d'étudier la présence de TFBS dans les séquences d'éléments transposables dans les génomes de l'homme et de la souris à l'aide du catalogue ReMap. Cependant nous décrirons uniquement les résultats chez l'Homme, car par manque de temps nous n'avons pas eu le temps d'analyser et comparer les résultats chez la souris.

Notre catalogue étant composées de 1210 TR chez l'homme dans 737 lignées cellulaires, cette analyse permettrait donc d'obtenir une vue plus globale de cet association TE-TFBS. Dans les parties suivantes, nous allons explorer différentes questions liées à l'interaction entre les TE et les TF. Pour ce faire, nous décrirons les étapes méthodologiques qui ont été nécessaires pour mener à bien cette analyse et les résultats qui en découlent.

Tout d'abord, nous nous demanderons si le chevauchement entre les séquences de TE et les sites de fixation des TF est dû au hasard. Pour répondre à cette question, nous avons utilisé un test d'enrichissement des TFBS dans les séquences de TE, en utilisant l'outil LOLA.

Après avoir identifié une association significative entre les TE et les TF, nous cherchons à caractériser le mécanisme responsable de cet enrichissement. En se basant sur la bibliographie existante, nous avons décidé de rechercher les motifs des TF dans les séquences de TE en utilisant l'outil FIMO [106]. Le but de cette analyse étant de déterminer si la présence de motifs de fixation dans les séquences de TE auraient introduit de nouveaux TFBS dans le génome au cours de l'évolution.

Par la suite, nous avons regroupé les TE et les TF sous formes de familles afin d'identifier si l'association des TE et TF avait une spécificité en termes de groupes de TF, et/ou de famille de TE.

Ensuite, nous voulions observer si les motifs de fixation des TF ont été conservés dans les séquences de TE qui contiennent des TFBS. Pour cela, nous avons extrait les séquences de TE associés aux TF, puis nous les avons alignées à l'aide de MUSCLE [76] afin de visualiser la position des motifs dans les alignements.

Enfin, nous avons tenté de déterminer si les TE avaient inséré des TFBS dans le génome au cours de l'évolution. Pour répondre à cette question, nous avons d'abord estimé l'âge des TE, puis nous avons analysé la présence des TFBS dans les TE en prenant en compte cet indicateur.

Notez que ces travaux ne sont pas encore publiés. Par conséquent, ils seront présentés dans un format d'article classique, comprenant les sections Matériels et Méthodes, Résultats et Conclusion.

3.3. Matériels et méthodes

3.3.1. Préparation des données

3.3.1.1. ReMap

Les données ChIP-seq ont été collectées sur le site de ReMap 2022 [1] sous forme de fichier BED pour chaque TF. ReMap propose plusieurs formats de données, les BED contenant toutes les données ainsi que les pics non-redondants. Dans le cadre de nos travaux, nous avons décidé d'utiliser les fichiers contenant les pics non redondants de chaque TF disponible.

3.3.1.2. Séquence de TE provenant de Repeatmasker

Les coordonnées des éléments transposables ont été collectées à l'aide de l'outil Table Browser du navigateur de génomes UCSC [2] [168]. Pour cela, j'ai téléchargé les fichiers de RepeatMasker [213] pour l'Homme (hg38) et la souris (mm10). Ces données ont été collectées sous forme de fichier TSV (Figure 3.2) puis converties en format BED pour les analyses ultérieures. J'ai également collecté la liste des TE, leur famille et leur superfamille, afin de les utiliser pour la classification des TE.

#bin	swScore	milliDiv	milliDel	milliIns	genoName	genoStart	genoEnd	genoLeft		strand	repName	repClass	repFamily	repStart	repEnd	repLeft	id
0	1892	83	59	14	chr1	67108753	67109046	-181847376	+	L1P5	LINE	L1	5301	5607	-544	1	
1	2582	27	0	23	chr1	8388315	8388615	-240567804	-	AluY	SINE	Alu	-15	296	1	1	
1	4085	171	77	36	chr1	25165803	25166380	-223790042	+	L1MB5	LINE	L1	5567	6174	0	4	
1	2285	91	0	13	chr1	33554185	33554483	-215401939	-	AluSc	SINE	Alu	-6	303	10	6	
1	2451	64	3	26	chr1	41942894	41943205	-207013217	-	AluY	SINE	Alu	-7	304	1	8	
1	1587	272	100	49	chr1	50331336	50332274	-198624148	+	HAL1	LINE	L1	773	1763	-744	9	
1	1393	280	82	51	chr1	58719764	58720546	-190235876	+	L2a	LINE	L2	2582	3418	-8	1	
2	5372	165	14	27	chr1	75496057	75497775	-173458647	+	L1MA9	LINE	L1	5168	6868	-30	1	
2	536	349	146	56	chr1	92274205	92275925	-156680497	+	L2	LINE	L2	406	2306	-1113	1	
2	25118	52	2	0	chr1	100662981	100669120	-148287302	-	L1PA4	LINE	L1	-2	6153	7	1	
3	2212	97	0	0	chr1	150994812	150995102	-97961320	+	AluSg	SINE	Alu	2	291	-19	2	

FIGURE 3.2. – *Huits premières lignes du fichier TSV de RepeatMasker.*

Certaines familles de TE ont été éliminées lors de l'analyse : Low Complexity, simple repeats, Unknown. Ce choix a été fait car il ne paraissait pas nécessaire de les prendre en compte. Nous avons choisi de nous focaliser sur les éléments transposables et pas sur les short tandem repeats (STR) car ceux-ci constituent un vaste domaine de recherches [92, 178, 124]. De plus, les TE de faible complexité sont généralement des séquences de nucléotides qui se répètent de manière régulière mais ne contiennent pas de séquences codantes [114] , tandis que les répétitions simples sont des séquences de nucléotides qui se répètent plusieurs fois de suite dans une région donnée du génome, comme les séquences polyA ou polyT [104].

1. https://remap.univ-amu.fr/
2. https://genome.ucsc.edu/cgi-bin/hgTables

3.3.1.3. ENCODE DNase-seq

Lors de l'étape d'enrichissement (section 3.4.2), nous sommes tenus d'utiliser un univers génomique pour identifier les régions génomiques "actives". Pour ce faire, nous avons choisi d'utiliser les DNases d'ENCODE en raison de leur capacité à représenter les régions du génome avec une chromatine ouverte, ce qui indique une "activité génomique".[200].

L'atlas DNase-seq d'ENCODE couvre 27 types tissulaires humains, ainsi que différents types de traitement et de perturbation des cellules. Ils permettent de cartographier les régions d'ADN accessibles dans les différents types de cellules et de comprendre comment ces régions varient en fonction des conditions expérimentales. Cet atlas est un outil précieux pour les chercheurs qui étudient la régulation, les mécanismes de certaines maladies et les effets des perturbations environnementales sur les cellules. Les données sont accessibles via le portail d'ENCODE [1] et sont largement utilisées dans la communauté de recherche pour des analyses bioinformatiques et en génétique fonctionnelle.

3.3.1.4. JASPAR

Afin d'identifier les motifs de fixation dans les séquences de TE, nous avons obtenu les motifs de TF de la base de données JASPAR [46]. Ces motifs ont été téléchargés au format MEME [14] et peuvent être obtenus sur le site internet de JASPAR [2]. Nous avons sélectionné les PFM non redondants des vertébrés pour réaliser notre analyse.

Pour simplifier le processus d'analyse, j'ai utilisé un script Python pour renommer les fichiers téléchargés avec le nom du facteur de transcription, sous le format "TF.meme". Cela a permis d'identifier plus facilement les motifs dans les données et de les associer aux facteurs de transcription correspondants pour mieux comprendre les interactions entre les facteurs de transcription et les séquences de TE.

JASPAR (JASPAR CORE vertebrates et metazoan) est une base de données de motifs de fixation de TF. Ces motifs ont été compilés à partir de différentes sources expérimentales, parmis lesquelles on retrouve les données ChIP-seq du catalogue ReMap. Les motifs sont mis à jour régulièrement pour inclure les dernières données expérimentales et sont classés selon la famille de facteurs de transcription à laquelle ils appartiennent.

1. https://www.encodeproject.org/
2. https://jaspar.genereg.net/downloads/

3.3.2. Enrichissement

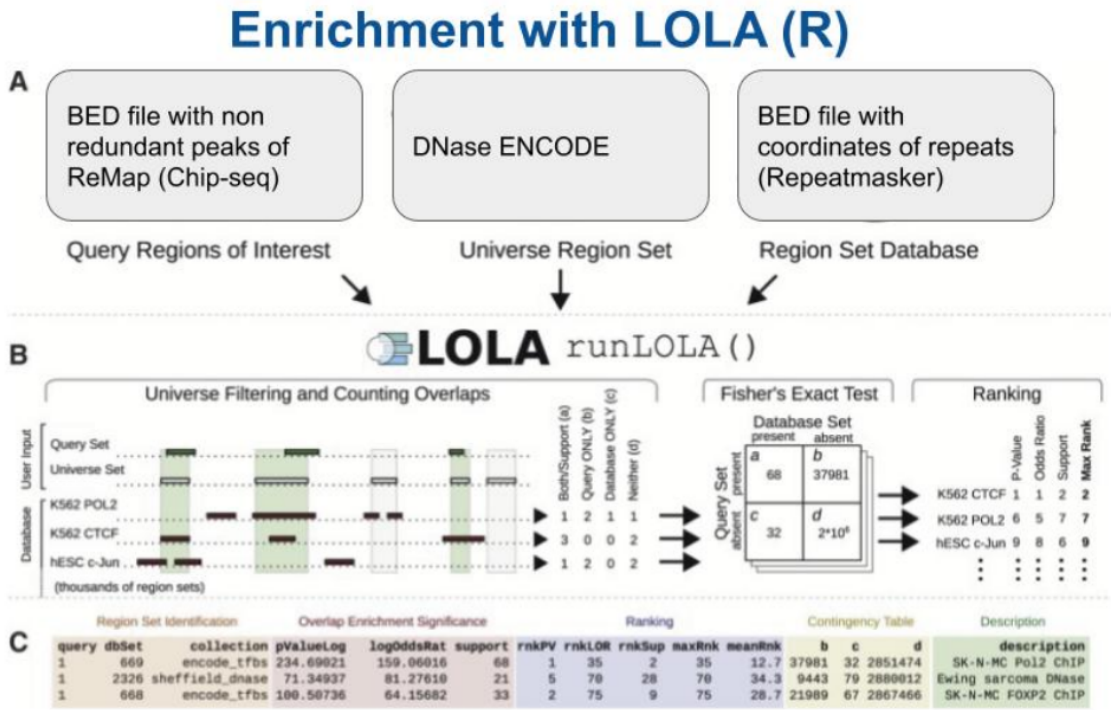
L'interprétation de différents types de données biologiques peut être réalisée en comparant les données à des bases de référence et en recherchant des modèles d'enrichissement pertinents. Poussée par des projets tels que ENCODE [274] et FANTOM [5], la communauté des chercheurs a établi des catalogues exhaustifs d'éléments régulateurs et d'autres caractéristiques génomiques dans divers types cellulaires, permettant ainsi une meilleure compréhension des mécanismes génétiques. Dans le cadre de ma thèse, un de nos objectifs consiste à évaluer l'enrichissement de régions régulatrices dans les séquences de TE. Ce type d'analyse est possible grâce à l'outil bioinformatique LOLA (Locus Overlap Analysis) [263]. LOLA est un outil permettant l'analyse d'enrichissement basée sur les régions génomiques. Ce package R/Bioconductor est conçu pour faciliter l'analyse rapide des ensembles de régions génomiques pour toutes les espèces disposant d'un génome de référence annoté. LOLA complète les outils existants pour l'analyse d'enrichissement génique [152], les outils de prédiction de fonction de régions génomiques tels que GREAT [197] et les outils de détection de régions régulatrices via l'enrichissement en pics ChIP-Seq [13]. La principale force de LOLA est son intégration avec R et Bioconductor ainsi que son interface en ligne de commande pour le traitement automatisé des données. De plus, il est compatible avec les pipelines haut débit et les scripts interactifs en R (Figue 3.3), et offre un temps d'exécution rapide même pour les grandes listes de régions et les bases de données de référence.

```
1  library("LOLA")
2  library(GenomicRanges)
3  library(rtracklayer)
4  regionDB = loadRegionDB("/shared/projects/remap/remap-repeats/LOLA/")
5  #!/usr/bin/env Rscript
6  args = commandArgs(trailingOnly=TRUE)
7  userSets =  import(args[1])
8  userUniverse = import("/shared/projects/remap/remap-repeats/ENCFF503GCK.bed")
9
10 locResults = runLOLA(userSets, userUniverse, regionDB, cores=1)
11 ## ----------------------------------------------------------------------
12 colnames(locResults)
13 head(locResults)
14
15 ## ----------------------------------------------------------------------
16 locResults[order(support, decreasing=TRUE),]
17
18 ## ----------------------------------------------------------------------
19 locResults[order(maxRnk, decreasing=TRUE),]
20
21 ## ----Write results----------------------------------------------------
22 writeCombinedEnrichment(locResults, outFolder= args[2])
```

FIGURE 3.3. – *Script R pour l'analyse d'enrichissement avec LOLA.*

Chaque analyse LOLA est basée sur trois composants (Figure 3.4) :

- L'ensemble de requêtes, une ou plusieurs listes de régions génomiques à tester pour l'enrichissement. Dans notre analyse, cela correspond aux pics ChIP-seq de ReMap.
- Un ensemble de régions univers, qui correspond à l'ensemble de régions qui auraient pu potentiellement être inclus dans l'ensemble de requêtes. Nous avons décidé d'utiliser les régions DNase d'ENCODE.
- Une base de données de référence d'ensembles de régions génomiques qui doivent être testés pour un chevauchement avec l'ensemble de requêtes. Ces données correspondent aux séquences des éléments transposables provenant de RepeatMasker.

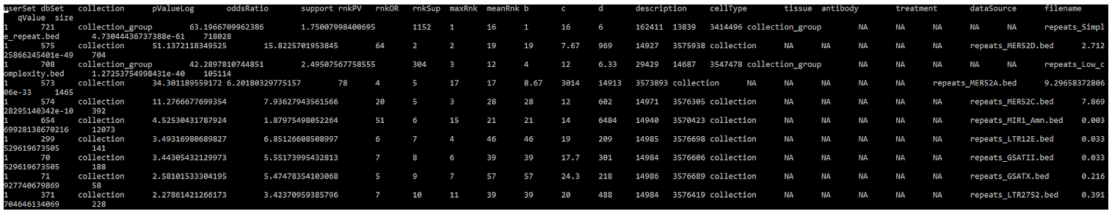


query	dbSet	collection	pValueLog	logOddsRat	support	rnkPV	rnkLOR	rnkSup	maxRnk	meanRnk	b	c	d	description
1	669	encode_tfbs	234.69021	159.06016	68	1	35	2	35	12.7	37981	32	2851474	SK-N-MC Pol2 ChIP
1	2326	sheffield_dnase	71.34937	81.27610	21	5	70	28	70	34.3	9443	79	2880012	Ewing sarcoma DNase
1	668	encode_tfbs	100.50736	64.15682	33	2	75	9	75	28.7	21989	67	2867466	SK-N-MC FOXP2 ChIP

FIGURE 3.4. – ***Workflow de l'outil d'enrichissement LOLA et ses résultats.****(A) L'ensemble de requêtes, les régions univers et la base de données de référence sont chargés dans R. (B) LOLA identifie le chevauchement, calcule l'enrichissement et classe les résultats. (C) Exemple de résultats d'enrichissement classés de LOLA obtenus avec runLOLA().*

LOLA comprend une base de données de référence de base assemblée à partir de données publiques, y compris, par exemple, la base de données CODEX [248] et l'annotation inter-tissulaire de données de DNase [264]. Nous avons pu créer des ensembles de régions personnalisées et construire une base de données de référence personnalisée, il suffit de créer une collection de fichier BED avec les coordonnées génomiques des éléments transposables. Ces fichiers sont collectés dans un dossier, annoté avec des noms descriptifs.

LOLA identifie toutes les régions génomiques des pics ChIP-seq de ReMap qui chevauchent chaque ensemble de séquences d'éléments transposables dans la base de données de référence. Nous avons gardé les paramètres par défaut, une seule paire de base partagée est suffisante pour que les régions soient considérées comme se chevauchant. Ensuite, en considérant chaque région comme indépendante, LOLA utilise le test exact de Fisher avec correction du taux de faux positifs pour évaluer l'importance du chevauchement dans chaque comparaison par paires TE/TF. Le score résultant pour chaque ensemble de régions d'un TE donné est ensuite calculé en lui attribuant le rang le plus mauvais (max) parmi trois mesures : la p-value du test de fischer, le rapport de cotes logarithmique et le nombre de régions qui se chevauchent. Les résultats sont renvoyés sous la forme d'un fichier csv (Figure 3.5).



userSet	dbSet	collection	pValueLog	oddsRatio	support	rnkPV	rnkOR	rnkSup	maxRnk	meanRnk	b	c	d	description	cellType	tissue	antibody	treatment	dataSource	filename	qValue	size
1	721	collection_group	63.1966709962386	1.75007998400695	1152	1	16	1	16	6	162411	13839	3414496	collection_group	NA	NA	NA	NA	NA	repeats_Simple_repeat.bed	4.73044436737388e-61	718028
1	575	collection	51.1372118349525	15.8225701953845	64	2	2	19	19	7.67	969	14927	3575938	collection	NA	NA	NA	NA	NA	repeats_MER52D.bed	2.7125866245401e-49	704
1	708	collection_group	42.2897810744851	2.49507567758555	304	3	12	4	12	6.33	29429	14687	3547478	collection_group	NA	NA	NA	NA	NA	repeats_Low_complexity.bed	1.27253754998431e-40	105114
1	573	collection	34.3011895591726.20180329775157	78	4	5	17	17	8.67	3014	14913	3573893	collection	NA	NA	NA	NA	NA	repeats_MER52A.bed	9.29658372806	06e-33	1465
1	574	collection	11.2766677699354	7.93627943561566	20	5	3	28	28	12	602	14971	3576305	collection	NA	NA	NA	NA	NA	repeats_MER52C.bed	7.869	2829514034 2e-10	392
1	654	collection	4.52530431787924	1.87975490852264	51	6	15	21	21	14	6484	14940	3570423	collection	NA	NA	NA	NA	NA	repeats_MIR1_Amn.bed	0.003	699281386702 16	12073
1	299	collection	3.49316980689827	6.85126608508997	6	7	4	46	46	19	209	14985	3576698	collection	NA	NA	NA	NA	NA	repeats_LTR12E.bed	0.033	529619673505	141
1	70	collection	3.44305432129973	5.55173995432813	7	8	6	39	39	17.7	301	14984	3576606	collection	NA	NA	NA	NA	NA	repeats_GSATII.bed	0.033	529619673505	188
1	71	collection	2.58101533304195	5.4747835410306 8	5	9	7	57	57	24.3	218	14986	3576689	collection	NA	NA	NA	NA	NA	repeats_GSATX.bed	0.216	927740679869	58
1	371	collection	2.27861421266173	3.42370959385796	7	10	11	39	39	20	488	14984	3576419	collection	NA	NA	NA	NA	NA	repeats_LTR2752.bed	0.391	704646134069	228

FIGURE 3.5. – ***Screenshot d'un fichier de sortie LOLA du facteur de transcription AATF*** *Ce fichier CSV contient les résultats d'enrichissement du logiciel LOLA pour chacun des TE testé.*

Nous avons développé un script Python pour obtenir la liste des paires TE/TF qui ont montré un score significatif d'enrichissement (p-value < 10-5). Les résultats obtenus ont permis de créer une heatmap où les TE sont représentées en colonnes, les TF en lignes et chaque case correspond au -log(p-value) d'enrichissement (Figure 3.15, section 3.4.2). Ce script nous a permis de mettre en évidence les paires TE/TF associées. LOLA a fait ressortir 15441 paires TE/TF avec un enrichissement significatif (p-val<10÷5). Nous avons ensuite décidé d'évaluer la présence des motifs des TF enrichies dans les séquences des TE. Pour ce faire, nous avons collecté les motifs de la base de données JASPAR pour 693 TF enrichies et utilisé l'outil FIMO pour rechercher ces motifs. Cette analyse sera expliquée en détail dans la section suivante.

3.3.3. Recherche de motifs

Un motif de fixation de TF, représente la séquence spécifique où un TF se lie à l'ADN. Les motifs de séquence sont un élément fondamental pour comprendre les processus évolutifs moléculaires. Par conséquent, l'identification des motifs de fixation de TF dans les séquences de TE enrichies par des pics de ChIP-seq d'un TF est cruciale pour comprendre comment ces phénomènes se produisent à l'échelle génomique. Pour ce faire, nous avons utilisé FIMO (Find Individual Motif Occurrences) [106], qui utilise une méthode statistiquement rigoureuse pour scanner les séquences d'ADN et trouver les occurrences précises des motifs ciblés.Dans le cadre de la suite MEME [15], FIMO peut être utilisé en ligne de commande. FIMO prend en entrée un ou plusieurs motifs de longueur fixe, représentés sous forme de matrices de fréquence spécifiques à la position. Ces motifs sont extraits d'une base de données de motifs, JASPAR au format MEME (voir Figure 3.6).

```
MEME version 4

ALPHABET= ACGT

strands: + -

Background letter frequencies
A 0.25 C 0.25 G 0.25 T 0.25

MOTIF MA0139.1 CTCF
letter-probability matrix: alength= 4 w= 19 nsites= 913 E= 0
 0.095290  0.318729  0.083242  0.502738
 0.182913  0.158817  0.453450  0.204819
 0.307777  0.053669  0.491785  0.146769
 0.061336  0.876232  0.023001  0.039430
 0.008762  0.989047  0.000000  0.002191
 0.814896  0.014239  0.071194  0.099671
 0.043812  0.578313  0.365827  0.012048
 0.117325  0.474781  0.052632  0.355263
 0.933114  0.012061  0.035088  0.019737
 0.005488  0.000000  0.991218  0.003293
 0.365532  0.003293  0.621295  0.009879
 0.059276  0.013172  0.553238  0.374314
 0.013187  0.000000  0.978022  0.008791
 0.061538  0.008791  0.851648  0.078022
 0.114411  0.806381  0.005501  0.073707
 0.409241  0.014301  0.557756  0.018702
 0.090308  0.530837  0.338106  0.040749
 0.128855  0.354626  0.080396  0.436123
 0.442731  0.199339  0.292952  0.064978
URL http://jaspar.genereg.net/matrix/MA0139.1
```

FIGURE 3.6. – *Fichier au format MEME du motif de CTCF provenant de la base de données JASPAR.*

Le programme calcule un score de rapport log-vraisemblance pour chaque motif par rapport à chaque position de séquence et convertit ces scores en p-value à l'aide d'une programmation dynamique, en supposant un modèle d'ordre zéro (hypothèse h0) dans lequel les séquences sont générées au hasard en prenant en compte les fréquences des bases données dans le fichier MEME du motif (Background letter frequencies, Figure 3.6). FIMO produit en sortie une liste classée d'occurrences de motifs, chacune avec un score de rapport de vraisemblance logarithmique associé, une p-value et une q-value. Cette liste est représentée sous forme de rapport au format CSV délimités par des tabulations.

motif_id	motif_alt_id	sequence_name	start	stop	strand	score	p-value	q-value	matched_sequence
MA1155.1	ZSCAN4	AluJb::chr19:18416316-18416477(-)	11	25	-	14.3966	1.95e-06	0.175	GGCACACGCTGCCAC
MA1155.1	ZSCAN4	AluJb::chr17:62877895-62878074(+)	17	31	+	13.9655	2.27e-06	0.175	ggcacacacttgcag
MA1155.1	ZSCAN4	AluJb::chr1:225276615-225276944(+)	131	145	+	12.7759	3.47e-06	0.175	tgcacacacacaaaa
MA1155.1	ZSCAN4	AluJb::chr7:127232249-127232549(-)	14	28	+	12.6724	3.58e-06	0.175	tgcacacacttgtaa
MA1155.1	ZSCAN4	AluJb::chr1:155736418-155736716(-)	238	252	+	12.4828	3.84e-06	0.175	tacacatactgtcac
MA1155.1	ZSCAN4	AluJb::chr17:143333-143644(-) 128	142	+	11.2241	5.82e-06	0.175	cacacacaatgacaa
MA1155.1	ZSCAN4	AluJb::chr10:119834534-119834888(+)	117	131	+	10.7414	6.81e-06	0.175	tgcacacacacaaac
MA1155.1	ZSCAN4	AluJb::chr7:43839923-43840089(+)	17	31	+	10.1552	8.22e-06	0.175	ggcacacacttatag
MA1155.1	ZSCAN4	AluJb::chr3:196016681-196016947(+)	109	123	+	9.89655	8.88e-06	0.175	tacacacacacaaaa
MA1155.1	ZSCAN4	AluJb::chr1:229249995-229250296(-)	138	152	+	9.7069	9.41e-06	0.175	cacacacacgcaaat
MA1155.1	ZSCAN4	AluJb::chr19:15411375-15411649(+)	110	124	-	9.56897	9.8e-06	0.175	GGCACATGCTGCCAC
MA1155.1	ZSCAN4	AluJb::chr20:56741630-56741809(-)	58	72	-	9.5	1e-05	0.175	TGCACACCATTACAC
MA1155.1	ZSCAN4	AluJb::chr6:20693875-20694042(-)	14	28	-	8.93103	1.19e-05	0.175	GGCACACACTAACAC

FIGURE 3.7. – *Fichier output de fimo du couple ZSCAN4/AluJb.*

Pour les paires TE/TF ayant présenté un enrichissement significatif, nous avons lancé une recherche des motifs correspondants en utilisant l'outil FIMO sur les séquences de TE coïncidant avec les pics de ChIP-seq ReMap du TF considéré. Cette approche nous a permis d'identifier les sites de fixation protéique spécifiques associés à chaque couple TE/TF significatif. Cette analyse a été réalisée à l'aide d'un workflow schématisé sur la Figure 3.8.

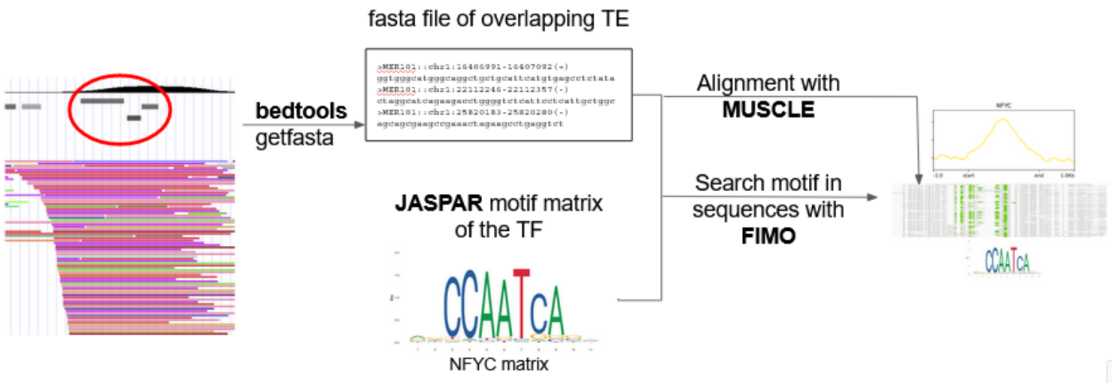


FIGURE 3.8. – *Étapes pour faire la recherche de motifs dans les séquences des TE enrichies en TFBS. Ici un exemple avec NFYC et les TE MER101.*

3.3.4. Calcul pourcentage de motifs

Afin de représenter les résultats obtenus lors de cette analyse nous avons créé une métrique basée sur le pourcentage de motifs identifiés dans les séquences de TE qui chevauchent les pics ChIP-seq de TF. Le pourcentage de motifs a été créé pour résoudre le problème suivant : nous ne pouvions pas observer les alignements de tous les éléments transposables (plus de 7000 alignements) et ainsi constater la présence des motifs de TF dans les séquences. Il était donc nécessaire de trouver un moyen de quantifier la présence des motifs dans les séquences de TE. Cette métrique nous permet de remédier à ce problème en fournissant une mesure permettant d'évaluer comment certains TE présentent plus ou moins de motifs de fixation pour les TF, témoignant ainsi de l'insertion des TFBS dans le génome.

Nous avons extrait les séquences de TE enrichies en pics ChIP-seq de TF de manière significative. Nous effectuons ensuite une recherche du motif du TF enrichie dans ces séquences de TE. Ensuite pour chaque couple TE/TF je calcule le pourcentage de ces motifs identifiées dans ces séquences de TE :

Pourcentages de motifs = $\frac{Nombredeséquencesavecunmotif}{NombredeTEenrichies}$
Les motifs ont été compté en suivant la méthode de la Figure 3.9.

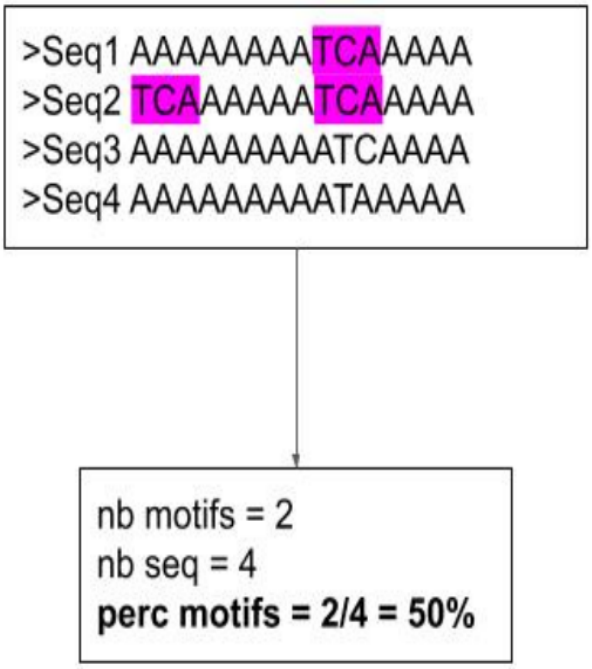


FIGURE 3.9. – ***Méthode pour le calcul du pourcentage de motifs.*** *Pour chaque séquence nous cherchons à savoir si le motif du TF est présent. Ensuite nous calculons la proportion de séquence avec motifs.*

3.3.5. Alignements

Les alignements réalisés dans le cadre de nos analyses ont été réalisés avec l'outil d'alignement multiple MUSCLE [76]. Le programme MUSCLE [76] est une méthode utilisée pour aligner des séquences d'acides aminés ou d'acides nucléiques. Il est considéré comme l'un des meilleurs outils pour l'alignement de séquences multiples car il est rapide et précis.

MUSCLE utilise une approche de l'optimisation par les scores de log-expectation pour aligner les séquences, ce qui permet d'obtenir des alignements plus précis que les méthodes basées sur des scores de similarité. Il utilise également une technique d'optimisation par étape pour améliorer la précision de l'alignement. Nous avons utilisé la version 5 de l'algorithme de MUSCLE [1], cette version de l'outil permet d'aligner des dizaines de milliers de séquences très rapidement. Apres des tests, Nous avons choisi cet algorithme au lieu de T-Coffee [215] et Clustal O [265] pour sa rapidité sur de très grands jeux de données.

Notre travail débute par la collecte des séquences de TE associées à un TF donné parmi les paires TE/TF ayant un enrichissement significatif. Pour cela, nous avons utilisé la fonction "intersect" de la suite d'outils bedtools [235] pour filtrer les séquences présentant un chevauchement avec les pics ChIP-seq du TF associé. Les séquences sont ensuite obtenues au format fasta via la fonction "getfasta" de bedtools.

Une fois les séquences collectées, nous les avons alignées avec l'outil d'alignement multiple MUSCLE [76]. Ensuite, les résultats de l'alignement et de FIMO ont été utilisés pour visualiser la position des motifs dans les séquences de TE (Figure 3.10). La visualisation de ces résultats a été rendue possible grâce au logiciel Mview [39], qui permet de visualiser des alignements et de colorer certaines séquences spécifiques. Dans notre cas, nous avons mis en évidence les séquences des motifs identifiées avec FIMO en les surlignant.

FIGURE 3.10. – ***Exemple d'alignement des séquences de TE.*** *Alignement des séquences de MER101 sur lesquelles le TF NFYCse fixent. L'outil Mview a permis de colorer en vert les séquences du motif de fixation de NFYC retrouvées par FIMO.*

1. https://www.drive5.com/muscle/

3.3.6. Workflow

Les analyses de ce chapitre ont été effectuées pour 1210 facteurs de transcription et 692 éléments transposables, il était donc nécessaire d'automatiser les étapes de cette analyse pour en faciliter la réalisation. La première étape de cette analyse consiste à lancer un Snakefile (fichier d'éxection de snakemake) qui exécute le script R de l'analyse d'enrichissement de LOLA. Ce script prend en entrée les fichiers ReMap pour chaque TF et les fichiers BED de chaque TE. En sortie, nous obtenons un dossier pour chaque TF contenant les résultats d'enrichissement pour chaque TE. Ces résultats sont ensuite traités par un script Python qui collecte la liste des couples TE/TF ayant un enrichissement significatif. Cette liste est ensuite utilisée en entrée d'un deuxième workflow qui permet d'obtenir les séquences des TE qui chevauchent les sites de fixation des facteurs de transcription, de les aligner et de rechercher les motifs des TFBS dans ces séquences. Les figures présentées dans la suite du manuscrit ont été créées avec le logiciel R, et les fichiers d'entrées pour ces figures ont été générées via un script Python.

3.3.7. Calcul de l'âge des TE

Le calcul de l'âge des repeats est basé sur les articles publiés de G.Bourque [35]. La formule utilisée pour estimer l'âge des TE est la suivante :
$\text{Âge} = \frac{Divergence}{Ratiodesubstitution}$ Le ratio de substitution chez l'Homme a été estimé à $2.2x10^{-}9$, tandis que pour la souris il est de $4.5x10^{-}9$ [2, 1]. Afin de calculer la divergence j'ai utilisé la méthode de Jukes-Cantor [79]. Le calcul de la divergence est donc réalisé à l'aide de la formule suivante :

$$t = \frac{-3}{4} ln(1 - \frac{4}{3}p)$$

La distance p entre la séquence ancêtre et la séquence actuelle du TE a été obtenue à partir de RepeatMasker et la divergence (t) a été calculée pour chaque séquence de TE. Une moyenne des divergences a ensuite été utilisée pour déterminer l'âge global du TE. Ce calcul a été effectué via un script Python pour l'Homme et la souris.
Afin de faciliter l'analyse nous avons divisé l'âge des TE en périodes différentes basé sur l'article publié par Pace et Feschotte [220] (Tableau 3.1).

TABLEAU 3.1. – *Période en millions d'années afin de classifier l'âge des TE basé sur [220].*

Nom de la période	Période en millions d'années (ma)
Post anthropoid	de 0 à 40ma en arrière
Anthropoid specific	de 40ma à 63ma en arrière
Primate specific	de 63ma à 80ma en arrière
Eutherian specific	de 80ma à 150 ma en arrière
Pré-Eutherian	à partir de 150ma

3.4. Résultats

3.4.1. Analyse préliminaire

La première analyse que nous avons réalisé consistait à déterminer l'étendue des pics ChIP-seq chevauchant les TE dans le génome humain. Pour cela, nous avons décidé de faire un comptage en paire de base de la séquence chevauchante entre le pic et le TE (Figure 3.12). Cette analyse s'inspire d'une figure de l'article de H. Nishihara [212], qui montre le pourcentage de TE qui se superposent aux événements de fixation des TF.

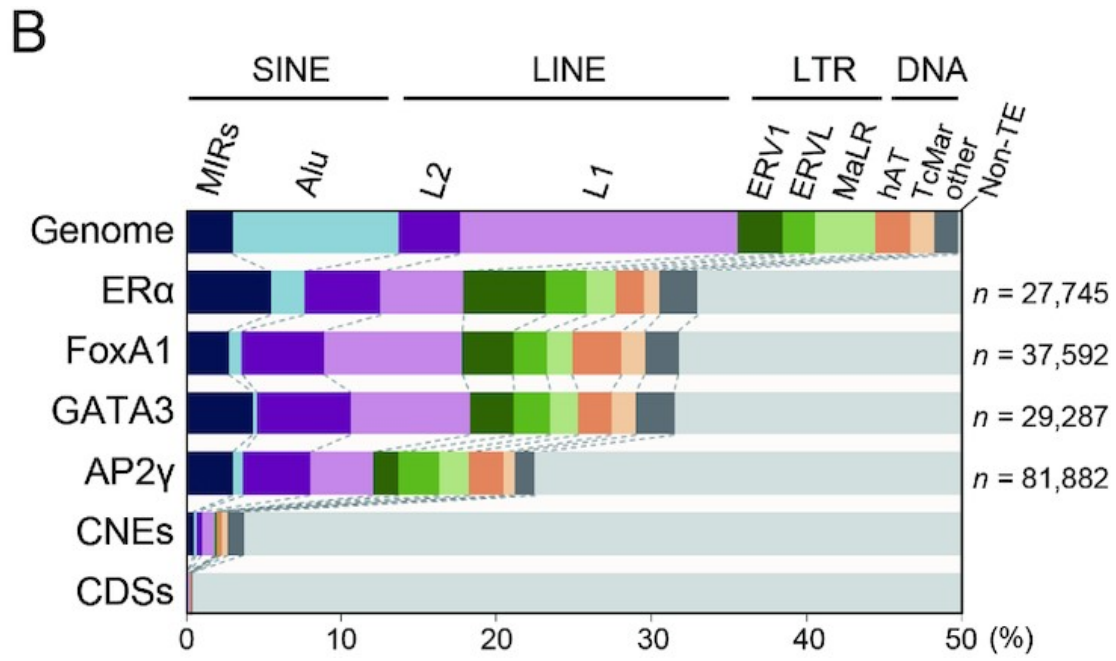


FIGURE 3.11. – ***Pourcentage des TE parmi le nombre d'événements de liaison (n) pour ERα, FoxA1, GATA3 et AP2γ.*** *Alignement des séquences de MER101 sur lesquelles le TF NFYCse fixent. Les proportions des TE dans le génome humain (hg19, à l'exclusion du chromosome Y) (Genome), les éléments non codants conservés (CNE) et les séquences codantes en protéines (CDS) sont présentées*[11].

11. Nishihara, H. (2019). Retrotransposons spread potential cis-regulatory elements during mammary gland evolution.

Nous avons d'abord représenté sous forme de barplot le pourcentage en pb du chevauchement TE/TF par rapport au nombre de pb des TE du génome (Figure 3.12). Cependant la figure montre une très grande différence entre le pourcentage de TE (100%) et le TF avec le pourcentage le plus élevé (BRD4 avec 14%). La proportion de TE sur le génome est également observable dans l'annexe A.

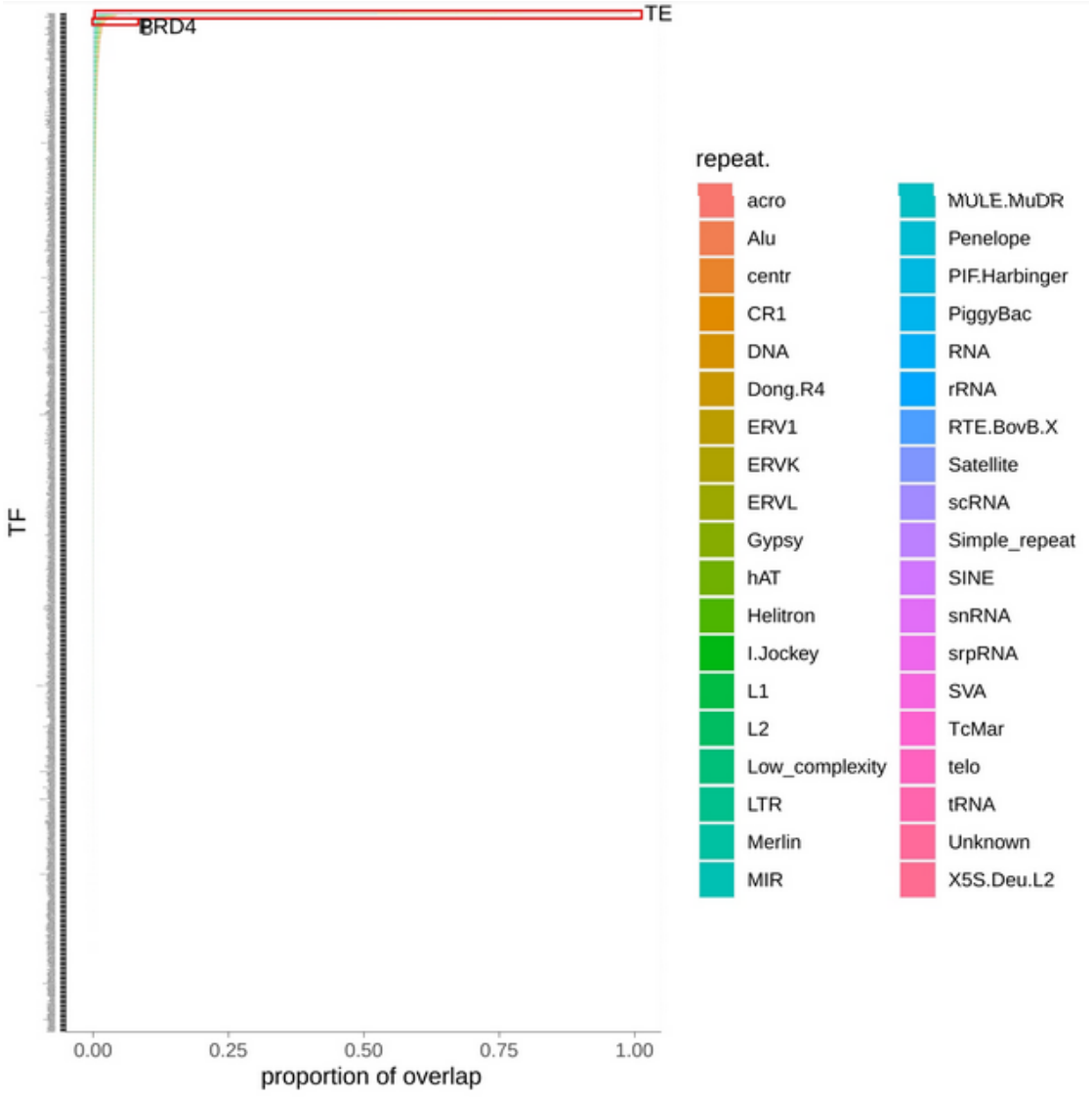


FIGURE 3.12. – ***Barplot représentant la proportion de pics TF chevauchant des TE par rapport aux pb de TE.*** *Barplot horizontal représentant pour chaque TF la proportion de chevauchement entre les pics ChIP-seq ReMap et les séquences de TE en paire de base par rapport au nombre de pb total des TE. La première ligne correspond à l'ensemble des TE dans le génome humain.*

Nous avons décidé de modifier la méthode utilisé pour construire ce graphique afin de normaliser la proportion en fonction des pics totale du TF. Pour calculer cette nouvelle proportion, nous utilisons les pics non redondants (NR) de ReMap. Nous avons décidé de nous baser sur la longueur totale des pics en pb, c'est-à-dire l'espace total en pb des pics dans le génome (Figure 3.13). Ceci permet d'éviter des biais inhérents aux nombres d'expériences de ChIP-seq, qui est variable selon le TF étudié. En effet pour certains TF la distribution des TE peut être affectée par la quantité de données disponible. En utilisant cette métrique, nous avons normalisé le chevauchement des TF.

Sur la Figuree 3.13 nous observons que pour certains TF ont plus de 50% de leurs pics qui chevauchent des séquences de TE. Ces données ne peuvent pour l'instant pas être interprétées car il est nécessaire d'étudier statistiquement cet enrichissement mais cette première analyse nous a confortés dans l'idée d'un lien entre TE et fixation des TF.

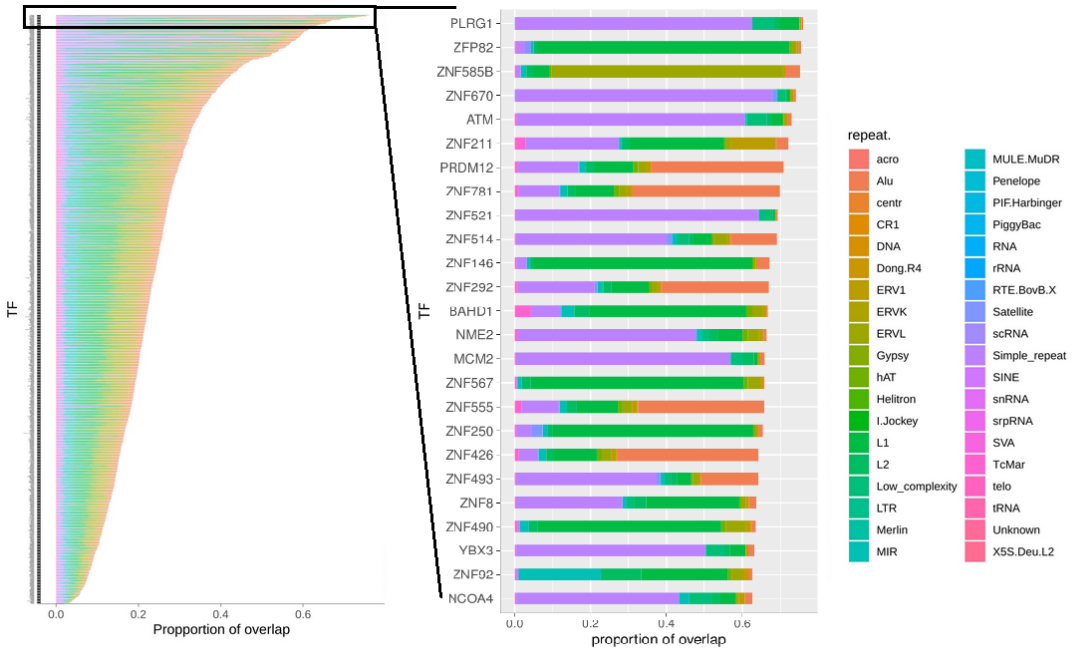


FIGURE 3.13. – ***Proportion de pics NR de ReMap qui chevauchent une séquence de TE en pb.*** *A gauche, les résultats pour les 1210 TF disponible dans ReMap. A droite, les résultats pour les top 25 des TF avec une proportion la plus élevé. Les résultats montrent la proportion pour chaque famille de TE.*

3.4.2. Enrichissement des TFBS sur les TE

Après avoir observé la distribution des TE au sein des TF, nous avons voulu évaluer l'impact réel de l'insertion des TE sur la régulation, plus particulièrement la fixation des TF sur les TE. En nous basant sur les travaux de Bourque et al. [35] nous avons réalisé une analyse d'enrichissement des pics ChIP-seq de TF. Cet analyse consistait à déterminer si les séquences des TE sont de manière significative enrichies en TFBS identifiées par ChIP-seq. Nous avons effectué cette analyse en réalisant les étapes suivantes :

- Télécharger le fichier de RepeatMasker disponible sur UCSC qui contient les séquences des éléments transposables. Ce fichier a ensuite été formaté pour sélectionner uniquement les colonnes utiles (coordonnées des séquences de TE, nom de la repeat, nom de la famille, etc.) et créer des fichiers BED pour chaque repeat. Nous avons donc obtenu 692 fichiers BED.

- Télécharger les fichiers de ReMap qui contiennent les pics non redondants de chaque TF identifié par ChIP-seq. Nous avons obtenu 1210 fichiers BED.

- Effectuer une analyse d'enrichissement avec l'outil LOLA qui a été intégré dans un script développé sous R. A l'aide de ce script et un pipeline Snakemake nous avons calculé l'enrichissement pour chaque paire TE/TF. Les résultats ont été obtenus sous forme de fichier pour chaque TF contenant la liste des éléments transposables testés et les métriques associées (le rank, la p-value d'enrichissement...).

- Ensuite, nous avons créé un fichier au format matrice à partir des fichiers de sorties de LOLA contenant les p-value d'enrichissement. Ce fichier stocke les p-values d'enrichissement pour chaque paires TE/TF, les lignes correspondant aux TF et les colonnes correspondant aux TE.

- A partir de cette matrice nous avons créé une heatmap avec l'outil R. Cette heatmap permet de visualiser les résultats de l'analyse d'enrichissement.

Au total nous avons identifié 15,441 paires TE/TF associés avec un enrichissement significatif (p-val $< 10^{-}5$), avec 693 TF et 622 TE. Dans l'ensemble, 70 TE n'ont aucun enrichissement en TF, et 479 TF ne sont enrichies dans aucun TE. Malheureusement le reste des analyses n'as pas pu aboutir pour 38 TF car ils ne contiennent pas de motifs dans la base de données de JASPAR.

La représentation des résultats de l'analyse se fait sous forme de heatmap afin de visualiser les enrichissements des TF sur les TE. La Figure 3.14 est inspirés d'une figure de l'article de Bourque al. [35]. Nous avons représenté les enrichissements en fonction des familles de TE. Les colonnes représentent les familles de TE et les lignes représentent les TF. La valeur utilisée pour construire la heatmap est le -log de la p-value d'enrichissement. Dans leur article les auteurs mettent en valeur un enrichissement du TF ESR1 dans les séquences de la famille des MIR. Cet enrichissement est également identifié dans nos analyses. Dans la Figure 3.14, on observe une association préférentielle des TF à certaines familles de TE tel que les MIR, L1, L2 et Alu.

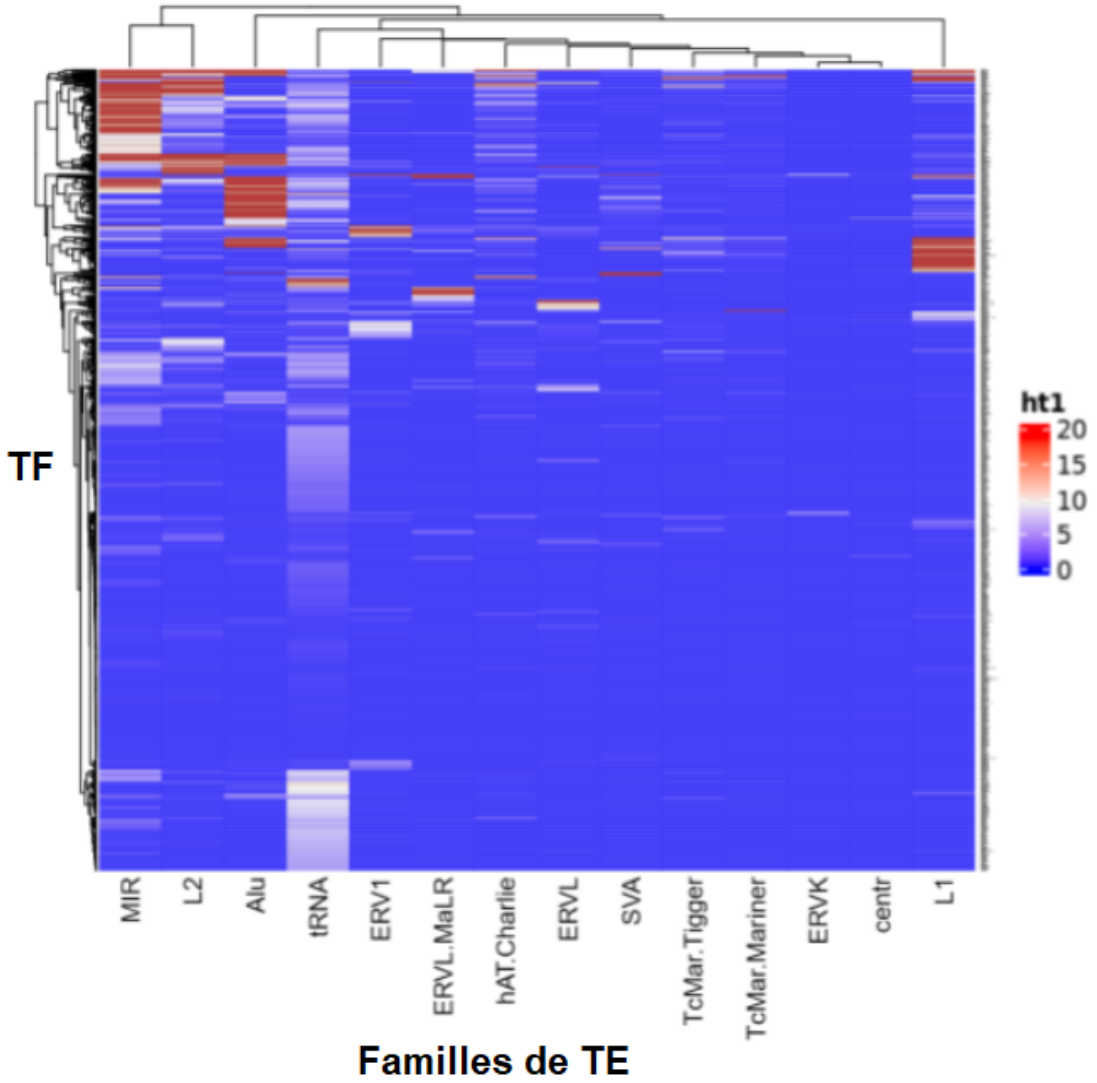


FIGURE 3.14. – *Heatmap représentant du -log(p-value) d'enrichissement des pics ChIP-seq de TF dans les séquences de famille de TE (ex : MIR, L2, etc...).*

Après avoir observé l'enrichissement des 14 familles de transposables éléments (MIR, L1, L2, etc.), nous souhaitons examiner l'enrichissement au niveau des 622 éléments transposables individuels (MER101, L1PA2, etc.) afin d'identifier si ces associations existent au niveau du TE.

La visualisation de la heatmap a permis de mettre en évidence la complexité des interactions entre les TE et les TF. En effet, en examinant les clusters de la heatmap plus en détail, il a été constaté que certains groupes de TF sont enrichis dans des séquences de TE spécifiques. Par exemple, les GATA (GATA1, GATA2, GATA4, GATA6...) sont sur-représenté dans les séquences des ERV, tandis que les CEBP (CEBPA, CEBPB, CEBPD, ...) et les ONECUT1 et 2 (Figure 3.15) sont enrichis dans les séquences des L1. Cette concentration a mis en évidence l'association de certains TF avec des TE de la même famille, et de la même façon certains TE sont associés à un groupe de TF spécifique.

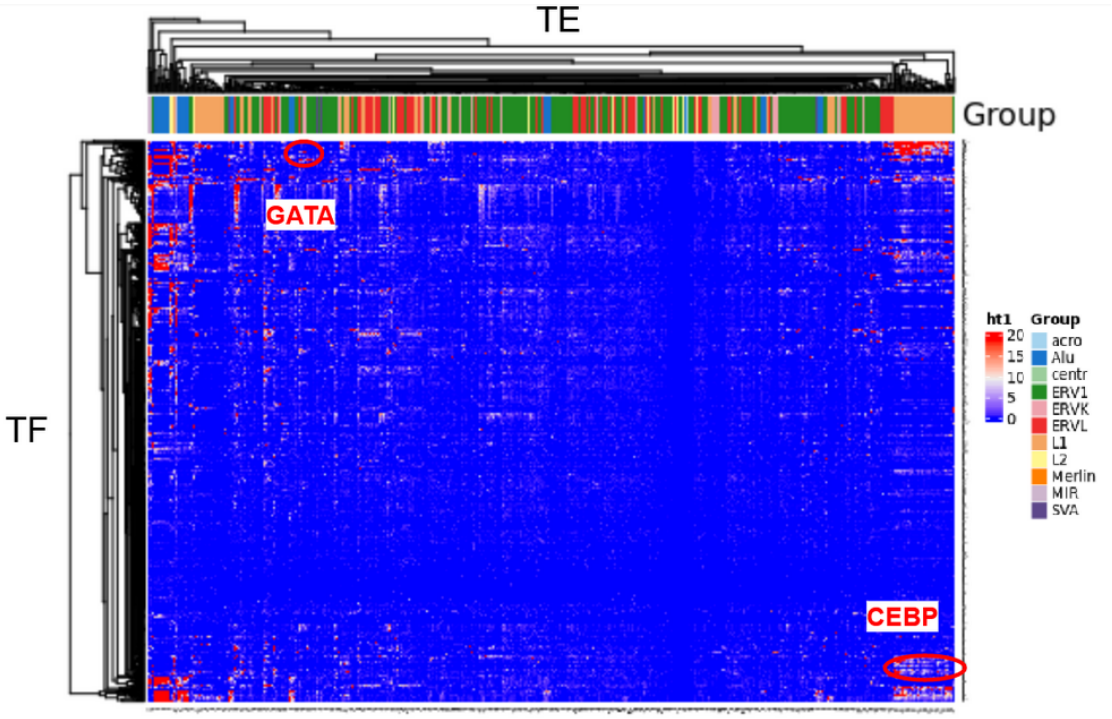


FIGURE 3.15. – ***Heatmap avec les résultats des p-value calculé par LOLA pour les paires TE/TF.*** *Les lignes représentent les 1210 TF et les colonnes représentent les 692 TE. Au-dessus des colonnes les couleurs représentent les familles de TE certaines familles sont d'ailleurs regroupées entre elles.*

Dans la Figure 3.16, nous pouvons apprécier de manière plus détaillée un cluster constitué uniquement des facteurs de transcription ONECUT*. Nous constatons des profils similaires au niveau des TE enrichis, lesquels appartiennent majoritairement à la même famille, les L1. Ce type de profils peut également être observé pour plusieurs autres groupes de TF. Cette observation renforce notre conviction selon laquelle les sur-représentation de TF ne sont pas aléatoires et sont spécifiques à la famille de TE.

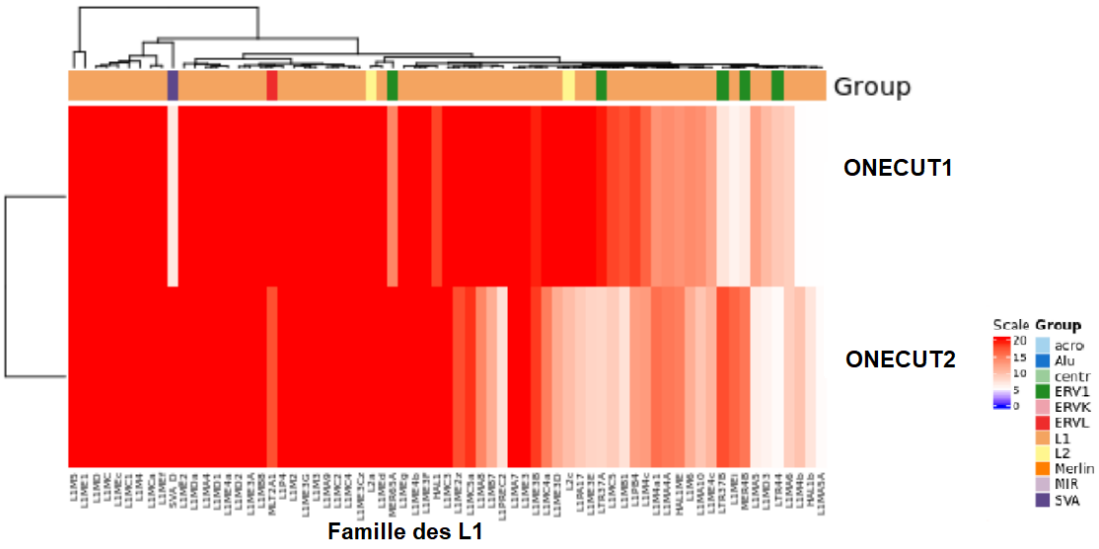


FIGURE 3.16. – ***Heatmap d'enrichissement des pics de ONECUT sur les L1.*** *Zoom sur un cluster composé uniquement des TF ONECUT*, avec des TE L1.*

Notre analyse se distingue de la bibliographie existante en ce qu'elle ne se limite pas à quelques TF (de 1 à 26 TF) [162, 35, 283] ou TE (ALU, L1) [231, 134]. Au contraire, nous avons étudié l'enrichissement pour un grand nombre de couples TE/TF, ce qui nous a permis d'avoir une vision globale de l'influence des TE sur l'insertion des régions régulatrices dans le génome. Les autres études ne fournissent pas véritablement cette information, car elles se concentrent sur des cas individuels. Cette analyse a confirmé l'hypothèse initiale et a incité à poursuivre les recherches pour caractériser cet enrichissement des pics de TF dans les séquences de TE.

3.4.3. Affinité de groupe de TF aux familles de TE

3.4.3.1. Présence de motifs dans les séquences de TE associés

Nous avons calculé l'enrichissement des pics de TF dans les séquences de TE et identifié 15,441 paires TE/TF associées à un enrichissement significatif. Dans cette partie nous nous concentrons sur la caractérisation plus spécifique de ce phénomène. En effet, nous savons maintenant que les TF se fixent sur les séquences de TE. Cependant, la fixation d'un TF sur une séquence d'ADN se fait sur la séquence de son motif de fixation. Cela suggère donc que les motifs de fixation des TF se trouvent potentiellement dans les séquences de TE, ce qui pourrait expliquer comment les TE peuvent insérer de nouvelles régions régulatrices dans le génome.

Nous voulons également déterminer s'il existe une spécificité d'associations entre les TE et les TF. Les heatmaps révèlent une telle spécificité, mais nous ne savons pas si cela est dû aux motifs présents dans les séquences de TE. Dans ce chapitre, nous tenterons donc de répondre à ces questions en déterminant la présence éventuelle de ces motifs dans les séquences tout en caractérisant la spécificité de l'association.

Pour ce faire nous avons procédé à la recherche de motifs dans les séquences de TE pour 7,757 de ces paires (pas les 15,441 car nous n'avions pas les motifs JASPAR de tous les TF). Le pourcentage de motifs présents dans les séquences de TE a été calculé pour chacune des paires présentant un enrichissement significatif.

La Figure 3.17 représente les pourcentages de motifs (PM) pour les paires TE/TF associées pour chaque famille de TE. Pour chaque famille de TE issue de paire TE/TF, nous calculons le pourcentage de séquences de TE avec un ou des motifs de fixation. Par exemple, dans la paire MIR/CTCF, 50% des TE contiennent un motif parmi les n=154 séquences de TE. Pour chaque famille de TE, nous évaluons donc la distribution globale des motifs de fixation identifiés dans les séquences (par paire de TE/TF) (voir section 3.3.4). Nos résultats révelent 1,841 paires de TE/TF avec des PM > 50%.

Nous constatons que les valeurs de ces pourcentages et les profils sont hétérogènes entre les familles de TE. Certaines familles ont des pourcentages de motifs par paires majoritairement supérieurs à 50% (ERV1 et les L1), tandis que d'autres ont des profils avec des pourcentages relativement faibles, comme les MIR (aux alentours de 25%). De manière générale les PM sont plus élevés pour les TE de la famille LTR (ERV1, ERVK et ERVL) et la famille des L1. Ces résultats peuvent être retrouvés dans les resultats de travaux similaires. Par exemple, l'article de Sundaram et al. [283] décris 710 associations de TE/TF avec une dominance des LTR. De plus, l'analyse menée par Jiang et al. [133] se concentre sur les TE de la famille des L1 et révèle la présence de motifs de fixation dans au moins 10% des TE pour 114 TF.

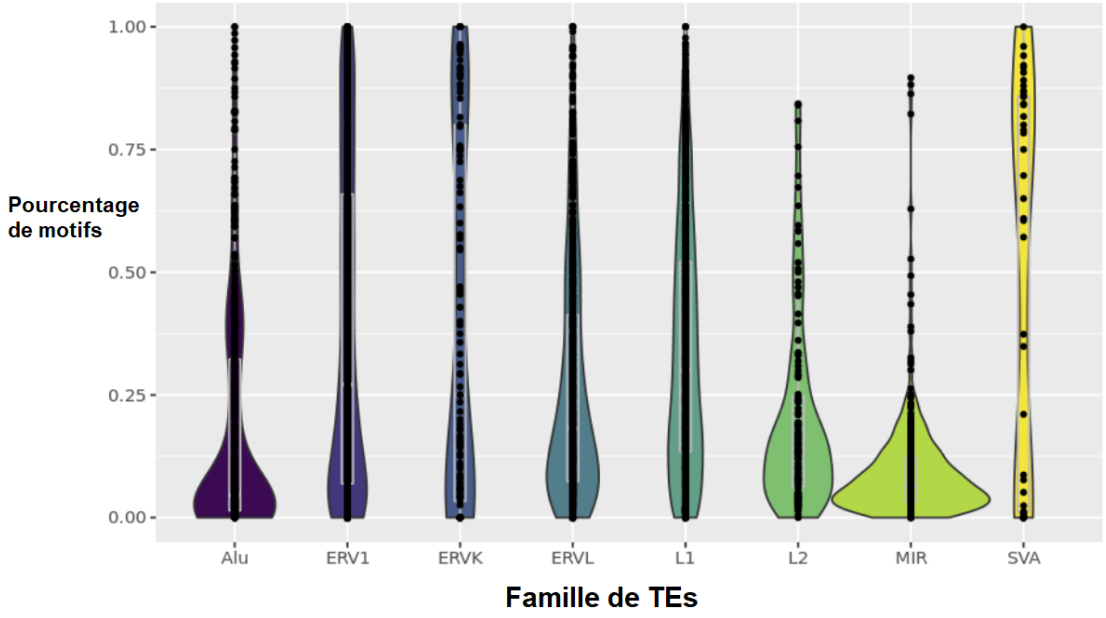


FIGURE 3.17. – ***Violin plot des pourcentages de motifs par familles de TE.*** *L'axe vertical représente le pourcentage de motifs identifiés dans les séquences de TE des couples TE/TF avec un enrichissement significatif. L'axe horizontal représente la famille du TE du couple TE/TF.*

3.4.3.2. Association TE/TF spécifique

Après avoir établi une vue d'ensemble de la présence de motifs dans les séquences de familles de TE, nous souhaitons à présent explorer la spécificité de ces motifs en fonction de la famille de TE et du groupe de TF. En effet, les groupes de TF ont souvent des séquences de motifs de fixation très similaires, ce qui pourrait expliquer une éventuelle spécificité.Il est important de rappeler que nous avons décidé de regrouper les facteurs de transcription en "groupes" en fonction de leur nom, par exemple en regroupant les facteurs GATA* (GATA1,2,3,4,5,6) ensemble.

La Figure 3.18 présente un diagramme en bulles avec les facteurs de transcription en colonnes et les familles et superfamilles d'éléments transposables en lignes. Certaines familles de facteurs de transcription se démarquent par le nombre de séquences d'éléments transposables associées, telles que les ZNF, avec plus de 160 séquences associées avec un pourcentage supérieur à 50% de motifs identifiés dans ces séquences. Nous constatons également que les groupes de TF ne sont pas surreprésentés de la même manière dans les familles de TE. Ces observations révèlent la spécificité de l'association entre les paires d'éléments transposables et de facteurs de transcription. Certains des résultats représentés sont déjà connus dans la littérature [283, 184, 303, 316, 304, 140, 35], tandis que d'autres sont nouveaux.

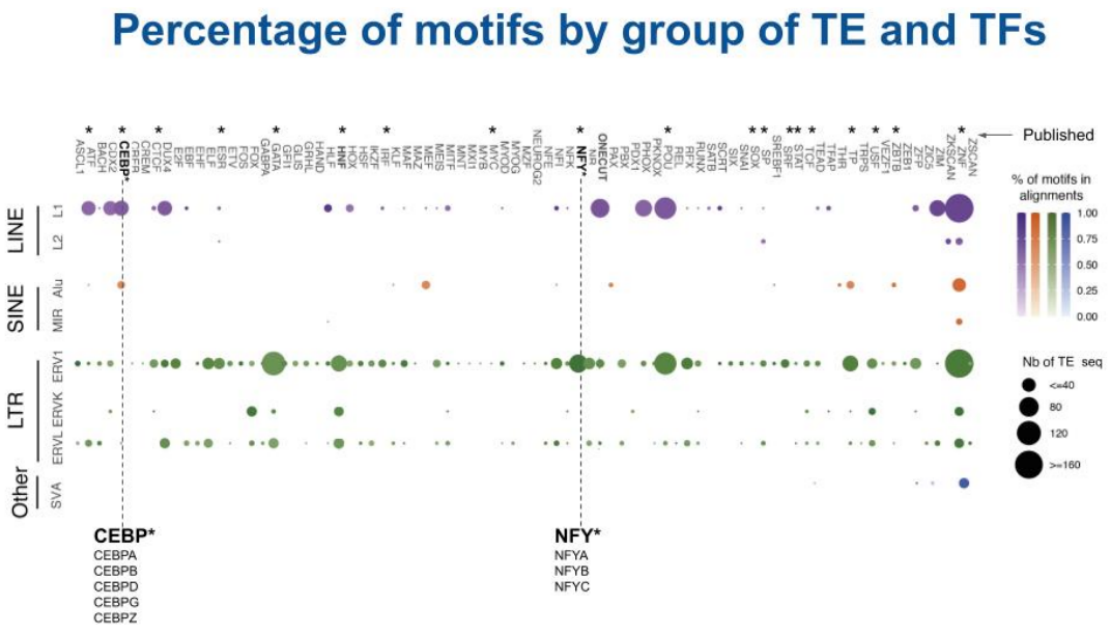


FIGURE 3.18. – ***Pourcentage de motifs par groupe de TF et familles de TE.*** *La couleur des bulles, allant du clair au foncé, représente le pourcentage de motifs identifiés dans les séquences d'éléments transposables. La largeur de la bulle indique le nombre de séquences d'éléments transposables chevauchant les pics du TF associé.*

Nous allons présenter deux cas concrets d'association spécifique détectée lors de notre analyse : la paire NFY*/ERV1 et la paire CEBP*/L1.

3.4.3.3. Exemple : NFY* et ERV1

Le premier exemple, qui coïncide avec des travaux antérieures, concerne les NFY* qui sont surreprésentés dans les séquences d'éléments transposables de la famille ERV1, et plus particulièrement les LTR12* (Figure 3.19).

Les facteurs de transcription Nuclear Factor Y (NFY) sont des complexes protéiques impliqués dans la régulation de l'expression génique. Ils sont formés par l'association de trois sous-unités, appelées NFYA, NFYB et NFYC, qui agissent ensemble pour contrôler la transcription de nombreux gènes.

Les NFY* se lient à des séquences spécifiques de l'ADN appelées CCAAT boxes, qui sont situés dans la région promotrice du gène cible [23]. Les NFY* peuvent fonctionner en tant qu'activateur [59] ou répresseur [148] en fonction des cofacteurs avec lesquelles ils interagissent. Ces TF sont souvent associées à des gènes impliqués dans des processus biologiques importants tels que la croissance cellulaire [23] et la différenciation cellulaire [219].

Les éléments transposables LTR12* font partie de la superfamille des LTR et de la famille des ERV1 [213].

Il existe plusieurs travaux qui indiquent un lien entre les éléments transposables LTR12* et les facteurs de transcription NFY* comme observé dans la Figure 3.19. Par exemple, un article de 2022 [128] révéle que les motifs de fixation des NFY* (CCAAT) était fortement conservé dans les séquences de 7 éléments transposables LTR12*. Ces travaux rejoignent une analyse de 2016 [159] ayant établit l'impact des NFY* sur le rôle régulateur des éléments LTR12*. Les travaux récents de Neumayr et al. [209] révèlent également que les TE de la famille LTR12* agissent comme des enhancers/promoteurs indépendants de BRD4 qui contiennent une combinaison de motifs TATA-box et de multiples motifs CCAAT-box.

En conclusion, les résultats des travaux précédents renforcent les observations faites lors de notre analyse. Ils montrent un lien avéré entre les facteurs de transcription NFY* et les éléments transposables LTR12*.

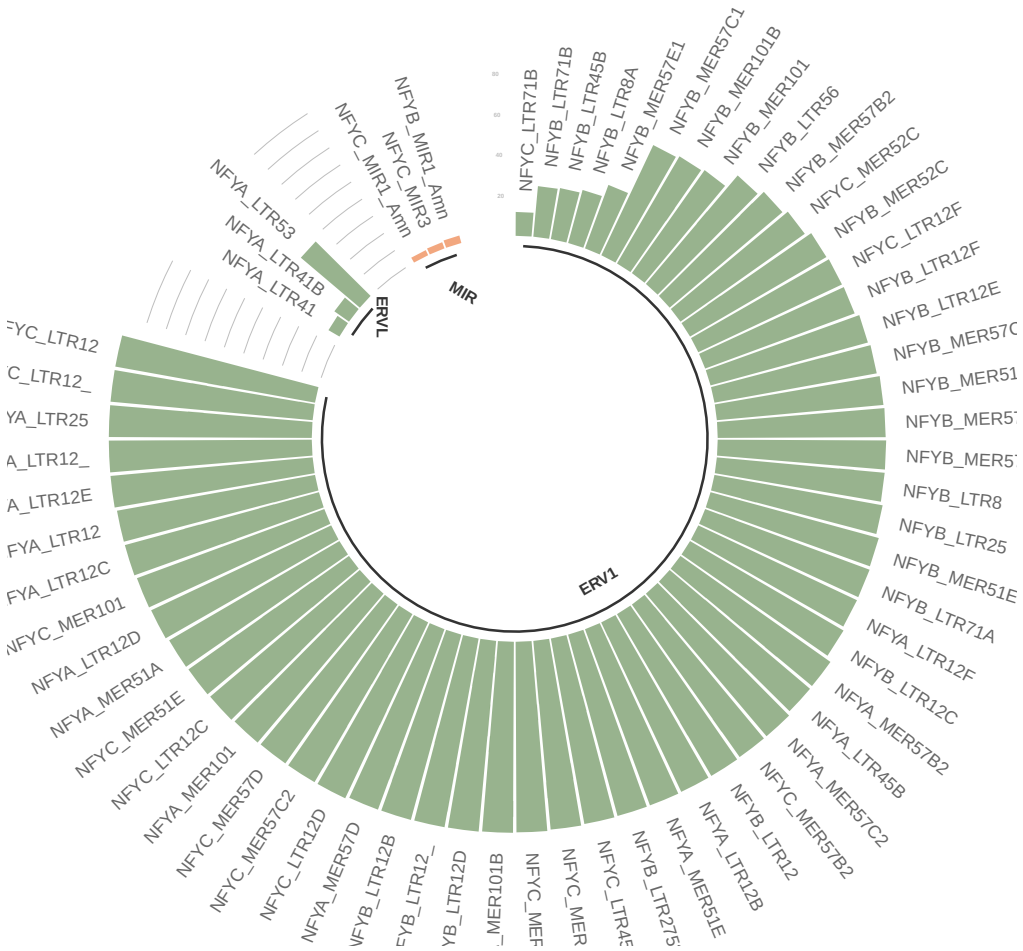


FIGURE 3.19. – ***Barplot circulaire du groupe de TF NFY* représentant les PM par TE.*** *Chaque bar représente le pourcentage de motifs d'une paire de TE/TF associé. Les superfamilles sont séparé avec un code couleur. En vert les LTR, en violet les LINE, en orange les SINE.*

3.4.3.4. Exemple : CEBP* et L1

Nous examinerons ensuite le cas des CEBP* qui sont fortement associés aux séquences d'éléments transposables de la famille L1 (Figure 3.20). Cette association n'étant pas explicitement décrite dans la littérature, nous proposons de fournir une interprétation de ces résultats à partir de nos connaissances.

D'après les résultats de la recherche de motifs, le motif de fixation des TF du groupe des CEBP* a été identifié dans plus de 50% des séquences de la famille L1 auxquelles les CEBP* se fixent (Figure 3.20). Contrairement à NFY*, cette association n'a pas encore été mise en avant par d'autres travaux. Nous allons ainsi tenter d'expliquer cette spécificité en décrivant les fonctions des CEBP*, afin d'établir un lien avec les L1.

Les protéines CEBP (CCAAT enhancer-binding proteins) sont une famille de 6 facteurs de transcription multifonctionnels avec un domaine bZIP (basic leucine zipper). Ils sont définis par des domaines conservés de la terminaison carboxylique comprenant une dimérisation de la fermeture bZIP et des domaines de liaison à l'ADN basiques. Il existe six isoformes différentes de CEBP*, notamment CEBPα, CEBPβ, CEBPγ, CEBPδ, CEBPϵ et CEBPζ. Chacune de ces isoformes peut jouer un rôle différent. Les isoformes CEBPα, β et δ sont impliquées dans la différenciation des cellules graisseuses et la régulation du métabolisme [66]. CEBPα joue également un rôle important dans la régulation de l'expression des gènes liés au cancer [138]. CEBPδ serait notamment impliqué dans plusieurs processus tel que la régulation de la réponse au stress, l'inflamation et la prolifération cellulaire[240].

Les rétrotransposons L1 (LINE-1) sans LTR sont la famille de TE la plus importante chez les mammifères. Le génome humain contient environ 500 000 éléments L1, représentant 17% de sa masse totale [2], témoignant l'effet profond que la réplication des L1 a eu sur les génomes des mammifères. La rétrotransposition des L1 produit principalement des copies défectueuses qui restent dans le génome et accumulent des mutations à un taux similaire à celui des pseudogènes. De nouvelles variantes de L1 compétentes pour la réplication peuvent également être produites, et peuvent ensuite donner naissance à une famille de plusieurs centaines ou milliers de copies qui ont des caractéristiques similaires à leur progéniteur (ou groupe de progéniteurs étroitement liés) [95].

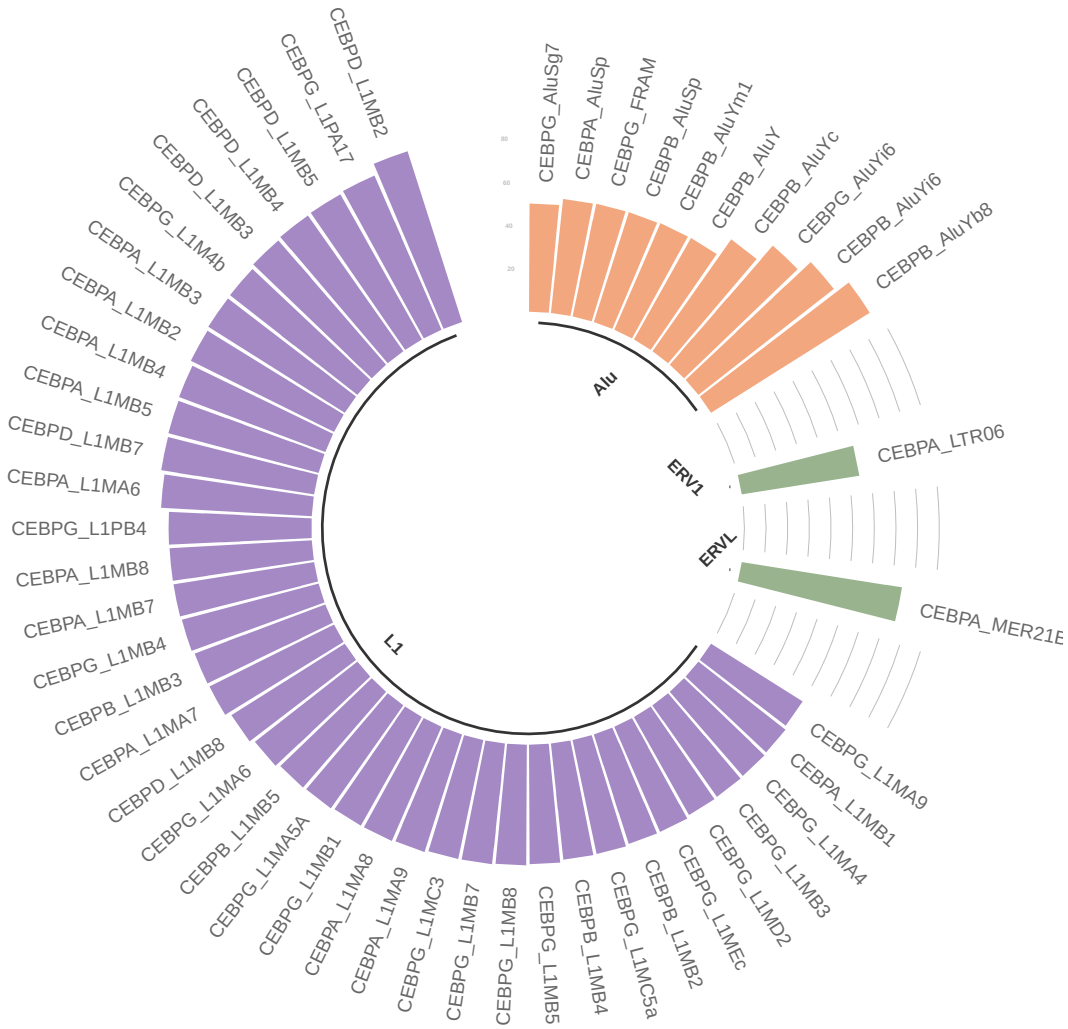


FIGURE 3.20. – ***Barplot circulaire du groupe de TF CEBP* représentant les PM par TE.*** *Chaque bar représente le pourcentage de motifs d'une paire de TE/TF associé. Les paires représentées ont un PM > 50%. Les superfamilles sont séparé avec un code couleur. En vert les LTR, en violet les LINE, en orange les SINE.*

Les éléments de la famille L1 peuvent avoir une influence sur l'expression génique par leur capacité à la transcription bidirectionnelle, en effet ils ont deux promoteurs dans leur régions UTR, un antisens et un sens [191]. L'expression anormale de ces éléments peut survenir en cas de modification de l'environnement épigénétique et peut conduire à l'expression anormale de gènes dans le cancer. De plus, la répression de ces éléments peut avoir un impact sur l'expression génique même à de grandes distances des promoteurs, en créant des îles hétérochromatiques qui peuvent réguler quantitativement les niveaux d'expression de nombreux gènes humains. Le génome hôte doit donc trouver un équilibre entre la suppression nécessaire de ces éléments et les conséquences collatérales qu'elle peut avoir sur l'expression génique. [284].

Dans un article de 2013 [305], Wanichnopparat et al. ont examiné le rôle des L1 intragéniques dans la régulation de l'expression. Ils ont obtenu les profils d'expression de 205 gènes à partir d'expériences de knockdown et ont comparé les niveaux d'expression des gènes avec et sans L1. Après une analyse statistique basée sur un test d'hypothèse multiple, 73 gènes ont été identifiés comme étant impliqués dans la régulation des gènes avec L1. Les régulations peuvent varier en fonction du type de cellule et de l'orientation des L1 intragéniques. De plus, les gènes régulés par siRNA qui contiennent des L1 jouent un rôle dans plusieurs fonctions moléculaires, des phénotypes cellulaires et sont associés à diverses maladies. Les résultats suggèrent que les cellules utilisent les L1 intragéniques comme éléments *cis*-régulateurs pour moduler l'expression génique et peuvent être impliqués dans la régulation de divers processus biologiques, y compris les dommages de l'ADN et la réparation, l'inflammation, la fonction immunitaire, l'embryogenèse, la différenciation cellulaire, la réponse cellulaire aux stimuli externes et les réponses hormonales.

Parmi les 73 gènes impliqués dans la régulation de gène contenant des L1 on retrouve le gène CEBPB que nous avons identifié. On peut supposer à partir de cet article et de nos résultats que les CEBPB se fixent sur ces séquences de L1 intragéniques et régulent potentiellement l'expression de ces mêmes gènes. La fixation se fait via le motif de fixation des CEBP* identifié dans les séquences de L1. Cette hypothèse est confortée par le fait que les CEBP* et les L1 sont également impliqués dans la différentiation cellulaire, l'inflammation et la réponse au stress.

3.4.4. Alignements des TE

Dans nos analyses antérieures, nous avons identifié des associations spécifiques entre certaines familles de TE et de TF, telles que les NFY* avec les ERV1. Cette analyse a été réalisée en calculant le pourcentage de motifs identifiés dans les séquences de TE d'une paire TE/TF associée avec un enrichissement statistiquement significatif.

Cependant, pour établir une relation plus directe entre les TE et les TF, nous devons avoir une vision plus claire de la position exacte des motifs identifiés dans les séquences de TE. Pour cela nous avons réalisé un l'alignement des séquences de TE sur lesquelles les TF se fixent. Cet alignement permet de visualiser avec précision la position des les motifs de fixation des TF dans les séquences de TE.

Cette analyse est basée notamment sur la figure 1d de l'article de Kunarso et al. [162]. Les auteurs ont calculé la surreprésentation des TFBS de quatres TF différents POU5F1, NANOG, OCT4 et CTCF se fixant sur les séquences de différentes familles de TE dans les génomes de l'Homme et de la souris. Ils ont identifié des associations spécifiques entre les TF et les TE plus fréquentes que ce qui était attendu par hasard. Par exemple, ils ont observé que 33,2% des TE de la famille ERV1 étaient fixés par OCT4. Les auteurs ont appelé ces TFBS associés à des répétitions "RABS". Dans de nombreux cas, l'alignement des séquences associées d'une famille de TE particulière a mis en évidence un haut degré de similitude d'une région provenant de la séquence ancestrale et renfermait le motif de fixation du TF.

En 2008 Bourque et al. [35] traite de l'expansion des régions régulatrices dans le génome par l'insertion de motifs de fixation des TF dans des TE. L'analyse a montré que des associations TE/TF sont médiées par des motifs de fixation et sont hautement ciblées. Les auteurs ont constaté que les TE associées abritent le motif de fixation de TF contrairement aux TE de la même classe non associé. L'alignement des 17 séquences associées de la répétition RLTR11B vérifient la présence du motif POU5F1-SOX2 et sa conservation.

Nous avons donc décidé d'appliquer la même analyse à nos données. Nous avons aligné les séquences de TE associées pour vérifier si elles sont alignées sur le motif de fixation du TF. Cette analyse sera présentée à l'aide de l'exemple de la paire MER101/NFYC.

La Figure 3.21 présente un exemple d'alignement de séquences de MER101 auxquelles se fixent les séquences du TF NFYC. Nous observons un alignement des séquences de MER101 sur le motif de NFYC. La fixation des NFYC dans les MER101 peut donc être largement expliquée par la présence de motifs NFYC.

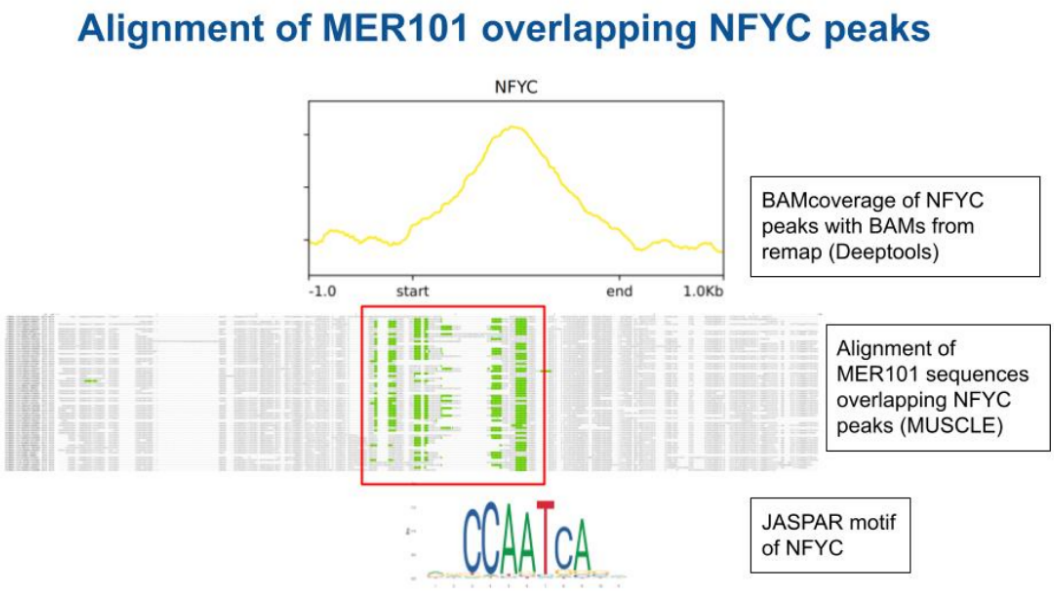


FIGURE 3.21. – ***Alignements des TE de paires TE/TF associé.*** *A) BAMcoverage des pics de NFYC avec l'outil deeptools, les BAMs ont été extrait avec ReMap. B) Alignements MUSCLE des séquences de MER101 chevauchant les pics ChIP-seq de NFYC. C) Motif JASPAR de NFYC. Dans l'encadré rouge, nous pouvons observer les motifs identifiés par fimo (correspondant au motif de JASPAR pour le TF NFYC), surlignés en vert à l'aide de MVIEW.*

Plusieurs hypothèses peuvent être formulées à partir de ces résultats. L'alignement montre une conservation des motifs dans les séquences de TE, renforçant ainsi les conclusions des analyses précédentes de Bourque et al. [35], qui ont découvert que les régions conservées qui présentent le motif font partie de la séquence ancestrale conservée. De plus, le fait que ces régions soient associées à une fixation indique que ces sites de régulation sont potentiellement fonctionnels.

La conservation du motif dans les séquences de TE, peut suggérer une pression sélective de ces régions régulatrices potentiellement fonctionnelles. On peut considérer que les TE ont donc contribué à l'expansion du répertoire des régions régulatrices du génome en introduisant de nouveaux TFBS dans le génome au fil de l'évolution.

3.4.5. Lien entre âge des TE et motifs associés

Après avoir établi que les motifs de fixation des TF ont été introduits par les TE, nous cherchons maintenant à savoir dans quelle mesure les éléments transposables ont eu un impact sur l'évolution de la régulation transcriptionnelle. L'objectif final est de comprendre si les TFBS dérivés des TE ont été insérés à des temps évolutif différent et en quoi cela a pu impacter les fonctions biologiques régulées au cours de l'évolution. On pourrait donc supposer que certains TE insérés à un temps spécifique évolutif pourraient expliquer l'ajout de nouveaux réseaux de régulation dans cette même période.

L'analyse initiale décrite dans Edward B. et al. [50] a révélé la présence de 27 familles de TE enrichies par des pics de fixation induits par IFNG dans au moins un des jeux de données examinés. Ces séquences contenaient des familles de TE jeunes à anciennes, la plupart d'entre elles provenant de la région promoteur de LTR d'éléments de ERV. Ces données suggèrent que les ERV, qui proviennent des infections rétrovirales anciennes et constituent actuellement 8% du génome humain, représentent une source de nouveaux sites de fixation liés par les facteurs de transcription inductibles par IFNG.

Ces résultats ainsi que nos résultats précédents ont inspiré nos travaux sur l'âge des TE associés aux motifs de facteurs de transcription. En effet, lors des analyses précédentes nous avons vu que pour certaines paires les motifs de fixation des TF étaient très conservés dans les séquences de TE. Nous avons donc décidé de représenter l'âge des TE en fonction des pourcentages de motifs des paires TE/TF afin de voir si par exemple les pourcentages de motifs seraient plus élevés dans les séquences les plus anciennes. Par conséquent, nous avons calculé l'âge des TE, basé sur l'article de Bourque et al. [35] et l'article de Pace et Feschotte [220]. La méthode de calcul a été décrite dans la section 3.3.7.

La Figure 3.22 est un diagramme en nuage de points représentant nos résultats. Cette figure permet de visualiser la relation entre l'âge des TE et le pourcentage de motifs de TF identifiés dans les séquences des TE associées. Les abscisses représentent l'âge en Mya des TE et les ordonnées représentent les pourcentages de motifs identifiés. Les points sur le graphique représentent les paires TE/TF associées et les couleurs des points indiquent les familles de TE.

D'après cette figure, il semble que les familles de TE soient regroupées les unes avec les autres en fonction de leur âge d'insertion, ce qui correspond à ce qui est rapporté dans la littérature [233, 95]. Par exemple, les TE de la famille des L2 sont regroupées entre 180 Mya et 200 Mya, ce qui concorde avec leur âge d'insertion [230]. Dans la Figure 3.22 on peut voir que la plupart des TE sont situés dans une période allant de 50 à 180 millions d'années. Par ailleurs, on observe moins de TE dans la période la plus récente, l'ère post-anthropoïde. Il semble qu'entre 50-100 Mya il y ait un nuage de points important. Ce qui pourrait signifier que la majorité des TFBS ont été inséré à ce moment-là lors de l'évolution. Cette figure rejoint les résultats de Jacques PE et al. [130] (décrit plus bas) quant à la distribution des ERV dans l'intervalle de 50 à 180 Mya. Mais on observe également des pourcentages de motifs plus élevé entre 50 et 100 Mya, tendance aussi observé dans la Figure 3.22 de l'article de Jacque PE et al. avec dans ce même intervalle des pourcentage de chromatine ouverte plus élevé. Les régions régulatrices actives sont situé dans les régions de chromatine ouverte. Les TE inséré entre 50 et 100 Mya seraient donc potentiellemnt sources de nouvelles régions régulatrices actives introduite dans ce même intervalle de temps sur le génome.

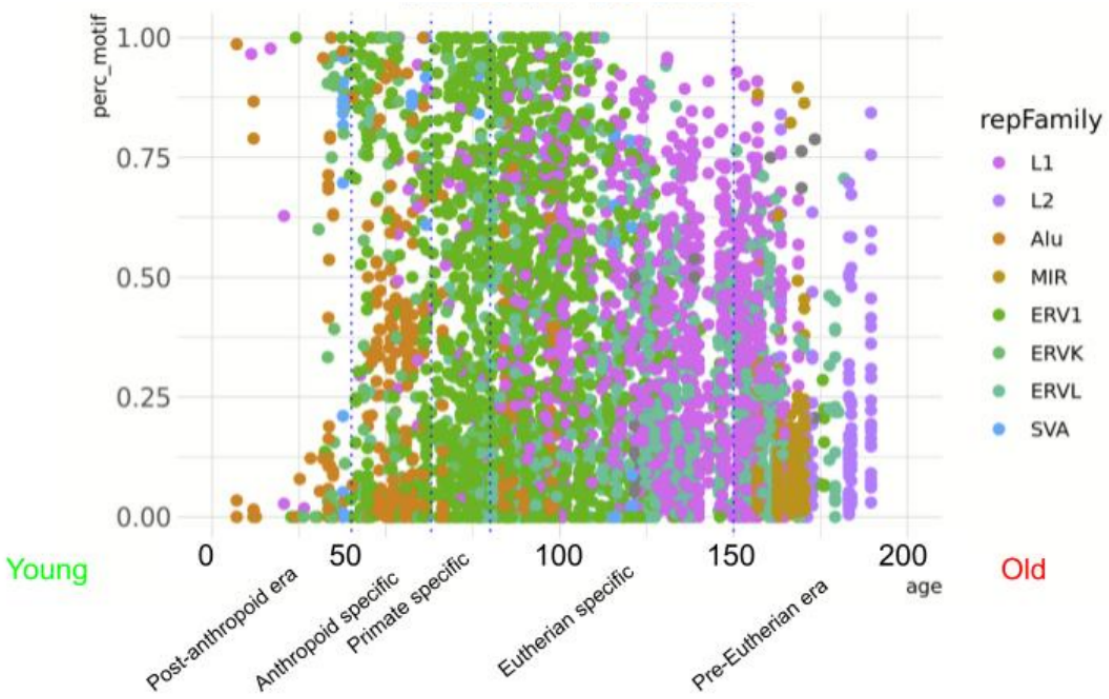


FIGURE 3.22. – *Age des repeats en fonction du pourcentage de motifs de fixation dans les séquences des TE.*

Ensuite, nous avons cherché un lien entre l'âge d'insertion des TE et les groupes de TF associés. Cette analyse est inspirée d'un article révélant un lien entre les périodes d'évolution des TE et leur lien spécifique à un TF.

Dans la Figure 3.23, nous avons examiné le lien entre l'âge des TE et les différents groupes de TF auxquels ils sont associés. Nous pouvons constater que les profils d'âge associés à chaque groupe de TF sont différents. Par exemple, les facteurs de transcription ZKSCAN sont associés à des TE anciens tandis que les PAX sont associés à des TE plus récents. Cette différence dans les profils d'âge n'est probablement pas attribuée au hasard, et il serait donc intéressant d'examiner plus en profondeur la signification biologique à travers des analyses que nous n'avons pas eu le temps de réaliser au cours de cette thèse et qui seront discutées dans la partie discussion et perspectives.

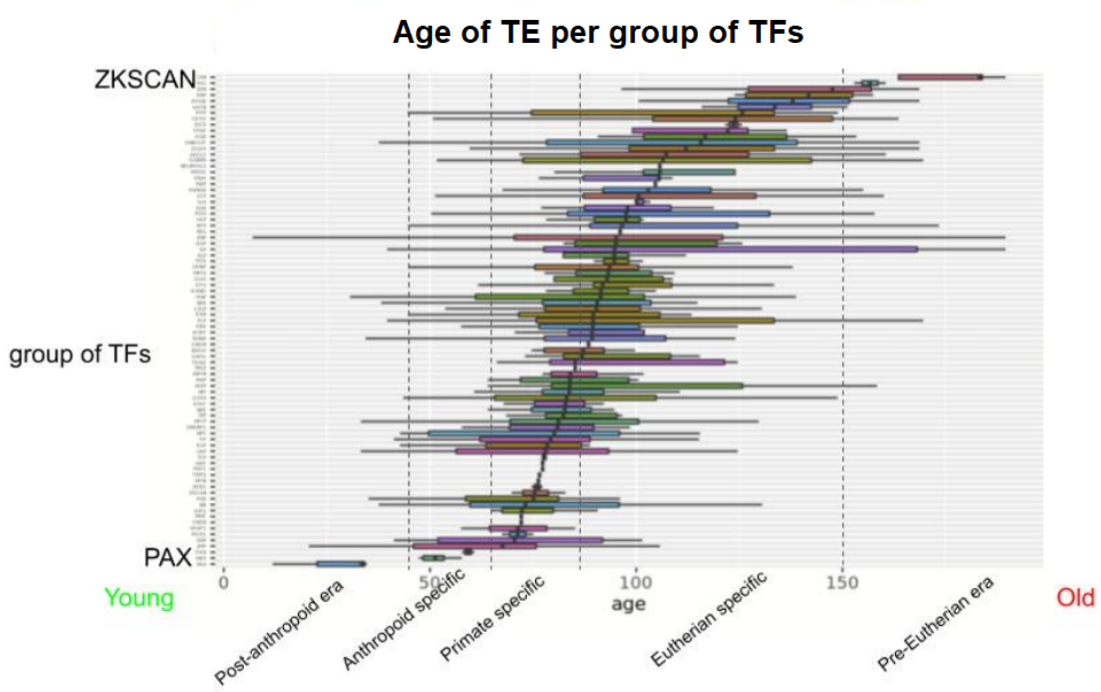


FIGURE 3.23. – ***Age des repeats par groupe de TF.*** *L'axe des ordonnées représente les groupes de TF et l'axe des abscisses représente l'âge des TE, qui ont été divisés en périodes d'évolution basé sur l'article de Pace et Feschotte 2007 [220] pour faciliter l'interprétation des résultats. Pour chaque groupe de TF, nous avons utilisé un boxplot pour représenter l'âge des TE associés à ce groupe de TF.*

Plusieurs hypothèses émanant de cette analyse, la première étant que l'âge d'insertion des TE peut jouer un rôle dans l'incorporation de nouveaux TFBS dans le génome humain au cours de l'évolution. Les TE les plus anciens peuvent être à l'origine de mécanismes de régulation anciens. Alors que les TE les plus récents seraient responsables de mécanismes de régulation plus récents.

3.5. Conclusion

Il y a plus de 50 ans, Barbara McClintock [194] a découvert pour la première fois les éléments transposables dans le maïs (*Zea mays*) et les a appelés "éléments de contrôle" car ils altèrent l'expression génétique. Roy Britten et Eric Davidson [38] ont alors proposé que les systèmes régulateurs coordonnés dans les génomes animaux soient en réalité encodés par des réseaux de TE. En revisitant et en élargissant les modèles antérieurs, Cedric Feschotte [89] décrit les TE, comme étant une source abondante de matériaux pour l'assemblage et la modification des systèmes régulateurs des gènes eucaryotes au cours de l'évolution. Des travaux en génomiques [304, 35, 162, 253] ont fourni davantage de preuves appuyant ce mécanisme. Certains TE spécifiques à une lignée ont été identifiés comme responsables de l'insertion de TFBS. Par exemple, certains ERV spécifiques à des primates contenaient plus de 30% des sites de fixation pour la protéine suppresseur de tumeur TP53 [304]. De manière similaire, environ 20% des sites de fixation de POU5F1 et NANOG sont situés dans les séquences de TE spécifiques à une lignée chez l'Homme et la souris [162], et une expansion spécifique aux rongeurs des sites de fixation de CTCF a également été reliée à l'insertion de rétrotransposons chez les rongeurs [253]. En dehors de la fixation de TF, des travaux ont déterminé que les TE contribuent à des sites de sensibilité à la DNase I [130]. Les progrès récents dans ce domaine ont révélé qu'une partie substantielle des éléments *cis*-régulateurs chez les mammifères sont dérivée de TE. Ils sont souvent spécifiques au type de cellule et à l'espèce et peuvent contribuer à la régulation de l'expression génique par de nombreux mécanismes [284]. Ainsi la grande partie des génomes mammifères contiennent des TE ayant contribué à l'évolution des éléments *cis*-régulateurs, peut-être par le biais du renouvellement des TFBS ou leurs expantions lors de la transpositions des TE[89].

En utilisant la capacité de ReMap à identifier la fixation des facteurs de transcription même dans les régions riches en TE, nous avons montré que, pour 693 des 1210 TF de ReMap, il existe un enrichissement significatif dans 622 TE, formant ainsi 15,441 paires TE/TF enrichies. Les résultats identifient une préférence de certains groupes de TF pour certaines familles de TE. Nous avons pris deux exemples illustrant cette spécificité. D'une part, les TE de la famille LTR12* ont récemment été rapportées comme étant associées au TF du groupe NFY* [128] cette association a été confirmée par nos résultats. D'autre part, les TF du groupe CEPB* présenteraientt également une affinité pour les TE de la famille L1.

Nos analyses révèlent la présence de motifs de fixation dans plus de 50% des TE pour 1,841 paires TE/TF. Les séquences des motifs sont non seulement incluses dans ces TE, mais que dans de nombreux cas, les motifs de fixation sont hautement conservés dans les séquences de TE, suggérant une sélection évolutive. Pour tenter d'expliquer ces phénomènes, nous avons calculé l'âge des TE. Les résultats révèlent que les familles de TE sont regroupées en fonction de leur âge et que la plupart des TE ont été inséré à une période allant de 50 à 180 millions d'années. Les profils d'âge associés à chaque groupe de TF sont distincts et n'ont pas été attribués au hasard. Les résultats suggèrent un lien entre l'âge des TE et les TF avec lesquels ils sont associés. En outre, la tendance selon laquelle une grande partie des TE ont été inséré entre 50 et 100 Mya rejoint les travaux de Jacques PE et al. [130], identifiant dans ce même intervalle des pourcentage de chromatine ouverte plus élevé. Ce qui supposerait que les TE inséré dans cet intervalle de temps d'évolution seraient sources de nouvelles régions régulatrices actives.

En résumé, les TE ont probablement joué un rôle clé dans l'évolution du système régulateur génétique des espèces. Ils ont probablement introduit de nouveaux TFBS dans le génome, ce qui a modifié le répertoire des éléments *cis*-régulateurs, tels que les enhancers. Nos travaux ont permis de découvrir un grand nombre d'associations TE/TF qui n'ont pas été publiées précédemment. Certaines de ces associations semblent être très récentes, tandis que d'autres peuvent être très anciennes.

Cependant, malgré l'importance potentielle de ces associations TE/TF, il reste beaucoup à comprendre sur leur fonctionnalité biologique. Certaines de ces associations peuvent être simplement de nature neutre et ne pas avoir d'influence sur l'expression des gènes, tandis que d'autres peuvent avoir des conséquences significatives pour l'évolution et les différences phénotypiques entre les espèces. Pour mieux comprendre le rôle des TE dans l'évolution, il est donc important de poursuivre les recherches sur leurs interactions avec les facteurs de transcription et leur impact sur les gènes.

4. Régulation des ALDH

Sommaire

4.1. Contexte

Ma thèse a été financée par une bourse INSERM-Région SUD qui implique un travail de collaboration entre un acteur académique et un laboratoire privé. Dans ce contexte nous avons collaboré avec le laboratoire Advanced BioDesign (ABD [1]), pour une étude de la régulation du gène aldéhyde déshydrogénase 1 de type A1 (ALDH1A1) dans les leucémies myéloïdes aigües (AML) notamment grâce au catalogue de région régulatrice ReMap.

Les AML sont des cancers graves du tissu hématopoïétique qui se caractérisent par une absence de différenciation des précurseurs hématopoïétiques. Les tissus atteints prolifèrent ensuite dans la moëlle osseuse et le sang. L'AML est plus fréquente chez les adultes. Les traitements existants sont la chimiothérapie d'induction, de consolidation ou la transplantation de moëlle osseuse. Le taux de survie après 5 ans est de moins de 50% [73].

L'ALdehyde DeHydrogenase (ALDH) est utilisé comme un marqueur de l'AML [300]. En effet, les patients atteints d'AML avec un mauvais pronostic ont une activité positive des ALDH sont plus élevés (citation). En conséquence, ABD travaille depuis 2010 sur un inhibiteur des ALDH le DIMATE. Le DIMATE a été testé en 2023 sur l'humain dans le cadre d'un éssai clinique de phase I. Ce traitement inhibe directement les ALDH [64].

Le but de cette collaboration était d'identifier le régulateur de la transcription de ALDH1A1. Ce projet pourrait mener à l'identification d'une nouvelle cible thérapeutique qui dans ce contexte pourrait être utilisée en complément du DIMATE.

Pour remplir ces objectifs nous avons décidé de construire un trackhub de données multi-omiques (ReMap, DNase, ATAC, ChIP-seq d'histones). Nous avons donc intégré ces différentes données à partir de datasets GEO des tissus AML. Ces données en combinaison du catalogue ReMap permettront de trouver les régions régulatrices du génome actives autour du gène ALDH1A1. La liste des pics ChIP-seq de ReMap permettras ensuite d'identifier les facteurs de transcriptions qui se fixent sur les régions activent et qui potentiellement régulent le gène ALDH1A1. Les résultats de cette analyse ont ensuite été communiqués à l'équipe d'ABD, afin de procéder aux validations expérimentales qui sont toujours en cours.

1. https://www.a-biodesign.com/

4.2. Données

Pour cette étude, les données utilisées sont les données ChIP-seq du catalogue ReMap, ainsi que des données DNase, ATAC et ChIP-seq d'histones d'AML provenant de GEO. Les données ont été utilisées pour construire un trackhub, qui est une ressource en ligne permettant d'accéder à des données multi-omiques pour une analyse approfondie des régions régulatrices. Les lignées cellulaires sélectionnées sont listées dans la Table 4.1.

TABLEAU 4.1. – *Tableau listant les lignées cellulaires avec des données multi-omiques disponibles utilisé dans ces travaux.*

Lignées cellulaires AML
AML
AMLPZ12
K-562
MV4-11
MV4-11-B
THP-1
MLL-AF9
MM-1
MM-6

4.2.1. ChIP-seq

4.2.1.1. ReMap

Les données de ReMap utilisées sont celles de la mise à jour de 2020, au moment de la construction du trackhub nous n'avions pas accès aux nouvelles données de 2022. Elles ont été filtrées via les filtres UCSC, les lignées cellulaires sélectionnées ont été listées dans le Tableau 4.1. Les données ReMap représentent le cœur de cette analyse, les pics permettent à la fois de détecter les régions régulatrices proches du gène ALDH1A1 ainsi que d'avoir une liste de TFs se fixant sur ses régions. Ces TFs peuvent être potentiellement les régulateurs du gène ALDH1A1.

4.2.1.2. Histones

J'ai intégré dans le trackhub des données ChIP-seq d'histones. Ces données sont composées de 8 expériences différentes d'histone H3K27 acétylé et H3K4 triméthyl. Il y a 22 tracks de H3K27ac et 19 tracks de H3K4me3. Les pics de H3K27ac sont utilisée pour détecter les régions enhancers [113].

accession	Name	Assembly	LINK	target	biotype	experiment_id	FORMAT
GSE118963	GSE118963.H3K27ac.AML_CEBPA_biallelic	hg19	https://ftp.	H3K27ac	**AML**	GSM3354812	BW
GSE118963	GSE118963.H3K27ac.AML_CEBPA_monoallelic	hg19	ftp://ftp.nc	H3K27ac		GSM3354813	BW
GSE118963	GSE118963.H3K27ac.AML	hg19	ftp://ftp.nc	H3K27ac		GSM3354814	BW
GSE128259	GSE128259.H3K27ac.MM-1_rep1	hg38	ftp://ftp.nc	H3K27ac	**MM-1**	GSM3669483	BIGWIG
GSE128259	GSE128259.H3K27ac.MM-1_rep2	hg38	https://ftp.	H3K27ac		GSM3669484	BIGWIG
GSE128259	GSE128259.H3K27ac.MM-6_rep1	hg38	https://ftp.	H3K27ac	**MM-6**	GSM3669485	BIGWIG
GSE128259	GSE128259.H3K27ac.MM-6_rep2	hg38	https://ftp.	H3K27ac		GSM3669486	BIGWIG
GSE128261	GSE128261.H3K27ac.K562	hg38	https://ftp.	H3K27ac	**K562**	GSM3669509	BIGWIG
GSE128261	GSE128261.H3K27ac.K562_KDM6A-KO	hg38	https://ftp.	H3K27ac		GSM3669510	BIGWIG
GSE128261	GSE128261.H3K27ac.K562_KDM6A-KO_doxy	hg38	ftp://ftp.nc	H3K27ac		GSM3669511	BIGWIG
GSE128261	GSE128261.H3K27ac.THP-1	hg38	https://ftp.	H3K27ac	**THP-1**	GSM3669512	BIGWIG
GSE128261	GSE128261.H3K27ac.THP-1_KDM6A-KO_doxy	hg38	https://ftp.	H3K27ac		GSM3669513	BIGWIG
GSE61785	GSE61785.H3K27ac.MV4-11_EZH2-high	hg19	https://ftp.	H3K27ac	**MV4-11**	GSM1513830	WIG
GSE61785	GSE61785.H3K27ac.MV4-11_EZH2-low	hg19	https://ftp.	H3K27ac		GSM1513831	WIG
GSE61785	GSE61785.IgG.MV4-11_EZH2-high	hg19	https://ftp.	H3K27ac		GSM1513834	WIG
GSE61785	GSE61785.IgG.MV4-11_EZH2-low	hg19	https://ftp.	H3K27ac		GSM1513835	WIG
GSE71809	GSE71809.H3K27ac.AML_pat8	hg19	https://ftp.	H3K27ac	**AML**	GSM2152602	BED
GSE71809	GSE71809.H3K27ac.AML_pat10	hg19	https://ftp.	H3K27ac		GSM2152605	BED
GSE71809	GSE71809.H3K27ac.AML_pat11	hg19	https://ftp.	H3K27ac		GSM2152609	BED
GSE71809	GSE71809.input.AML_pat8	hg19	https://ftp.	H3K27ac		GSM1846177	BED
GSE71809	GSE71809.input.AML_pat11	hg19	https://ftp.	H3K27ac		GSM1846188	BED
GSE71809	GSE71809.input.AML_pat10	hg19	https://ftp.	H3K27ac		GSM1846207	BED
GSE79899	GSE79899.H3K27ac.MV4-11	hg19	https://ftp.	H3K27ac	**MV4-11**	GSM2108039	WIG
GSE79899	GSE79899.H3K27ac.THP-1	hg19	https://ftp.	H3K27ac	**THP-1**	GSM2108046	WIG
GSE89336	GSE89336.H3K27ac.MLL-AF9-pos_pat5	hg19	https://ftp.	H3K27ac	**MLL-AF9**	GSM2366242	WIG

FIGURE 4.1. – *Liste des données ChIP-seq d'histone H3K27ac.*

Les pics ChIP-seq d'histone H3K4me3 permettent d'identifier les régions promoteurs [20].

accession	Name	Assembly	LINK	target	biotype	experiment_id	FORMAT
GSE54580	GSE54580.H3K4me3.PBSC_pat1	hg19	ftp://ftp.nc	H3K4me3		GSM1612049	WIG
GSE54580	GSE54580.H3K4me3.PBSC_pat2	hg19	https://ftp.	H3K4me3		GSM1612062	WIG
GSE54580	GSE54580.H3K4me3.AML_pat6	hg19	https://ftp.	H3K4me3	**AML**	GSM1612063	WIG
GSE54580	GSE54580.H3K4me3.AML_pat1	hg19	ftp://ftp.nc	H3K4me3		GSM1612068	WIG
GSE54580	GSE54580.H3K4me3.AML_pat5	hg19	https://ftp.	H3K4me3		GSM1612070	WIG
GSE54580	GSE54580.H3K4me3.AML_pat4	hg19	ftp://ftp.nc	H3K4me3		GSM1612071	WIG
GSE54580	GSE54580.H3K4me3.AML_pat3	hg19	https://ftp.	H3K4me3		GSM1612073	WIG
GSE54580	GSE54580.H3K4me3.AML_pat2	hg19	https://ftp.	H3K4me3		GSM1612075	WIG
GSE61785	GSE61785.H3K4me3.MV4-11_EZH2-high	hg19	https://ftp.	H3K4me3	**MV4-11**	GSM1513832	WIG
GSE61785	GSE61785.H3K4me3.MV4-11_EZH2-low	hg19	https://ftp.	H3K4me3		GSM1513833	WIG
GSE61785	GSE61785.IgG.MV4-11_EZH2-high	hg19	https://ftp.	IgG		GSM1513834	WIG
GSE61785	GSE61785.IgG.MV4-11_EZH2-low	hg19	https://ftp.	IgG		GSM1513835	WIG
GSE79899	GSE79899.H3K4me3.MV4-11	hg19	https://ftp.	H3K4me3	**MV4-11**	GSM2108040	WIG
GSE79899	GSE79899.H3K4me3.THP-1	hg19	https://ftp.	H3K4me3	**THP-1**	GSM2108047	WIG
GSE89336	GSE89336.H3K4me3.MLL-AF9-pos_pat3	hg19	https://ftp.	H3K4me3	**MLL-AF9**	GSM2366240	WIG
GSE89336	GSE89336.H3K4me3.MLL-AF9-pos_pat4	hg19	https://ftp.	H3K4me3		GSM2366241	WIG
GSE89336	GSE89336.H3K4me3.MLL-AF9-pos_pat5	hg19	https://ftp.	H3K4me3		GSM2366243	WIG
GSE89336	GSE89336.H3K4me3.MLL-AF9-pos_pat6	hg19	https://ftp.	H3K4me3		GSM2366245	WIG
GSE89336	GSE89336.H3K4me3.MLL-AF9-pos_pat7	hg19	https://ftp.	H3K4me3		GSM2366246	WIG
GSE89336	GSE89336.H3K4me3.MLL-AF9-pos_pat8	hg19	https://ftp.	H3K4me3		GSM2366247	WIG
GSE89336	GSE89336.H3K4me3.MLL-AF9-pos_pat9	hg19	ftp://ftp.nc	H3K4me3		GSM2366248	WIG

FIGURE 4.2. – *Liste des données ChIP-seq d'histone H3K4me3.*

4.2.2. ATAC-seq

Nous avons également intégré dans le trackhub des données d'ATAC-seq afin de détecter les régions de la chromatine accessible [40]. Les données GEO proviennent d'une expérience faite sur hg19 que nous avons lifté grâce à l'outil LiftOver de kentutils [150].

accession	NAME	ASSEMBLY	LINK	biotype	experiment_id	FORMAT
GSE107072	GSE107072.MEF2C.AML_S222A-induced_rep1	hg19	ftp://ftp.ncbi.r	AML	GSM2860578	NARROWPEAK/BIGWIG
GSE107072	GSE107072.MEF2C.AML_S222A-induced_rep2	hg19	ftp://ftp.ncbi.r		GSM2860579	NARROWPEAK/BIGWIG
GSE107072	GSE107072.MEF2C.AML_S222A-induced_rep3	hg19	ftp://ftp.ncbi.r		GSM2860580	NARROWPEAK/BIGWIG
GSE107072	GSE107072.MEF2C.AML_S222A-uninduced_rep1	hg19	https://ftp.ncb		GSM2860581	NARROWPEAK/BIGWIG
GSE107072	GSE107072.MEF2C.AML_S222A-uninduced_rep2	hg19	ftp://ftp.ncbi.r		GSM2860582	NARROWPEAK/BIGWIG
GSE107072	GSE107072.ME2C.AML_S222A-uninduced_rep3	hg19	ftp://ftp.ncbi.r		GSM2860583	NARROWPEAK/BIGWIG
GSE107072	GSE107072.MEF2C.AML_control-induced_rep1	hg19	https://ftp.ncb		GSM2860584	NARROWPEAK/BIGWIG
GSE107072	GSE107072.MEF2C.AML_control-induced_rep2	hg19	ftp://ftp.ncbi.r		GSM2860585	NARROWPEAK/BIGWIG
GSE107072	GSE107072.MEF2C.AML_control-induced_rep3	hg19	https://ftp.ncb		GSM2860586	NARROWPEAK/BIGWIG
GSE107072	GSE107072.MEF2C.AML_control-uninduced_rep1	hg19	https://ftp.ncb		GSM2860587	NARROWPEAK/BIGWIG
GSE107072	GSE107072.MEF2C.AML_control-uninduced_rep2	hg19	ftp://ftp.ncbi.r		GSM2860588	NARROWPEAK/BIGWIG
GSE107072	GSE107072.ME2C.AML_control-uninduced_rep3	hg19	ftp://ftp.ncbi.r		GSM2860589	NARROWPEAK/BIGWIG

FIGURE 4.3. – *Liste des données ATAC-seq.*

4.2.3. DNase-seq

Les données DNase intégrées dans le trackhub proviennent d'une expérience faite sur des patients sains et des patients atteints d'AML. Ces données permettent également d'identifier les régions de chromatines ouvertes et donc potentiellement actives[36, 122].

accession	sample_title	Name	assemblage	Link	biotype	experiment_id	FORMAT
GSE64864	CD14_PBSC_DNAse-Seq	GSE64864.CD14_PBSC	hg18	https://ftp.	CD14-pos	GSM1581819	WIG
GSE64864	CD34_PBSC_DNAse-Seq	GSE64864.CD34_PBSC	hg18	https://ftp.	CD34-pos	GSM1581820	WIG
GSE64864	FLT3_ITD3_DNAse-Seq	GSE64864.FLT3-ITD.AML_3	hg18	https://ftp.	AML	GSM1581827	WIG
GSE64864	FLT3_ITD4_DNAse-Seq	GSE64864.FLT3-ITD.AML_4	hg18	https://ftp.		GSM1581828	WIG
GSE64864	FLT3_WT1_DNAse-Seq	GSE64864.FLT3.AML_1	hg18	https://ftp.		GSM1581831	WIG
GSE64864	FLT3_WT2_DNAse-Seq	GSE64864.FLT3.AML_2	hg18	https://ftp.		GSM1581832	WIG
GSE64864	FLT3_WT3_DNAse-Seq	GSE64864.FLT3.AML_3	hg18	ftp://ftp.nc		GSM1581833	WIG
GSE64864	FLT3_WT4_DNAse-Seq	GSE64864.FLT3.AML_4	hg18	ftp://ftp.nc		GSM1581834	WIG
GSE64864	FLT3_WT5_DNAse-Seq	GSE64864.FLT3.AML_5	hg18	https://ftp.		GSM1581835	WIG
GSE64864	FLT3_WT8_DNAse-Seq	GSE64864.FLT3.AML_8	hg18	https://ftp.		GSM1581836	WIG

FIGURE 4.4. – *Liste des données DNase-seq d'AML.*

4.2.4. Données Hi-C

Les données de Hi-C ont été collectées à partir de UCSC sur des données d'AML. Ces données étaient disponibles uniquement pour l'assemblage hg19, nous avons donc fait un lift des données de hg19 à hg38 pour l'intégrer au trackhub.
Cela nous a permis de collecter les coordonnées des bordures TADs dans la région d'intérêt. Ces bordures nous ont permis dans la première version du trackhub de collecter la limite des régions interagissant entre elles. Dans la deuxième version du trackhub nous avons décidé d'aller plus loin que les limites TADs pour sélectionner les régions d'intérêt.

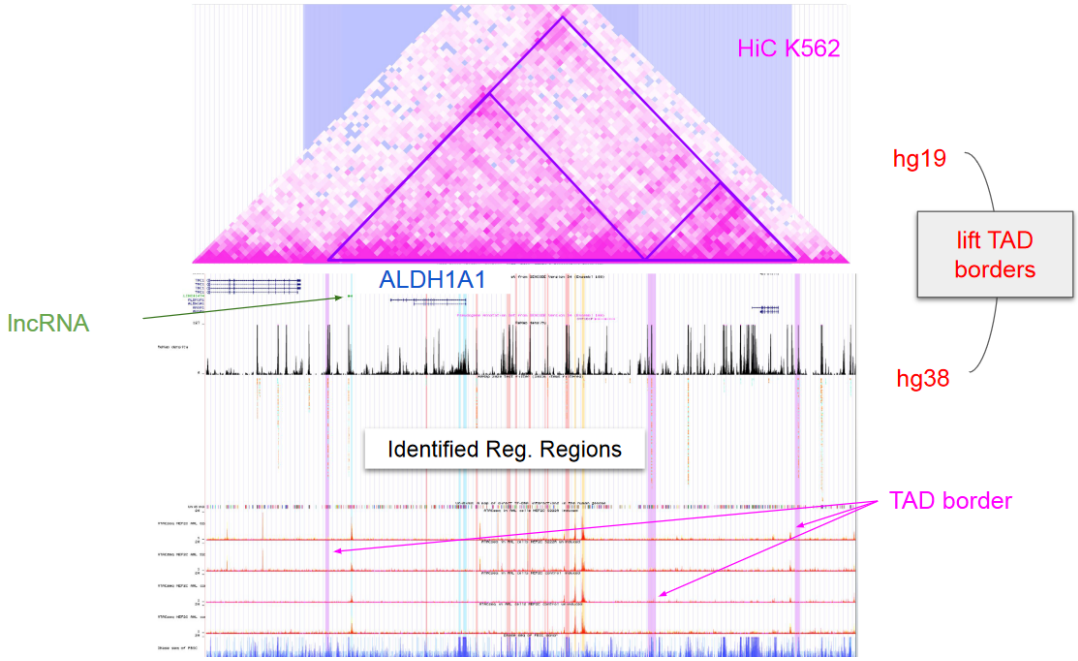


FIGURE 4.5. – ***Données Hi-C d'AML.*** *Ces données ont permis de détecter les bordures TADs autour du gène ALDH1A1. Ces données étant en hg19 nous avons liftés les cooredonnées des bordures en hg38.*

4.3. Construction d'un trackhub

Un trackhub est une plateforme en ligne qui permet de visualiser des données génomiques de plusieurs types sur un génome de référence. Un trackhub est construit en utilisant un logiciel de visualisation des données génomiques tel que le navigateur de genome UCSC.

Dans le contexte de ma thèse, impliquant une collaboration avec le laboratoire Advanced BioDesign, il est important d'utiliser un trackhub pour partager les données multi-omiques portant sur la régulation du gène ALD1HA1 dans les AML. Ce trackhub permet de centraliser les données multi-omiques (ReMap, DNase, ATAC, ChIP-seq d'histones) à partir de datasets GEO des tissus AML et de les rendre accessibles à nos collaborateurs. Nos collaborateurs peuvent ainsi visualiser les régions régulatrices du génome actives autour du gène ALDH1A1, ainsi que les facteurs de transcription qui régulent potentiellement ce gène, ce qui pourrait mener à l'identification d'une nouvelle cible thérapeutique pour le traitement de l'AML. De plus, cela garantit la transparence et la qualité des données en les conservant à un endroit centralisé. Cela peut également aider à éviter les erreurs et les incohérences dans les données, car tout le monde travaille à partir de la même source de données.

Pour construire notre trackhub multi omique nous nous sommes placées sur la région génomique autour du gène ALDH1A1. Nous avons ensuite superposé toutes les données multi-omiques nécessaires à l'analyse. La première piste représente l'annotation GENCODE des gènes afin de visualiser la position du gène ALDH1A1 ainsi que des gènes autour comme le lncRNA présent en amont du gène ALDH1A1. ReMap est en deuxième position dans le trackhub afin de visualiser les régions régulatrices du génome. Ensuite, nous avons ajouté les pistes d'ATAC-seq, DNAse-seq, et de ChIP-seq d'histones. Les données ReMap ont été filtrées afin d'afficher les pics provenant de lignées cellulaires AML. Après avoir collecté ces données dans le trackhub nous avons identifié des régions du génome potentiellement régulatrice du gène ALDH1A1.
Nous avons établi un code couleur permettant d'identifier les différentes régions car ces données sont ensuite partagées avec ABD afin de procéder aux validations expérimentales, il nous permet donc de différencier chaque type de données (ATAC-seq, DNase-seq. . .) ainsi que les catégories de régions (Enhancers ou promoteurs).

Nous sélectionnons des régions présentant un signal dans ReMap mais également dans les autres tracks. Nous émettons l'hypothèse que si la région présente un signal de Chip-Seq, de DNase-seq, et de ChIP-seq d'histones, il y a de fortes chances que le signal n'est pas dû au hasard et que c'est une région potentiellement active sur laquelle se fixent des TFs. Ce qui pourrait donc indiquer que la région est régulatrice du gène ALDH1A1. Les régions ont été sélectionnées entre les limites TAD délimitées par l'expérience Hi-C (Figure 4.6). En délimitant ces bordures, il est possible de localiser les régions régulatrices potentielles qui sont situées à l'intérieur du TAD contenant le gène ALDH1A1. Ces régions sont plus susceptibles d'interagir avec le gène ALDH1A1, ce qui en fait des cibles potentielles pour la régulation de son expression.

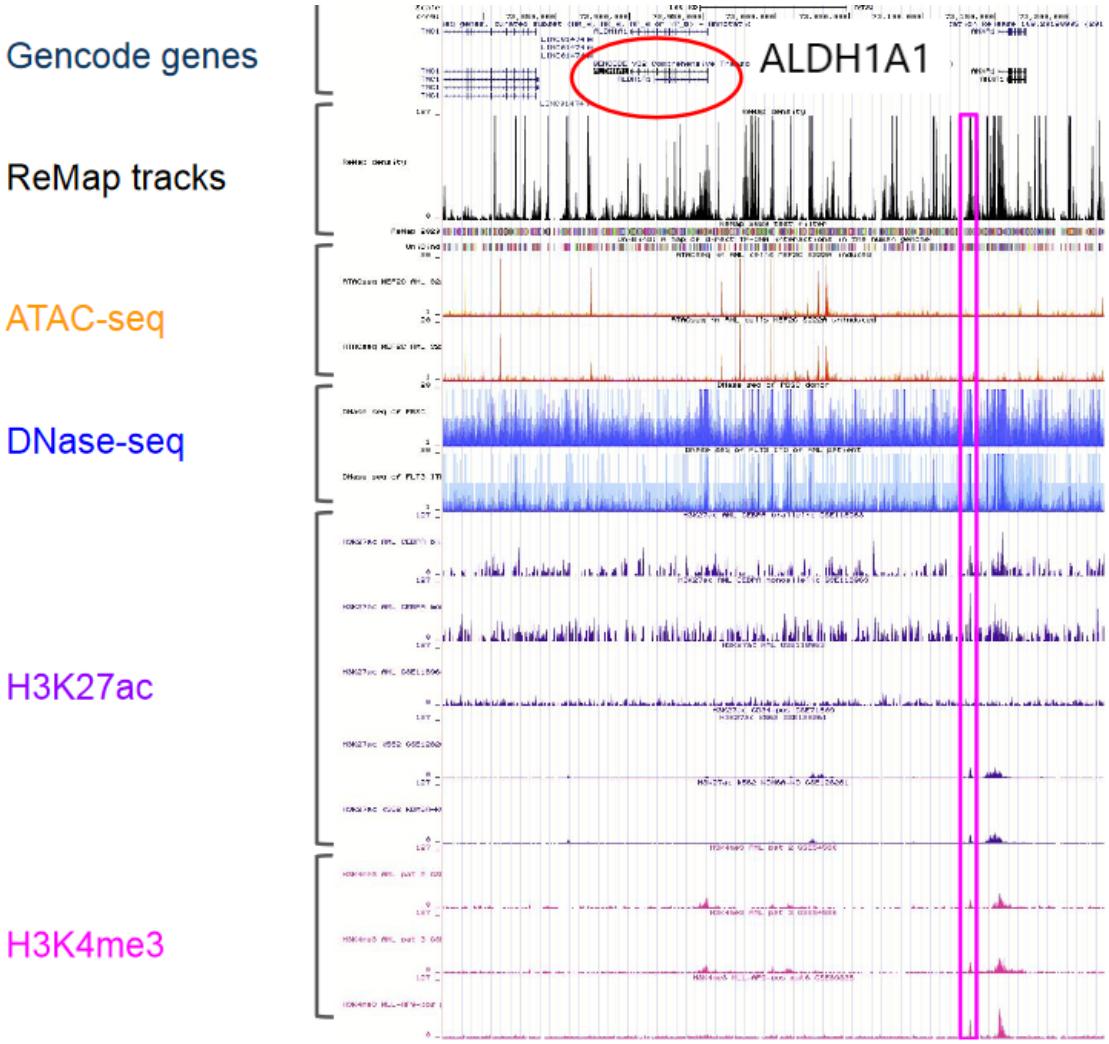


FIGURE 4.6. – ***Trackhub multi omique des AML.*** *Le gène ALDH1A1 est entouré en rouge. Encadré en violet, un exemple de région potentiellement régulatrice car elle présente un signal sur la piste ChIP-seq de ReMap, la DNase-seq (en bleu) et les ChIP-seq d'histone (en violet et en rose).*

A l'aide de ce trackhub nous avons d'abord identifié deux régions potentiellement promotrice du ALDH1A1 (Figure 4.7). Ces régions ont été identifiées à l'aide des pics ReMap mais également les pics ChIP-seq de H3K4me3 qui sont des marqueurs de régions promotrices. Elles sont également potentiellement situées dans une zone de chromatine ouverte car elles présentent un signal DNase-seq. Nous supposons que la région potentiellement promotrice située dans le gène ALDH1A1 pourrait-être un promoteur alternatif. Nous avons également marqué le promoteur d'un lncRNA à proximité du gène ALDH1A1 car il pourrait agir en tant que promoteur bidirectionnel et co-réguler ces deux gènes. Ces régions ont été marqué comme promotrice car elles présentent une fixation de TFs (pics ReMap) et un signal de ChIP-seq H3K4me3.

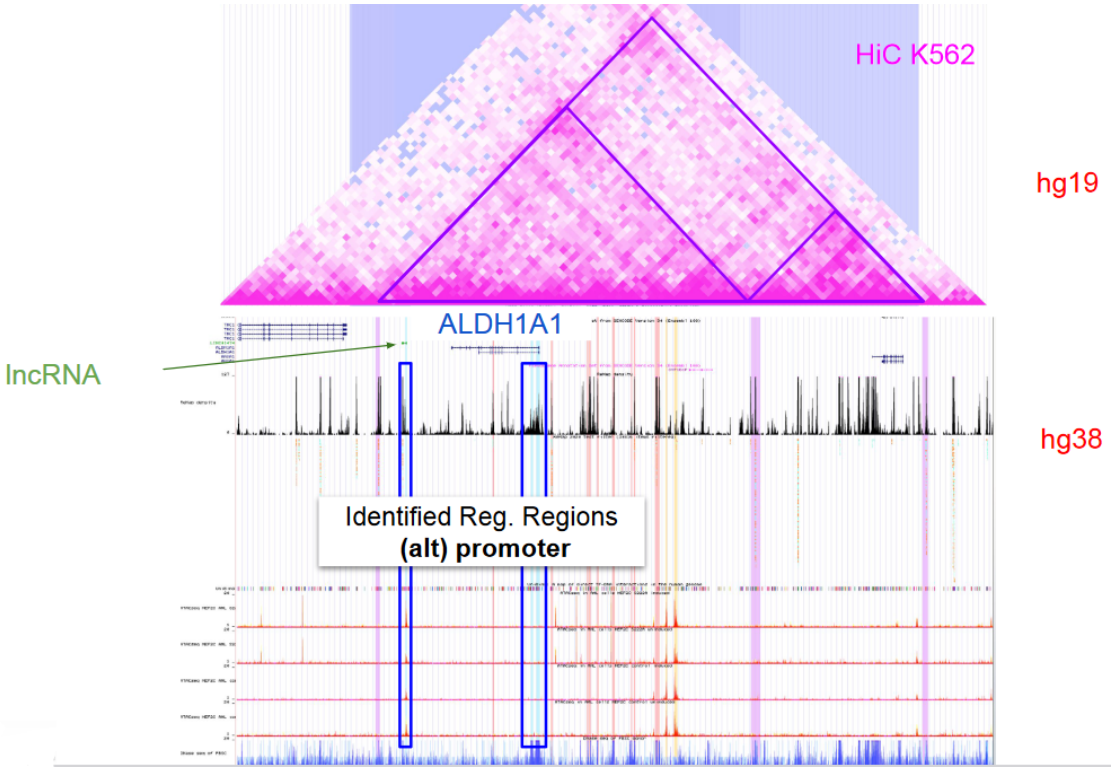


FIGURE 4.7. – *Régions promoteurs du trackhub multi omique des AML.*

Nous avons également identifié une région présentant un signal ReMap et un signal ATAC-seq fort. Cette région paraissait intéressantes car elle représente une région de chromatine ouverte, qui suggère que la région de fixation est bien active et donc potentiellement régulatrice.

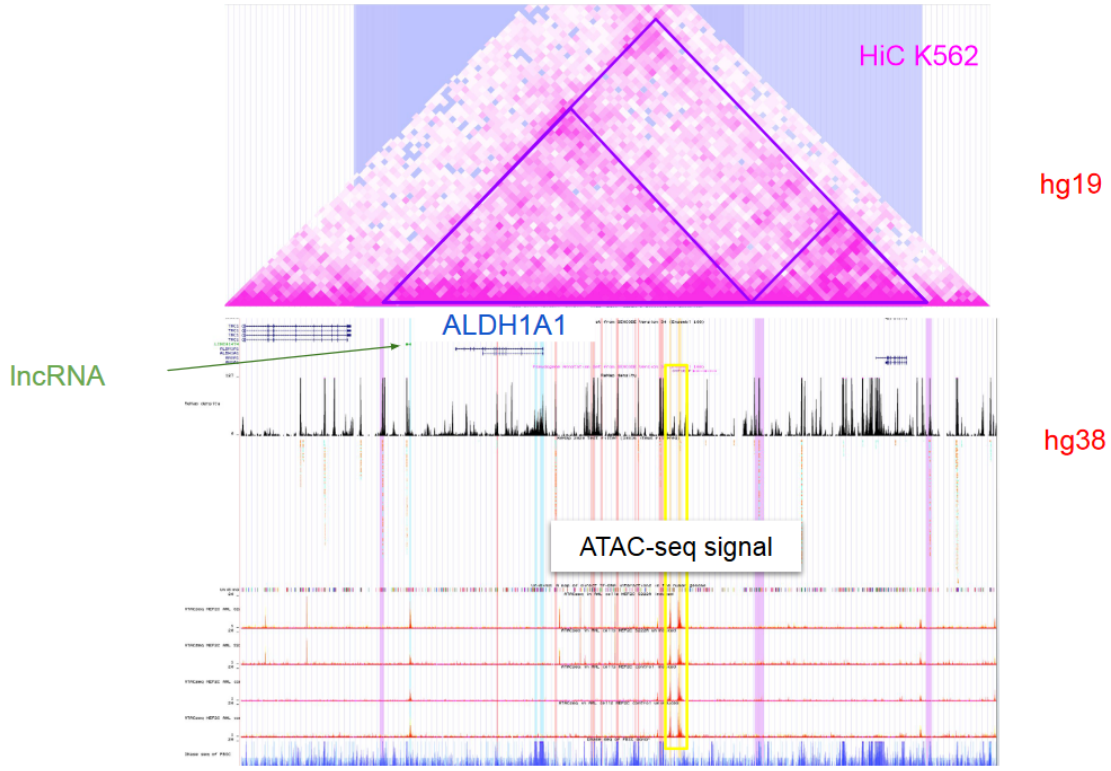


FIGURE 4.8. – ***Régions avec un signal ATAC-seq dans les AML.*** *Cet région est encadré en jaune. Dans cette région on observe à la fois un fort pic ReMap et du signal sur toutes les pistes ATAC-seq (pistes orange).*

Enfin les régions les plus importantes sont les régions enhancer qui présentent notamment un signal fort de ChIP-seq ReMap, ces régions sont donc fixées par des facteurs de transcription et donc potentiellement fonctionnelles. Nous avons également observé des pics ChIP-seq de H3K27ac, qui sont des marqueurs de régions enhancers. Pour la plupart de ces régions on retrouve également des signaux ATAC-seq et DNase-seq qui suggèrent que la chromatine est ouverte et donc que ces régions sont actives. Ces régions se situent dans les régions intergéniques. Elles sont au nombre de 11.

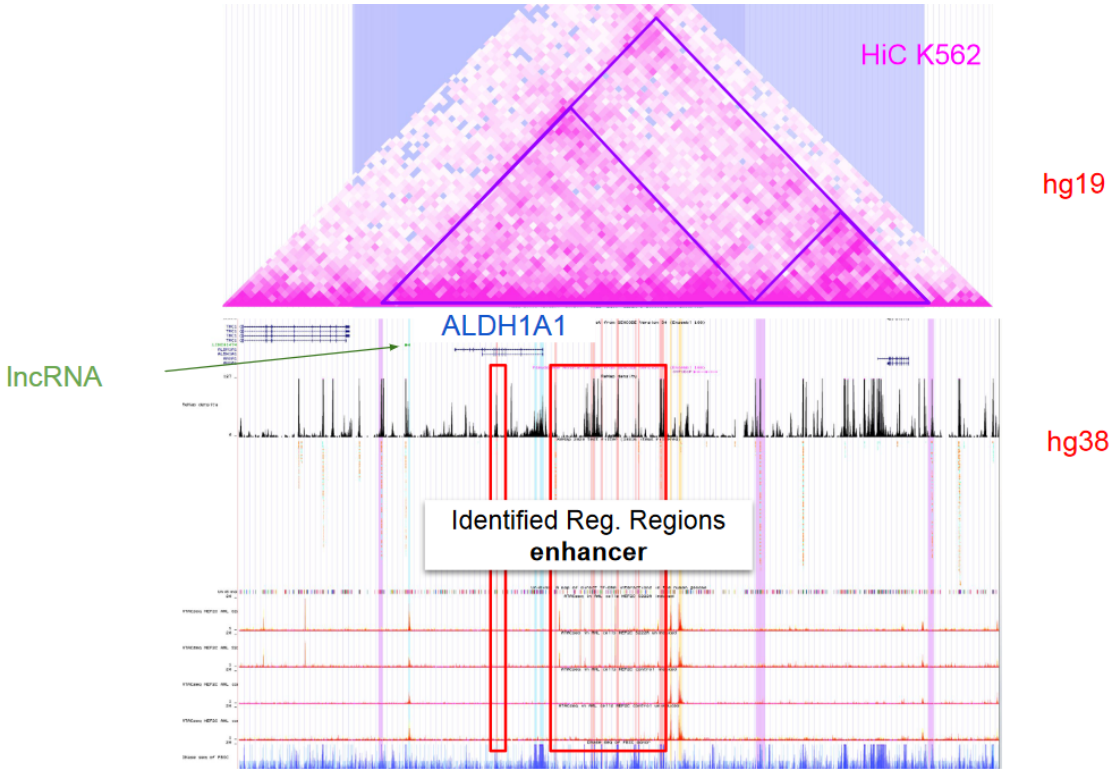


FIGURE 4.9. – ***Régions potentielemnt enhancer des AML.*** *Ces régions sont encadrés en rouge. Ces régions présentent du signal ChIP-seq histone et également des pics de fixation ReMap.*

En conclusion, nous avons identifié 14 régions qui pourraient agir en tant qu'enhancers dont deux régions qui pourraient agir en tant que promoteurs autour du gène ALDH1A1. Nous avons communiqué un trackhub multi-omique à nos collaborateurs afin qu'ils procèdent aux validations expérimentales. Les validations ont été réalisé par notre collaboratrice Yasmine LABIAD, qui a éffectué des tests pour confirmer les résultats obtenus à partir de ce trackhub multi-omique. Ces expériences seront détaillées dans la section suivante.

4.4. Validations expérimentales

4.4.1. Contexte

Dans les AML, les ALDH sont dérégulé dans 42% des cas entrainant négativement la survie des patients. Le laboratoire ABD a donc dévelopé un traitement (le DIMATE) qui inhibe ces protéines. En complément de ce traitement nos collaborateurs cherchent à indentifier le régulateur des ALDH. Pour ce faire, nous avons cartographié les régions régulatrices autour du gène ALDH1A1 à l'aide de données AML multi-omique.

Ces données nous ont permis d'identifier 14 régions potentielemnt régulatrices de gène ALDH1A1. Ces Résultats ont ensuite été communiqué à nos collaborateurs. Dans la suite de cette partie, nous présenterons les étapes de la validation experimentale. Ce travail a été éffectué par notre collaboratrice d'ABD, Yasmine LABIAD.

4.4.2. Matériels et Méthodes.

4.4.2.1. Screening de l'expression des ALDH

Culture cellulaire

Un panel de 10 lignées cellulaires a été utilisé pour réaliser le screening de l'expression des gènes et des protéines des 19 isoformes d'ALDH. Les cellules HL60, THP-1, KG1, K562, Kasumi-1, Kasumi-3, OCI-AML3 et MOLM-14 ont été achetées auprès de DSMZ; Monomac-6 et MV4.11 ont été aimablement fournies par le Dr Meritxel Alberich, IMG, Prague. Toutes les lignées cellulaires, à l'exception de OCI-AML3, toutes les lignées cellulaires ont été cultivées à 37°C, 5% de CO2 en utilisant du RPMI1640 (HYCLSH30096.01, Hyclone) complété par 2 mM de L-Glutamine (X0550-100, VWR), 1x de Pénicilline/Streptomycine (K952-100, VWR), et 10% de Sérum de veau fœtal - SVF (VWR) (20% pour Kasumi-1 et Kasumi-3). Les cellules OCI-AML3 ont été cultivées à 37°C, 5% de CO2 en utilisant du DMEM (11500416, Fisher Scientific) complété par 2 mM de L-Glutamine (X0550-100, VWR), 1x Pénicilline/Streptomycine (K952-100, VWR), et 20% de SVF.

Screening de l'expression des gènes de l'ALDH par RT qPCR

Enfin d'évaluer les niveaux d'expression des 19 isoformes de l'ALDH, la qPCR a été réalisée en tilisant les paires d'amorces correspondantes au tableau ainsi que le PerfeCTa SYBR®Green fastmix (733-1386, VWR). La qPCR a été réalisée via le CFx96 (Biorad) en utilisant le protocole suivant : 30 secondes à 95°C, 5 secondes à 95°C, 15 secondes à 60°C (45 cycles) puis programme de melt curve du CFx96.

TABLEAU 4.2. – *List des séquences primer pour chaque isoforme de l'ALDH*

ALDH Isoforms	**Primer Sequences**
ALDH1A1-Forward	TTGGAAATCCTCTGACCCCA
ALDH1A1-Reverse	CCTTCTTTCTTCCCACTCTC
ALDH1A2-Forward	CATTGGAGTGTGTGGACAGA
ALDH1A2-Reverse	GGAGCTATTTTCCAGGCA
ALDH1A3-Forward	TTTTCATCGACCTGGAGG
ALDH1A3-Reverse	GACGTTGTCATCTGTGGG
ALDH1B1-Forward	ACTTGGCCTCACTCGAGA
ALDH1B1-Reverse	CCAGCAAAGTACCGATAC
ALDH1L1-Forward	ACACAGTGGTGATCAAGC
ALDH1L1-Reverse	GCTCTGCAAACTTCAAGG
ALDH1L2-Forward	TTATCACCCATCCATCCTGC
ALDH1L2-Reverse	CCCAGCTTTCTTATCTCCCA
ALDH2-Forward	GTCAGATGCCGATATGGAT
ALDH2-Reverse	GCCCTGGTTGAAGAACAG
ALDH3A1-Forward	CACATCACCTTGCACTCTCT
ALDH3A1-Reverse	AGCTCTTCTTGCCATGGT
ALDH3A2-Forward	TAGCTTTTGGTGGGGAGA
ALDH3A2-Reverse	CTTGCATCACCTTGGTTT
ALDH3B1-Forward	TATCTAATCACGGGCCAC
ALDH3B1-Reverse	AGCTGCTTGTTTTCTTGC
ALDH3B2-Forward	TTCTCCAACAGCAGCCAG
ALDH3B2-Reverse	CGGACAGCAGAGATATGTAG
ALDH4A1-Forward	TCTCGCCCTTTAACTTCACT
ALDH4A1-Reverse	CTTCCATAGGACCACGTTGC
ALDH5A1-Forward	AGGATGACCTTGCCAGAA
ALDH5A1-Reverse	GAGAACCACTCTAGGAAAAAG
ALDH6A1-Forward	CCCCTGATGGAACATTAAACA
ALDH6A1-Reverse	TGCTTTGATGTCCGGATG
ALDH7A1-Forward	GACCTATTGCCCTGCTAA
ALDH7A1-Reverse	CCATGCTTCTCTTGCTTTC
ALDH8A1-Forward	CCACTCTACTTGCTGACCT
ALDH8A1-Reverse	GTTTGCACAACATCCACG
ALDH9A1-Forward	GCTGAGGTTCTAGAAAGAGC
ALDH9A1-Reverse	TATGAGCCCGTTGGATGT
ALDH16A1-Forward	CTATGTGAATGGGAAGTGGT
ALDH16A1-Reverse	AAGTTCTCTCCTGTGATGGG
ALDH18A1-Forward	GCTGATGGCCTTGTATGA
ALDH18A1-Reverse	GCTTCTGCTCATCATGGA

Screening de l'expression des gènes de l'ALDH par RT qPCR

Un million de cellules a été utilisé pour extraire les protéines. Les cellules ont été collectées par centrifugation à 600g, pendant 5 minutes. Les culots cellulaires ont été lavés 3 fois par centrifugation à 600g, à+4°C, pendant 5 minutes avec du PBS 1x froid. A partir du culot cellulaire lavé, les protéines ont été extraites en utilisant le tampon RIPA (50 mM Tris-HCl pH7, 1% NP-40, 0,5% Na-deoxycholate, 0,1% SDS, 150mM NaCl et 2mM EDTA) contenant 1/200 d'inhibiteur de protéase (539134-1ML, Millipore Sigma) et 1/100 d'inhibiteur de phosphatase (524625-1SET, Millipore Sigma). Les cellules ont été incubées dans le tampon RIPA pendant 30 minutes sur la glace, et centrifugées 10 minutes à 10 000 g, à +4°C. Les protéines ont été récupérées dans le surnageant et quantifiées par la méthode de Lowry (Pierce™ BCA® Protein Assay, 23225, Thermo Fisher Scientific). L'analyse Western Blot a été réalisée sur un système WES (ProteinSimple) selon les instructions du fabricant et en utilisant le module de séparation 12-230 kDa (ProteinSimple, SM-W004) et le module de détection Assay Module Anti-rabbit HRP (ProteinSimple, DM-001). La liste des anticorps utilisés est résumée dans le tableau suivant.

Antibody	Reference	Ab Dilution	Specificity	Secondary Antibody	Protein concentration (3 µL loaded)
Anti- ALDH1A1	15910-1-AP, Proteintech	1/5	Human	Anti-rabbit (Wes Detection Kit)	1 mg/mL
Anti- ALDH1A2	13951-1-AP, Proteintech	1/50	Human	Anti-rabbit (Wes Detection Kit)	0.2 mg/mL
Anti- ALDH1A3	ABGEAP7847A, VWR	1/5	Human	Anti-rabbit (Wes Detection Kit)	1 mg/mL
Anti- ALDH1B1	15560-1-AP, Proteintech	1/500	Human	Anti-rabbit (Wes Detection Kit)	0.5 mg/mL
Anti- ALDH2	15310-1-AP, Proteintech	1/50	Human	Anti-rabbit (Wes Detection Kit)	0.5 mg/mL
Anti- ALDH3A1	15578-1-AP, Proteintech	1/50	Human	Anti-rabbit (Wes Detection Kit)	0.2 mg/mL
Anti- ALDH3A2	15090-1-AP, Proteintech	1/10	Human	Anti-rabbit (Wes Detection Kit)	0.2 mg/mL
Anti- ALDH3B1	ab236673, Abcam	1/10	Human	Anti-rabbit (Wes Detection Kit)	0.5 mg/mL
Anti- ALDH3B2	15746-1-AP, Proteintech	1/10	Human	Anti-rabbit (Wes Detection Kit)	0.5 mg/mL
System Control	042-196, Protein Simple	1/10	Rabbit	Anti-rabbit (Wes Detection Kit)	0.5/1 mg/mL

FIGURE 4.10. – *Tableau avec les référence des anticorps utilisé lors du screening.*

4.4.2.2. Identification des facteurs de transcription

En partant de l'hypothèse suivante : « Etant donné que l'ALDH1A1 est un marqueur de la AML, ils se pourraient que les facteurs de transcriptions importants et/ou dérégulés dans AML réguleraient potentiellement l'expression de l'ALDH1A » Nous avons réalisé une bibliographie qui décrivaient l'implication, importance, dérégulation des facteurs de transcription dans la leucémie. Nous avons alors établi une short liste de 14 TFs : **GATA2, CEBPB, BRD4, EP300, ERG, IKZF1, TAL1, FOXM1, CEBPA, SPIA (PU.1), RUNX1, c-MYC, c-MYB, NRF2.**

4.4.2.3. Culture des lignées cellulaires

K562, KG1 et Kasumi-4 ont été cultivées s à 37°C, 5% de CO2 dans un milieu RPMI 1640 $GlutaMAX^{T}M$ Supplement ($Gibco^{T}M$), et 10% Serum de Veau Fœtal. Un culot de chaque lignée et trois répliquats techniques ont été collectés, lavés 3 fois avec du PBS à 1 100 rpm pendant 5min et stockés en attendant l'extraction de l'ARN.

4.4.2.4. Extraction de l'ARN, quantification et contrôles de qualité

L'ARN total a été extrait à l'aide du kit RNeasy Mini (Qiagen, France) selon les instructions du fabricant. Les ARN ont été quantifiés en utilisant le spectrophotomètre NanoDrop (Thermoscientific, France) et leur qualité a été évaluée avec le bioanalyseur Agilent 2100 (Agilent Technologies).

4.4.2.5. qRT-PCR

En partant de l'hypothèse suivante : « Si un facteur de transcription régulerait potentiellement l'ALDH1A1, son expression doit être corrélée à l'ALDH1A1 » .

La qRT PCR de l'expression basale des 14 TFs a été réalisée en utilisant le protocole SuperScript VILO MasterMix décrit par Invitrogen. Un microgramme d'ARN de chaque échantillon a été rétrotranscrit en utilisant 4 μL de SuperScript VILO MasterMix et de l'eau RNase DNase free pour un volume final de 20 μL. Le mélange a été incubé pendant 10 min à 25°C, 60 min à 42°C, et 5 min à 85°C.

La qRT-PCR a été réalisée avec l'appareil QuantStudio™ 6 Flex Real-Time PCR en utilisant Applied Biosystems Power SYBR Green Master Mix. Chaque réaction RT-qPCR de 20 μl comprenait 10 μl de Master Mix SYBR Green, 1 μl d'amorce F+R à 100pmol/μl de concentration, 2 μl d'ADNc dilué (1 :10) et 7 μl d'eau RNase DNase free. Les conditions de PCR consistaient en une étape de hold de 1 cycle à 95°C pendant 10 min, suivi de 40 cycles à 95°C pendant 15 s, 60°C pendant 60 s. Après la procédure d'amplification, nous avons soumis toutes les réactions de PCR à une analyse de melt curve avec une mesure de fluorescence continue de 60 °C à 95 °C pour assurer une amplification unique.

Les primers utilisés pour chaque facteur de transcription sont résumés dans le tableau suivant :

TABLEAU 4.3. – *Primers utilisés pour chaque facteur de transcription*

TF	Primers (Fw-Rv)
GATA2	CTACCACAAGATGAATGGGC
	GACAATTTGCACAACAGG
CEBPB	CAAAACTTTGGCACTGGG
	ATGTGCGGTTGGTTTGGA
BRD4	AAAGACCCGTGTAGGATG
	CAGACATGCTAGTGATCCCA
EP300	CGGCCTAAACTCTCATCT
	CGTGCTCCAAGTCAAATAG
ERG	GACAGACTTCCAAGATGAGC
	TCCTTGAGCCATTCACCT
IKZF1	GGACCATGGATGCTGATGA
	CCTCATCTGGAGTATCGCTT
TAL1	CGAAAAAGGGGGAAAGCA
	CCTGGGGCATATTTAGAGAG
FOXM1	GTGAATCTTCCTAGACCACC
	TGCCTCACCATCTGCACA
CEBPA	CCCGGCAACTCTAGTATT
	ATTTGCTCCCCCTACTCA
SPI1 (PU.1)	GTATTACCCCTATCTCAGCAG
	GCTCCGTGAAGTTGTTCT
RUNX1	TGATGAAAACTACTCGGC
	GCTTTTCCCTCTTCCACT
c-MYC	ATTCTCTGCTCTCCTCGA
	TCTTGTTCCTCCTCAGAGTC
c-MYB	ACTATGATGGGCTGCTTC
	CCACCAGCTTCTTCAGTT
NFE2	ATCCCAACCCCAGCATAGTT
	GCTCTAGAAACCTCCTCTCT

4.4.2.6. ChIP qPCR

Dans le but de confirmer implication potentielles des facteurs de transcription identifiés dans l'expression du gènes ALDH1A1 ainsi que leur fixations sur nos régions d'intérêt, nous avons réalisés une Chromatine immunoprécipitation qPCR (ChIP qPCR). Le ChIP qPCR a été réalisée en utilisant le protocole iDeal ChIP-qPCR kit de Diagenode selon le schéma suivant :

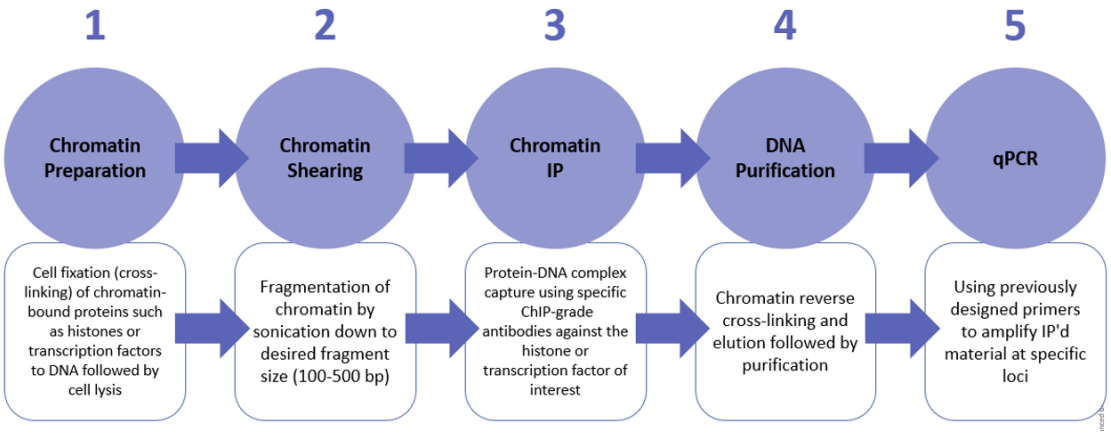


FIGURE 4.11. – *Protocole ChIP-qPCR.*

Etant donné que KG1 exprime le plus ALDH1A1 (résultats qPCR et WES), nous avons décidé de réaliser le ChIP en utilisant cette lignée cellulaire. Nous avons cultivé 300 millions de cellules en les divisant en 12 culots pour la fixation au formaldehyde avec une concentration finale de 11%. Cette étape est suivie par la lyse cellulaire et la fragmentation de la chromatine. Une étape d'optimisation a été réalisée afin d'identifier le nombre de cycle de fragmentation pour obtenir un smir de fragment d'ADN après purification entre 200 – 500 pb et cela en réalisant un gel d'agarose a 1,5%.

La 3eme étape consistait à incuber des billes magnétiques avec chacun des anticorps : c-MYB (AB226470), cMYC (AB32072), RUNX1 (AB272456), IKZF1 (AB229275) plus deux contrôles, un control négatif IgG et un contrôle positive CTCF. Le complexe (Bille + AC) a été incubé avec l'ADN fragmenté. L'immunoprécipitation a été faite en utilisant un rack magnétique.

L'étape 4 consistait à purifier le complexe ADN/TFs + bille suivie par une étape de descrosslink et purification d'ADN en utilisant des billes magnétiques. La quantification de l'ADN purifié a été réalisée par Qubit.

La dernière étape était de quantifier l'ADN par qPCR (protocole décrit au-dessus) en utilisant 2 couples de primers différents pour chaque région d'intérêt :

Primers	Seq
ALDH1A1-Pro-1-L	GCAAACCCGAGTCAAAGCAG
ALDH1A1-Pro-1-R	AATGTCATCCTCAGGCACGC
ALDH1A1-Pro-2-L	CTGGCCAGGTGTCTTCAGG
ALDH1A1-Pro-2-R	GCCTTTCTTCCCCAAACAGC
ALDH1A1-AltPro-1-L	TCACAGCCACTTTCCAAGGT
ALDH1A1-AltPro-1-R	CCAAGCTTCTCATCCCCTGG
ALDH1A1-AltPro-2-L	TCCAGCCAGATTAGCCTTTGG
ALDH1A1-AltPro-2-R	AGGTGGCCAAGAGTAGGGAA
ALDH1A1-Enh8-1-L	TGCCTGGCCTGGACATAAAG
ALDH1A1-Enh8-1-R	TTCCAAACAGTGAGGGAGGC
ALDH1A1-Enh8-2-L	CCTATCAGGGCTTGCCAGAG
ALDH1A1-Enh8-2-R	TCCCCTTGAAGATCAGCAGC
ALDH1A1-Enh7-1-L	ACAAGAAGGTGTTGGGATGGG
ALDH1A1-Enh7-1-R	CCCTGAGCTGTGTGGATCTC
ALDH1A1-Enh7-2-L	GCAAGTAGGGTGGAGTCTGC
ALDH1A1-Enh7-2-R	TCTCCCTTGCCAGATTGTGTC

FIGURE 4.12. – *Primers des régions d'intérêt.*

La séquences et coordonnées de chacune des chacune des régions testées sont résumés dans le tableau suivant :

Nom de la région	Coordonnée	Séquence
Promoteur	GRCh38:9:72951834:72951974:1	CATGTTAAAGGCACAATCTGATCTCTGCTGCTACTTTCCTGAGGAAAAGAGCCAATCAGCTGAGCAACTTCCTCATTTGTTGTGCTTTATGAAAATGAAAATGCAACAGACCTTGGGATTTCAAAGCCATCCATGGTTACA
Promoteur Alternative	GRCh38:9:72948128:72948327:1	GCCAAGGTCATATAATTAAATACATAGAATAGCTGGAAATTGAAGCCAGGTTTAGCTTGTTCAGTCACTGCACCTTGCAGCCAGTCAGCTATTTTTTTCACCTGCTTCCAAGCCCAAAGCAAACAAAATATTGCTTCAGTTGGAAGGAAGTTAAAAACATTTCCCAAGTGAAAGCAGCATCACAGCACAATCAGCTTTGC
Enhancer 7	GRCh38:9:73008039:73008195:1	AATACAGTATCTGGCCCTTGTAATAAGTACTCATATCAGTCAGGAAATCAGAAACCACAATAATTATAACAGAAAGAAGTTAGTATGATGAATAGCTAAAGTAGTAAAAAAGGAACATGAAGGGAACAGAGAAATTAACTGCAGGAAGTAGCTTTGC
Enhancer 8	GRCh38:9:73009988:73010150:1	AAAAGAAGCAAATATGTTGAAGACATAGATTTGCTCCCAAACCCCCCTAGAGGTCTGCATTGTAATCCTCCTATCAGGGCTTGCCAGAGACTCCATGTGGTGCCTCTGTATACCACAGACACTTAAGATCTTAGGAACAGCTGCTGATCTTCAAGGGGAAGTG

FIGURE 4.13. – *Séquences des régions d'intérêt.*

4.4.2.7. Luciférase assay

Dans le but de tester l'implication des régions identifiées dans la régulation de l'expression de l'ALDH1A1 nous avons réalisés un luciferase assay. Pour ce faire, 12 différentes constructions ont été designées par nos soins et synthétisés par la société GeneCust.

Une construction pour le promoteur, une construction pour le promoteur alternatif, 4 constructions sous le promoteur constitutive SV40 dans le but de valider l'activité des régions sélectionnées en forward et en reverse. 6 autres constructions ont été réalisées pour tester l'activité des deux enhancers sélectionnées (enhancer 7 et 8) ainsi que le promoteur alternatif sous le promoteur du gène ALDH1A1 toujours en forward et en reverse dans le but de valider l'implication de ces regions dans la régulation de l'ALDH1A1.

La première étape était de transformer des bactéries electrocompétantes E. coli à 2500v, incubation over night à 37°C. La production des plasmides a été réalisé dans 1ml de LB + Carbenicillin 100μg/μl, avec incubation over night à 37°C. La purification des plasmides a été réalisée selon le protocole du kit Qiagen plasmid plus midi kit. La quantification des différents plasmides a été faites par spectromètre NanoDrop. L'électroporation des plasmides (les constructions) et le plasmide renilla (contrôle pour normaliser les résultats) a été réalisé en utilisant le système de transfection Néon de ThermoFisher Scientific ainsi que du kit Neon transfection system 100μL.

Les paramètres d'électroporation sont les suivant : Pulse voltage = 1450v, Pulse Width : 10ms, pulse number : 3.

Au total, 42 échantillons ont été déposés dans des plaques 24 puits contenant 1ml de milieu RPMI 10% SVF par puit. Les plaques ont été incubées 24h à 37°C , 5% de CO2. la dernière étape était la révélation à l'aide d'un lecteur de plaque.

4.4.2.8. siRNA

Dans le but de réprimer l'expression des facteurs de transcription d'intérêt et tester ce silencing sur l'expression du gène de l'ALDH1A1 nous avons réalisés une mapie de siRNA. Trois siRNA par TF ont été commandés chez thermofisher scientific, nous les avons nommés comme suit :

- Anti-MYC : S9108 = MYC-A
- Anti-MYC S9110 = MYC-B
- Anti-MYC S910 = MYC-C

- Anti-MYB S9129 = MYB-A
- Anti-MYB S9131 = MYB-B
- Anti-MYB S9130 = MYB-C

- Anti-RUNX1 S229352 = RUNX1-A
- Anti-RUNX1 2460 =RUNX1-B
- Anti-RUNX1 229351 =RUNX1-C

Deux contrôles ont été aussi commandés : un contrôle négatif : negatif silencer select et siRNA GAPDH qui représente notre contrôle positif.

L'électroporation des siRNA dans les cellules a été réalisée selon le protocole de Neon transfection system Protocole de thermofisher scientific. Les cellules K562 ont été cultivées dans un milieu RPMI 10% SVF à 37°C, 5% de CO2. A J-1, les cellules ont été diluées à une concentration de 0,5M/ml. A J0, nous avons collecté un million de cellules par condition. Pour chaque condition, les cellules ont été re-suspendues dans 100 μl de buffer R. L'électroporation des différents siRNA a été faites en utilisant des concentrations différentes pour chaque siRNA. Au total 31 conditions ont été testées :

- Contrôle négatif, négatif : Cellules sans électroporation
- Contrôle (Véhicule) : cellules avec électroporation sans siRNA
- Contrôle négatif : Silencer select
- Contrôle positif : siRNA GAPDH
- 3X siRNA MYC : 100 – 150 – 200 pmol
- 3X siRNA MYB : 100 – 150 – 200 pmol
- 3X siRNA RUNXI : 100 – 150 – 200 pmol

Les paramètres électroporation utilisés sont les suivant : Pulse voltage = 1450v, Pulse Width : 10ms, pulse number : 3. Les échantillons électroporés ont été déposés dans des boites de Pétri contenant 5 ml de milieu RPMI 10% de SVF à 37°C. Les boites ont été incubées durant 48 et 72h. Après 48h ainsi que 72h d'incubation, les cellules ont été collectées et lavées 3 fois au PBS 1X. Les ARN ont été extraits, la rétrotranscript puis la qPCR a été réalisée .(protocole et primers utilisés : décrit au-dessus).

4.4.3. Résultats

4.4.3.1. Screening de l'expression des ALDH

Nous avons réalisé une analyse du niveau de transcription des 19 isoformes et établi un profil d'expression pour chacune des 10 lignées cellulaires. Les résultats sont résumés dans la Figure 4.14.

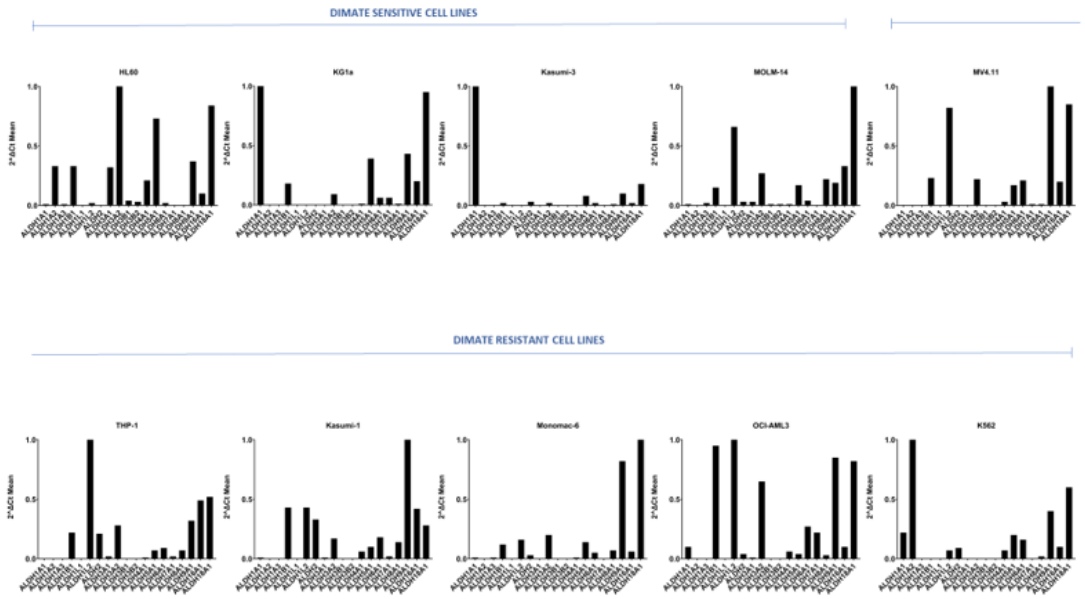


FIGURE 4.14. – ***Niveaux transcriptionnels relatifs des 19 isoformes de l'ALDH dans un panel de 10 lignées cellulaires de AML.*** *Dans les graphiques, les niveaux d'expression sont exprimés en moyenne $2^{\delta} * Ct$, par rapport au transcrit ALDH le plus abondant pour chaque lignée cellulaire individuelle.*

Nous avons réalisé une analyse westernblot à l'aide de la technique d'électrophorèse par capillaire WES dans le but d'établir le niveau protéique des 19 isoformes et établir un profil d'expression protéique pour chacune des 10 lignées cellulaires. Pour la majorité des lignées cellulaires, les niveaux d'expression protéique des isoformes ALDH1 et 3 ont confirmé le profil transcriptomique hétérogène pour ces deux classes d'ALDHs à travers les différentes lignées cellulaires (Figure 4.15). Cette analyse a montré une expression forte de l'ALDH1A1 dans les lignées HL60, KG1, Kasumi3 THP1 et K562.

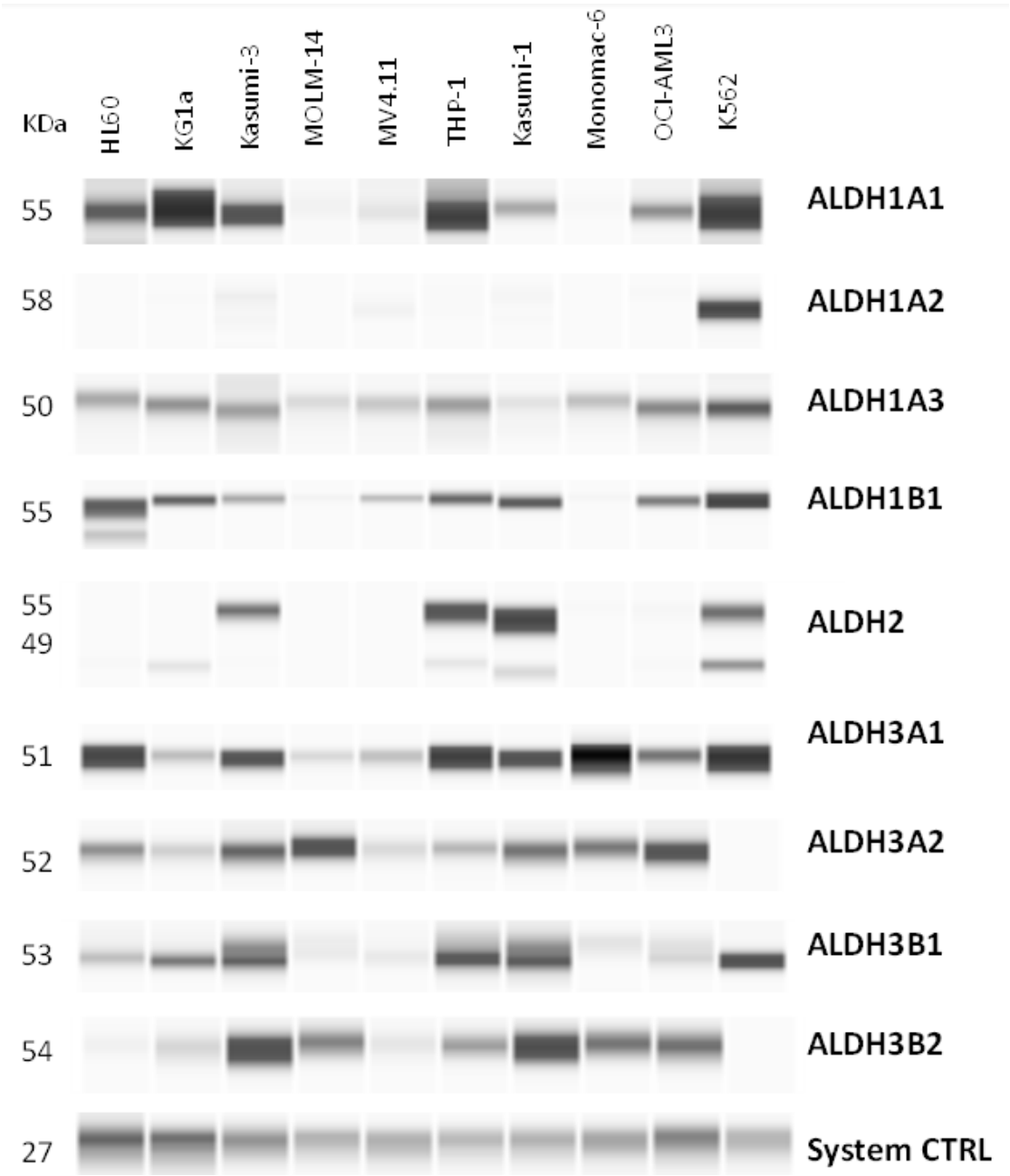


FIGURE 4.15. – *Niveaux d'expression protéique de l'ALDH de classe 1 et de classe 3 dans les 10 lignées cellulaires de AML. Ces données sont qualitatives.*

4.4.3.2. Résultat du RT qPCR

Dans le but d'identifier les facteurs de transcription parmi les 14 présélectionnés qui sont le plus exprimés dans les trois lignées KG1, Kasumi3 et K562. Nous avons réalisé une RT qPCR (Figure 4.16). Dans la lignée Kasumi3, les TFs les plus exprimés sont : RUNX1, c-MYC, c-MYB, EGR. Dans la lignée K562 les TFs plus exprimés sont : C-MYC et c-MYB. Dans la lignée K562 les TFs les plus exprimés sont RUNX1 et c-MYC.
Après le screening de l'expression basale de gène et de protéine des 19 ALDH par RT qPCR et par WES (western blot automatisé). Nous avons sélectionné KG1, K562 ainsi de Kasumi3, ces 3 lignées expriment le plus ALDH1A1.

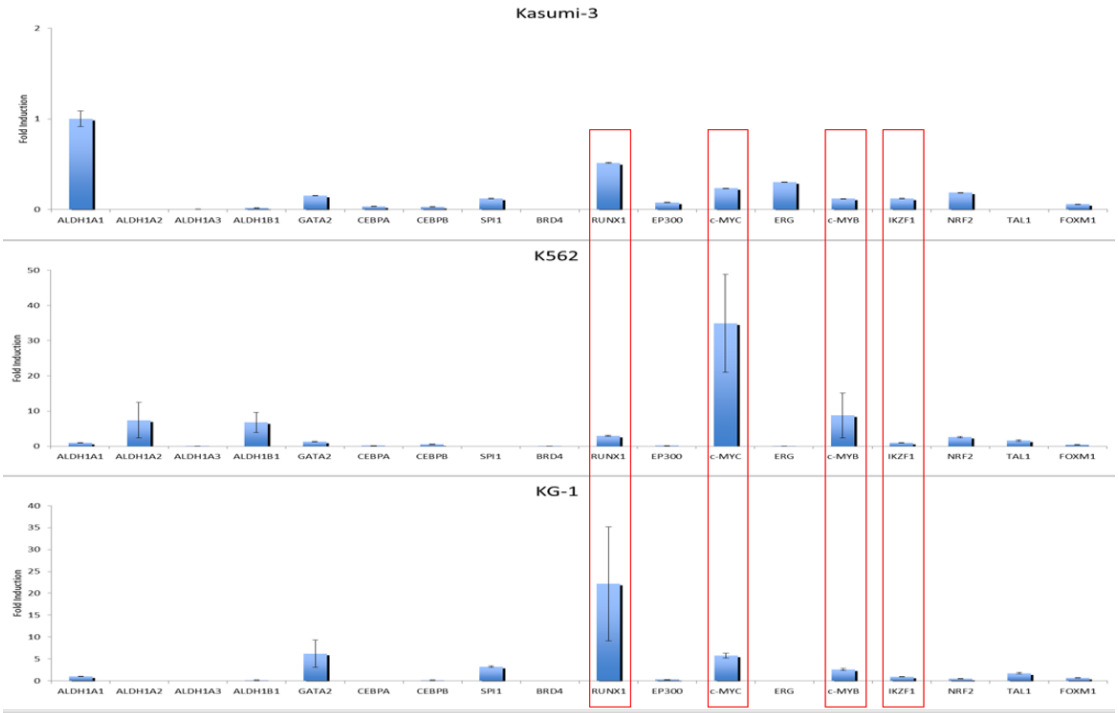
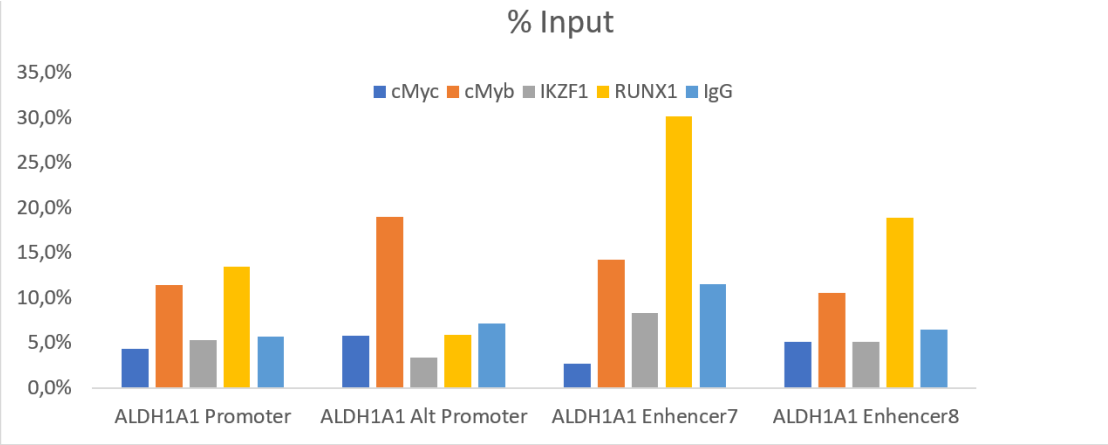


FIGURE 4.16. – ***Niveau d'expression de l'ALDH1A1 ainsi que les 14 TFs d'interet.*** *Dans les graphiques, les niveaux d'expression sont exprimés en moyenne* $2^{\delta} * Ct$*, par rapport au transcrit ALDH1A1.*

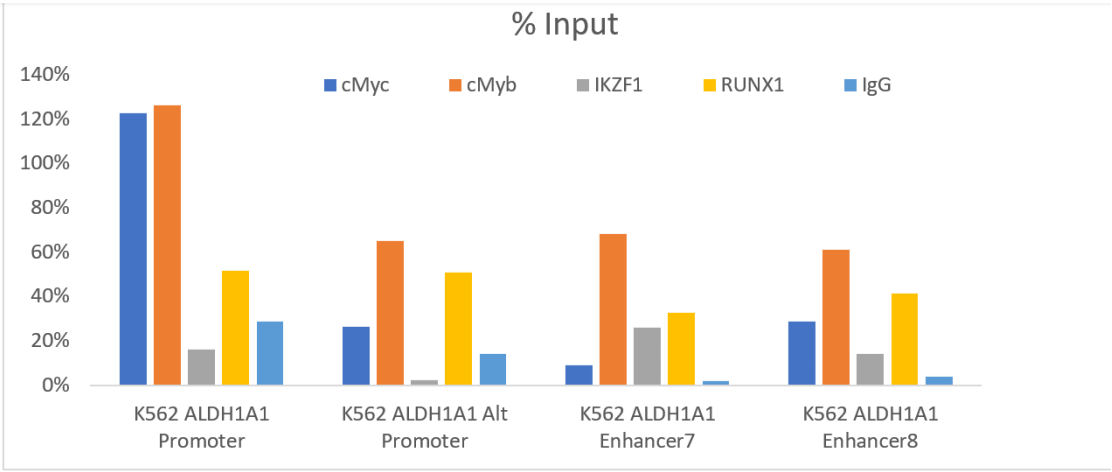
4.4.3.3. Résultat du ChIP qPCR

Les résultats du ChIP qPCR pour les lignées K562 et KG1 sont résumés dans les Figures 4.17 et 4.18. Pour la lignée KG1, les niveaux du pourcentage de l'input pour les facteurs c-Myb et RUNX1 sont plus élevés que le contrôle négatif IgG, c-Myb est plus élevé pour le promoteur alternatif par rapport à IgG, c-Myb et RUNX1 sont plus élevé pour l'enhancer 7 et 8 en comparaison avec le contrôle IgG Pour la lignée K562, les niveaux du pourcentage de l'input pour les facteurs c-Myb, c-Myc et RUNX1 sont plus élevés pour le promoteur ainsi que pour le promoteur alternatif. Pour l'enhancer 7 et 8 tous les facteurs de transcription testés (C-Myc, c-Myb, IKZF1, RUNX1) sont plus élevé que le contrôle négatif IgG.



	cMyc	cMyb	IKZF1	RUNX1	IgG
ALDH1A1 Promoter	4,4%	11,4%	5,3%	13,5%	5,7%
ALDH1A1 Alt Promoter	5,8%	19,0%	3,3%	5,8%	7,2%
ALDH1A1 Enhencer7	2,7%	14,2%	8,3%	30,1%	11,6%
ALDH1A1 Enhencer8	5,1%	10,6%	5,1%	18,9%	6,4%

FIGURE 4.17. – *Résultat du ChIP qPCR pour les cellules KG1 pour nos 4 régions d'intérêt pour les facteurs c-Myc, c-Myb, IKZF1 et le contrôle négatif IgG.*



	cMyc	cMyb	IKZF1	RUNX1	IgG
K562 ALDH1A1 Promoter	123%	126%	16%	52%	29%
K562 ALDH1A1 Alt Promoter	26%	65%	2%	51%	14%
K562 ALDH1A1 Enhancer7	9%	68%	26%	33%	2%
K562 ALDH1A1 Enhancer8	29%	61%	14%	41%	4%

FIGURE 4.18. – *Résultat du ChIP qPCR pour les cellules K562 pour nos 4 régions d'intérêt pour les facteurs c-Myc, c-Myb, IKZF1 et le contrôle négatif IgG.*

4.4.3.4. Résultat de la Luciferase Assay

Les résultats du luciferase assay sont résumés dans la Figure 4.19. Le plasmide pGL4 représente notre contrôle négatif qui a un fold change très bas, le pGL4-pSV4 notre contrôle positif avec un fold change de 1. Le promoteur de 'ALDH1A1 et le promoteur alternatif de l'ALDH1A1 ont des valeurs très faibles. Pour les constructions sous le promoteur constitutif sv40 SV40 enhancer 7 (F/R) SV40 enhancer 8 (F/R), ont des valeurs un peu plus élevées que le promoteur alternatif mais des valeurs inferieur a 1. Toutes les constructions sous le promoteur de l'ALDH1A1 sont au même niveau que celui le contrôle négatif.

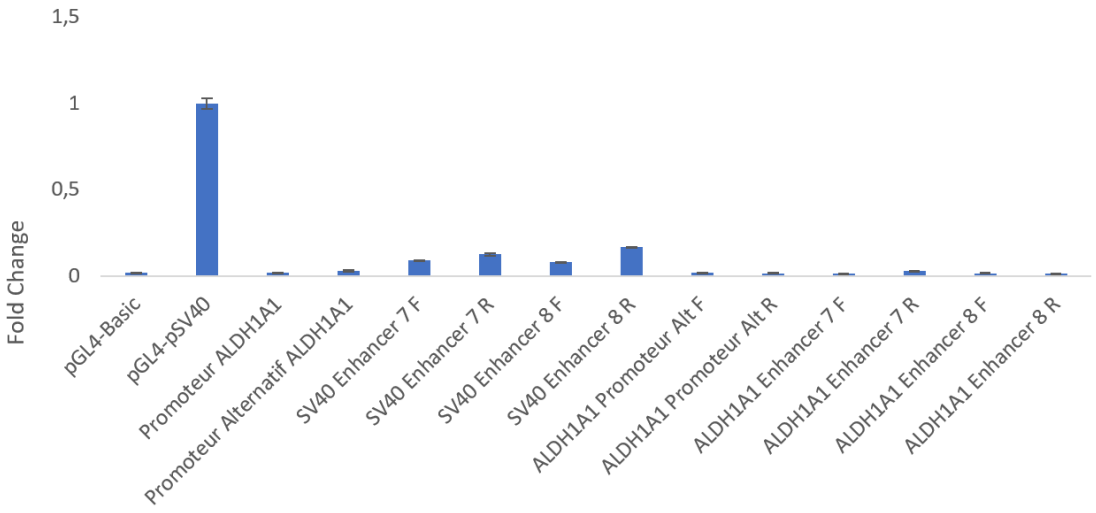


FIGURE 4.19. – *RTest de l'activité des différentes constructions en fold change.*

TABLEAU 4.4. – *Test de l'activité des différentes constructions en fold change.*

Samples	**Fold Change**
pGL4-Basic	0,020117157
pGL4-pSV40	1
Promoteur ALDH1A1	0,018946523
Promoteur Alternatif ALDH1A1	0,030319249
SV40 Enhancer 7 F	0,091922967
SV40 Enhancer 7 R	0,125997731
SV40 Enhancer 8 F	0,082523958
SV40 Enhancer 8 R	0,166435121
ALDH1A1 Promoteur Alt F	0,018844975
ALDH1A1 Promoteur Alt R	0,017505004
ALDH1A1 Enhancer 7 F	0,013504368
ALDH1A1 Enhancer 7 R	0,032177177
ALDH1A1 Enhancer 8 F	0,017281386
ALDH1A1 Enhancer 8 R	0,013101518

4.4.3.5. Résultat du siRNA

Dans le but testé le potentiel l'effet du silencing des facteur de transcription MYB, MYC et RUNX1 sur l'expression de l'ALDH1A1 nous avons réalisé une manipulation de siRNA. Avant de tester l'effet du silencing sur l'expression de l'ALDH1A1 nous devons tout d'abord tester l'effet de chaque siRNA sur le TF qui lui correspond. Nous avons testé 3 siRNA par TF ainsi qu'un anti-GAPD qui représentera nôtre contrôle positif. Plusieurs conditions de concentrations pour chaque TF sont testées et une incubation à 48 et à 72h. Les résultats sont résumes dans les Figures 4.20, 4.21, 4.22 et 4.23.

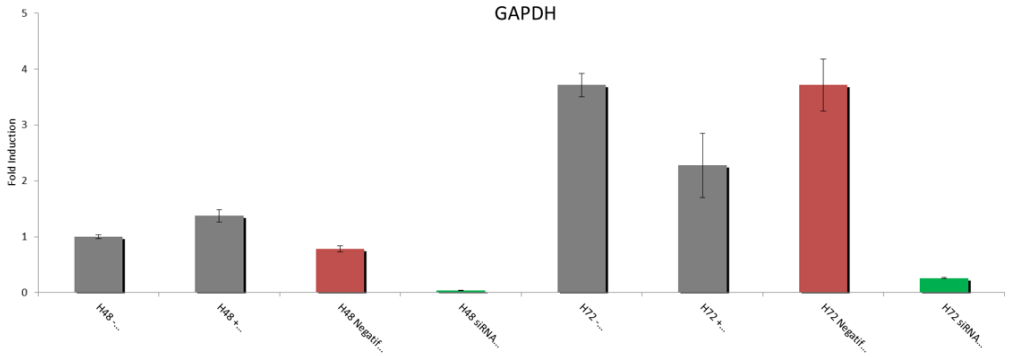


FIGURE 4.20. – ***Silencing du gène GAPDH.*** *Expression du GAPDH (en rouge) par rapport a l'échantillon avec siRNA GAPDH (en vert). En gris les contrôles négatifs : cellules sans électroporation et cellules avec électroporation san siRNA. Les analyses ont été faites a 48 et 72h.*

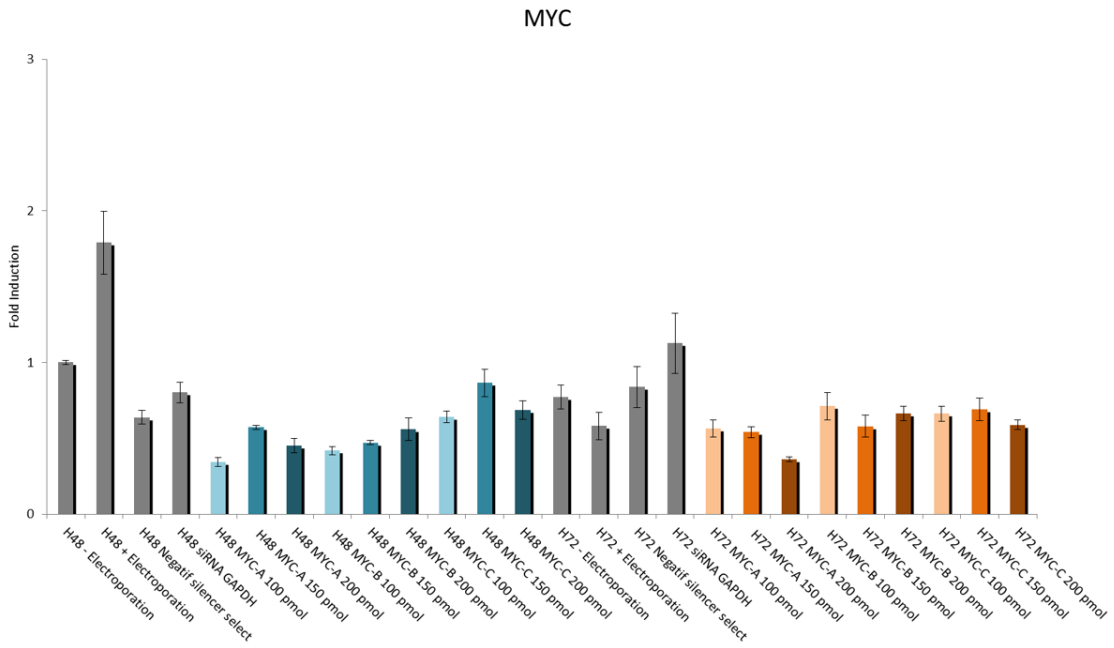


FIGURE 4.21. – ***Silencing du gène MYC.*** *Inhibition de l'expression de MYC (dégradé de vert à 48h et dégradé d'orange à 72h) par rapport aux contrôles négatifs et positif (en gris).*

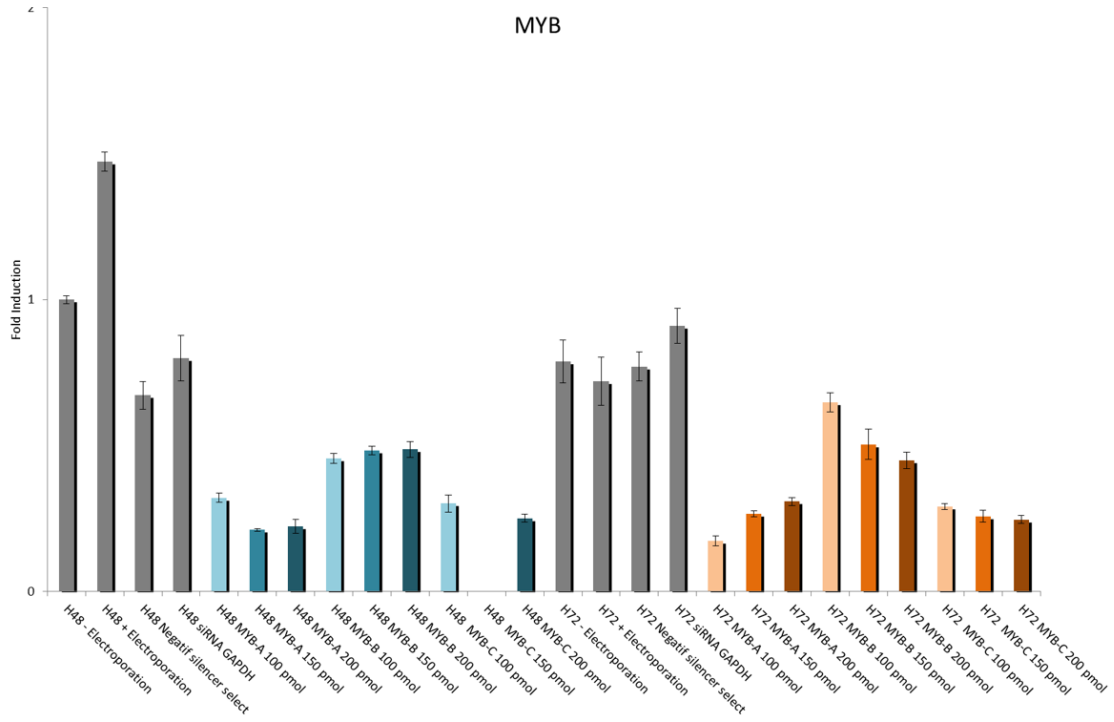


FIGURE 4.22. – ***Silencing du gène GAPDH.*** *Inhibition de l'expression de MYB (dégradé de vert à 48h et dégradé d'orange à 72h) par rapport aux contrôles négatifs et positif (en gris).*

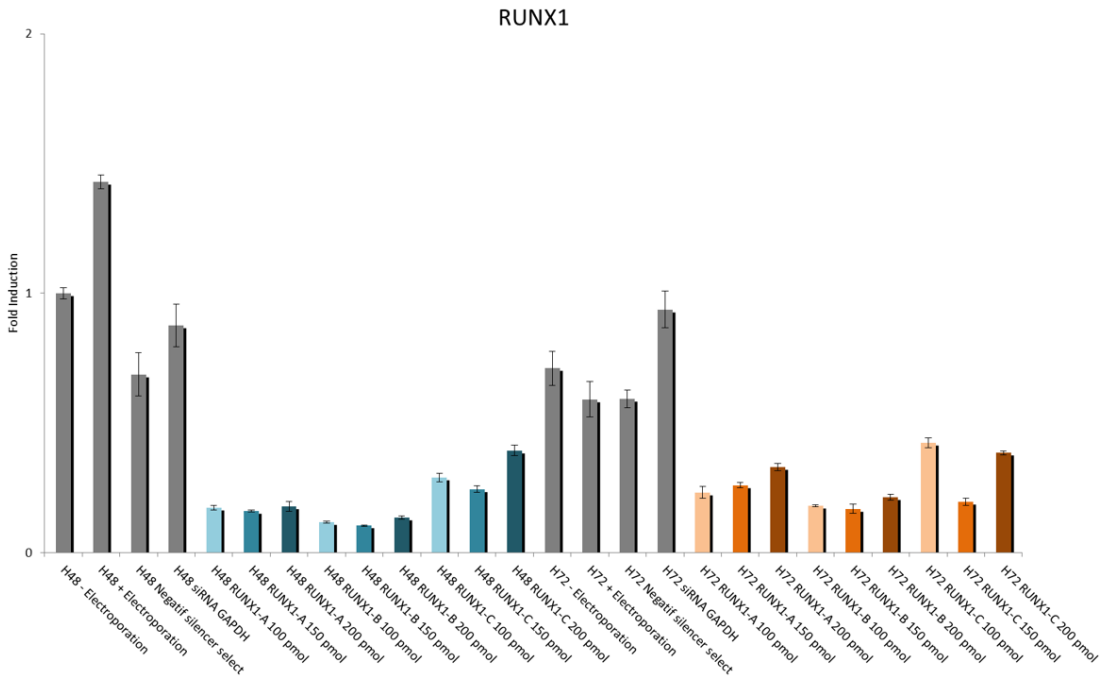


FIGURE 4.23. – ***Silencing du gène RUNX1.*** *Inhibition de l'expression de MYB (dégradé de vert à 48h et dégradé d'orange à 72h) par rapport aux contrôles négatifs et positif (en gris).*

4.4.4. Conclusion

Le screening de l'expression de l'ALDH1A1 dans les différentes lignées cellulaires ainsi que la quantité de protéine montrent une surexpression de l'ALDH1A1 dans les lignées KG1, K562 et KASUMI3 pour cette raison que nous avons choisi ces trois lignées cellulaires pour toutes nos expérimentations. Notre hypothèse de travail se basait sur le fait que si un facteur de transcription régulerait potentiellement l'ALDH1A1 son expression doit être corrélée à l'expression de ce dernier, autrement dit, dans les cellules qui expriment le plus l'ALDH1A1 selon notre expérience (KASuMI3, KG1, K562) nous devons choisir les facteurs de transcription qui ont une expression basale la plus haute.

En ce qui concerne le screening des 14 facteurs, nous avons décidé de tester les facteurs cMyC, cMyb, RUNX1 et IKZF1. Ces quatre derniers jouent un rôle important et/ou dérégulés et/ou présentent des altérations génétiques dans la AML. Par exemple 8 - 10% des AML secondaires montrent des mutations de RUNX1 dû au traitement aussi, 6 à 33% des cohortes hétérogènes montrent aussi des mutations de RUNX1. Ces mutations sont associées à un mauvais pronostic dans les cohortes de patient avec une cytogénétique hétérogène [107]. Les résultats de ChIP qPCR montrent une fixation potentielle des TFs c-Myb et RUNX1 sur le promoteur, c-Myb pour le promoteur alternatif, c-Myb et RUNX1 pour enhancer 7 et 8 pour la lignée KG1 Pour la lignée K562, nous avons une fixation potentielle des TFs c-Myb, c-Myc et RUNX1 sur le promoteur et le promoteur alternatif, c-Myc, c-Myb, RUNX1 et IKZF1 pour l'enhancer 7 et 8. Ces résultats sont prometteurs, en effet nous avons des valeurs qui sont bien au-dessus du contrôles négatif IgG comme pour les facteurs c-Myb, et c-Myc sur le promoteur dans la lignée K562 et d'autre qui sont moins franches comme cMyb pour les enhancer 7 et 8 pour la lignée KG1. Ce qui nous a posé problème dans cette manipulation, c'était les valeurs du pourcentage de l'input pour le contrôle négatif IgG qui étaient assez importantes dans certains échantillons. Cela peut être expliqué par le choix du contrôle négatif (IgG), ce dernier faisant partie d'un kit commercial, il a été utilisé comme tel. Pour valider ces résultats nous proposons d'utiliser une autre technique telle que le CRISPER cas 9.

L'analyse luciferase assay n'est pas du tout concluante même si nous avons des valeurs un signal pour le contrôle positif et un contrôle négatif avec une valeur proche de 0 cela veut dire que techniquement l'expérience a bien fonctionné. Nous pouvons expliquer ce résultat par le choix de la lignée cellulaire K562 pour cette expérience. En effet, la lignée K562 était pour nous le meilleur model pour faire cette expérience, K562 sont des cellules lymphoblastes isolées de la moelle osseuse d'un patient de 53 ans atteint de leucémie myéloïde chronique de plus cette lignée est facilement transfectable, avec le recul nous constatons que ce modèle n'est peut-être pas adapté pour l'étude de l'ALDH1A1. Après une analyse de l'expression de l'ALDH1A1 sur PROTEIN ATLAS (Figure 4.24) nous constatons que les K562 ont très peu d'ALDH1A1 par rapport à d'autres lignées telles que les A549. D'un autre côté, ces dernières sont des cellules de cancer du poumon et non pas des cellules myéloïdes, ce qui nous amène à se demander si elles feraient un modèle adéquat pour répondre à notre hypothèse.

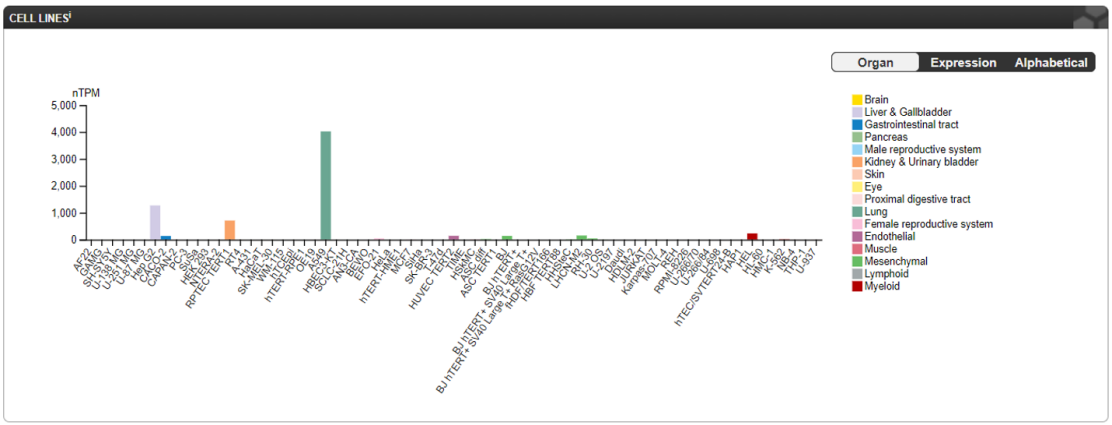


FIGURE 4.24. – *Analyse de l'expression de l'ALDH1A1 sur PROTEIN ATLAS.*

L'expérience de silencing montre que celle-ci a très bien fonctionné techniquement, en effet nous avons une très bonne inhibition du gène GADPH à 48 et à 72h. Pour les facteurs de transcriptions MYC MYB et RUNX1 nous constatons une meilleure inhibition à 48h comparé à 72h, on peut dire que l'inhibition maximale est à 48h même si cette dernière est partielle et n'a jamais était complétement efficace malgré les tests multiples, les concentrations multiples ainsi que les temps d'incubation multiple.

5. Discussion et perspectives

La thématique de ma thèse se focalise sur la régulation du génome sur plusieurs niveaux. Nous avons d'abord identifié les régions régulatrices des génomes de quatres espèces, par une approche à grande échelle. Nous avons ensuite réalisé des travaux en recherche fondamentale et évolutive en analysant l'impact des TE dans l'insertion des TFBS sur les génomes. Et enfin par une approche multi-omique, nous avons cartographié l'environnement régulateur d'un gène spécifique, l'ALDH1A1. Dans cette partie, je discuterais des différentes perspectives portant sur ces travaux.

5.1. Optimisation du catalogue ReMap

5.1.1. Caractérisation des éléments régulateurs

Actuellement, ReMap permet de cartographier la fixation des facteurs de transcription sur le génome, et ainsi de localiser les régions régulatrices. Cependant ces régions régulatrices ne suffisent pas à la classification des éléments *cis*-régulateurs (CRE) identifiés. Une compréhension détaillée de chaque CRE dans le génome permettra de caractériser la régulation de gènes qui contrôlent de multiples processus tels que le développement, la différenciation cellulaire et l'adaptation des espèces à leur environnement. Une telle compréhension est également cruciale pour interpréter l'impact des régions non codantes dans les maladies humaines et phénotypes complexes.

Plusieurs travaux à grande échelle, notamment le projet Roadmap Epigenome et ceux menés par le consortium ENCODE, ont caractérisé l'épigénome de centaines d'échantillons de tissus, de cellules primaires ou de lignées cellulaires pour annoter des millions de CRE candidats (cCREs) [54, 31, 200, 163] dans le génome humain et de souris. Les cCREs résultants ont été classés comme des éléments associés à des promoteurs ou à des enhancers en fonction de l'accessibilité de la chromatine, de l'hypométhylation de l'ADN et de certaines modifications d'histones tel que H3K4me3 pour les promoteurs actifs ou latents, H3K4me1 pour les enhancers latents, amorcés et actifs, ou H3K27ac pour les enhancers et les promoteurs actifs, ou comme des éléments insulateurs basés sur la liaison de CTCF [54, 31, 200, 163].

Ces catalogues de cCREs, couplées aux profils d'intéraction de la chromatine, fournissent une ressource précieuse pour étudier la régulation des gènes dans des tissus et des types cellulaires distincts chez l'homme et d'autres espèces. Ils permettent de déterminer le rôle des régions non codantes dans l'étiologie des maladies humaines et des phénotypes complexes [110].

Lorsque nous comparons les régions régulatrices ReMap avec les cCREs d'ENCODE, nous constatons que certaines régions de ReMap sont similaires à celles d'ENCODE, et d'autres permettent d'identifier de nouveaux éléments par rapport à ENCODE (Figure 5.1). Nous pourrions nous demander s'il existe un moyen de classer ces nouveaux éléments, pour apporter un information complémentaire aux cCREs d'ENCODE.

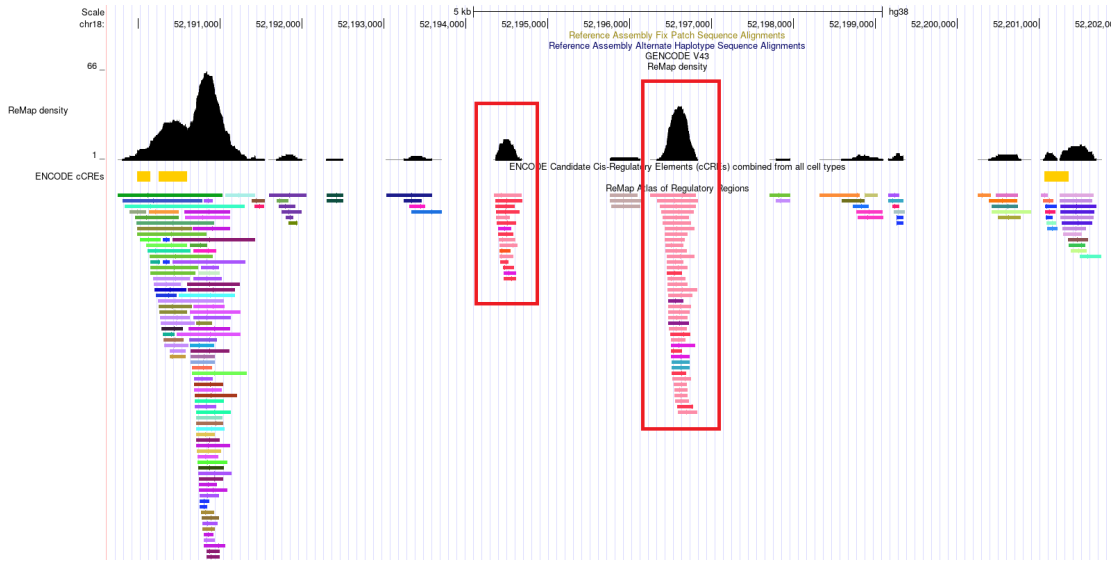


FIGURE 5.1. – ***Régions régulatrices identifiées par ReMap mais pas par les cCREs ENCODE.*** *Navigateur de génome UCSC de la région chr18 :52,189,620-52,203,543 du génome humain hg38. Piste 1 : Annotation GENCODE des gènes, Piste 2 : Courbe de densité ReMap, Piste 3 : cCREs ENCODE, Piste 4 : Pics ChIP-seq ReMap. Les encadrés rouges montrent les régions identifiées uniquement par ReMap.*

Ces observations permettent de supposer que la caractérisation des régions ReMap permettrait de détecter de nouveaux CREs non identifiés par ENCODE. Cependant les données ChIP-seq actuellement disponibles dans le catalogue ReMap ne permettent pas à elles seules de classifier avec certitude les régions (promoteur proximal, enhancer distal, etc.). En effet, bien que les régions soient déjà très informatives, il faudrait complémenter ces données avec d'autres types de données génomiques qui seront énoncés dans le paragraphe suivant.

Les promoteurs et les enhancers dirigent l'expression génique de manière spécifique à chaque type cellulaire en interagissant avec des combinaisons de TF pour faciliter la transcription. Ces interactions sont également régulées par d'autres mécanismes, notamment l'accessibilité de la chromatine, qui peut être profilée à l'aide de méthodes telles que le séquençage des sites hypersensibles à la DNase I (DNase-seq) [36, 122] et l'ATAC-seq [40] ; la méthylation de l'ADN, qui peut être profilée à l'aide du séquençage des modifications des histones, qui peuvent être identifiées par expérience ChIP-seq [20, 202, 137]. En outre, la régulation transcriptionnelle par les promoteurs et les enhancers dépend également de leur organisation spatiale dans le noyau. Les fibres de chromatine dans le noyau des cellules eucaryotes sont séparées en domaines d'association topologique (TADs) [94]. Les insulateurs, qui délimitent les limites des TADs et jouent un rôle critique dans leur formation, peuvent être identifier par la présence de pics ChIP-seq de CTCF. À travers leur rôle dans la formation de TAD, les insulateurs facilitent les interactions entre les enhancers et les promoteurs à l'intérieur du même TAD et réduisent les contacts entre les promoteurs et les enhancers situés dans des TAD séparés. La fréquence de ces contacts peut être utilisée pour déduire l'architecture de la chromatine et peut être mesurée à l'aide de méthodes de capture de conformation de chromosome à haute résolution telles que Hi-C [175, 237].

Le catalogue ReMap pourrait bénéficier de l'ajout de ces différentes sources de données pour compléter les informations existantes. En effet, les données DNase-seq et ATAC-seq permettraient de détecter les régions de chromatine ouvertes et donc potentiellement actives. Les données ChIP-seq de marques d'histones permettraient d'identifier le type de CRE détecté par ReMap, tel que les promoteurs centraux ou les enhancers distales. Et enfin les données Hi-C couplées aux données ReMap permettraient d'identifier les régions insulatrices et bordures de TADs. En effet le catalogue montre des régions CTCF spécifiques qui sont très probablement des bordures de TADs (Figure 5.2).

Le catalogue ReMap a permis d'identifier 3,4 millions CRMs, parmi ces régions nous pouvons déterminer lesquelles sont actives grâce aux données ChIP-seq de polymérase II (polII). H. Sun et al. [282] ont permis l'identification des promoteurs alternatifs des gènes de souris à l'aide d'expériences ChIP-seq de polII. L'analyse a permis d'identifier 38,639 promoteurs fixés par l'ARN Pol-II, dont 12 270 sont nouveaux. Les résultats indiquent que 37% des gènes codants pour des protéines utilisent des promoteurs alternatifs dans les cinq tissus de souris étudiés. Les annotations de promoteurs et les données ChIP-seq obtenues dans cette étude contribuent à la caractérisation des régions régulatrices des gènes dans les génomes mammifères. Les travaux de R. Gupta et al. [112] ont également identifié des promoteurs alternatifs dans le génome de souris à l'aide de données ChIP-seq de polII. Un étudiant en thèse dans notre équipe travaille actuellement sur la cactérisation des régions régulatrices "actives" grâce à l'intégration de données ChIP-seq de polII.

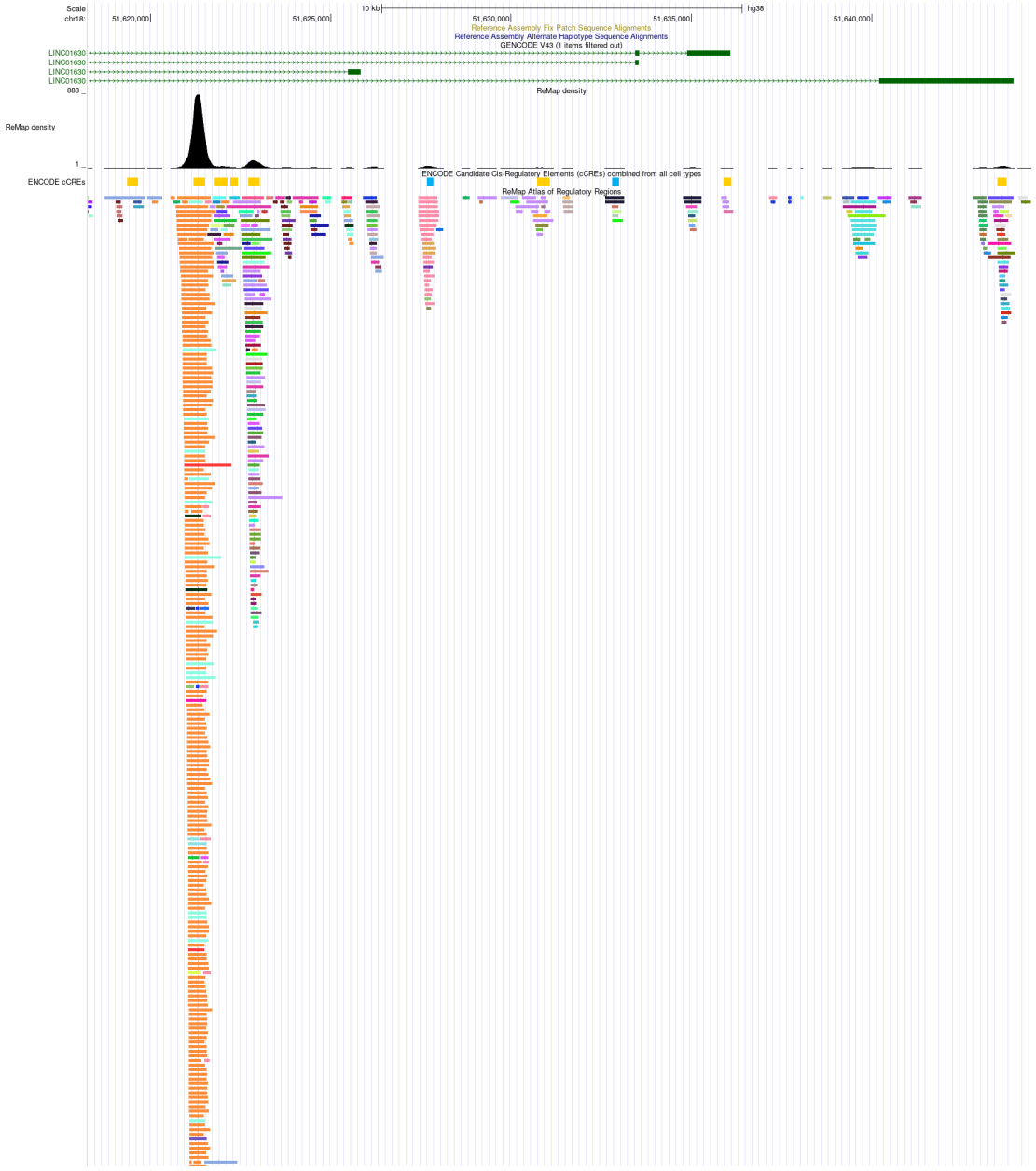


FIGURE 5.2. – ***Région du catalogue ReMap riche en CTCF.*** *Capture d'écran du navigateur de génome UCSC du catalogue de ReMap. Zoom sur la région chr18 :51,618,263-51,644,540. Sur cette image on peut voir un pic ReMap composé très majoritairement par des pics CTCF (en orange). On peut supposer que cette région est une bordure TAD identifiée par ReMap.*

5.1.2. Utilisation de données Single Cell ChIP-seq

Le catalogue ReMap a plusieurs limitations qui pourraient être en partie résolues par du séquençage à cellules uniques (scChIP-seq, scATAC-seq... etc). En effet, le catalogue manque de résolution au niveau du type cellulaire car les datasets ont été générés à partir de tissus ou lignées cellulaires "bulk" (en vrac). De plus, le ChIP-seq mesure le comportement moyen au travers de cellules dans un échantillon biologique constitué de milliers à des millions de cellules. Lorsqu'un échantillon hétérogène (par exemple, un échantillon de tissu) constitué de plusieurs types de cellules ou d'états de cellules est analysé, ces technologies en vrac peuvent manquer des signaux biologiques importants portés uniquement par un sous-ensemble de cellules. En effet, les éléments régulateurs sont souvent actifs uniquement dans des types cellulaires, des stades de développement ou des états physiologiques, spécifiques, dont beaucoup sont difficiles à récolter en quantité suffisante pour le séquençage [188, 320, 143]. Seuls les types cellulaires présents en grand nombre et avec des anticorps bien caractérisés, tels que les cellules sanguines, sont aptes au séquençage en quantités suffisantes, tandis que les types de cellules rares ou non caractérisés échapperont à l'analyse [45].

Le développement de techniques à cellule unique offre un moyen de surmonter certaines de ces limitations en générant un catalogue ReMap plus complet qui permet l'étude des régions régulatrices sur des types cellulaires spécifiques. Les techniques single-cell rendent possible la cartographie des éléments régulateurs dans les cellules individuelles. Par exemple, le séquençage de l'ATAC à cellule unique (scATAC-seq) [41, 62] et le séquençage de la DNase à cellule unique (scDNase-seq) [136] sont deux technologies pour analyser la chromatine ouverte, une marque des éléments *cis*-régulateurs actifs, dans les cellules individuelles. Le ChIP-seq à cellule unique (scChIP-seq) [245], permet l'analyse à cellule unique de la modification des histones et de la fixation des TF. Le Consortium Human Cell Atlas (HCA) utilise le scATAC-seq comme un outil majeur pour caractériser le paysage régulateur des cellules humaines [238]. Nous pourrions faire de même et construire l'un de premiers catalogues de régions régulatrices basé sur des techniques de séquençage à cellule unique.

5.1.3. Ajout d'espèces modèles à ReMap

Le projet Remap est une initiative qui vise à faciliter l'analyse des régions régulatrices de différentes espèces en fournissant un catalogue complet disponible publiquement. L'ajout de catalogue de régions régulatrices d'espèces modèles permettrait de rendre ces données accessibles aux chercheurs travaillant sur ces espèces et de faciliter l'analyse de leur régulation. En effet ChIP-atlas [326] par exemple, offre l'accès aux données ChIP-seq de 6 espèces modèles (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*).

Ces espèces modèles sont largement utilisées en recherche de biologie pour étudier différents processus biologiques et pour comprendre la génétique sous-jacente. Les données générées à partir d'espèces modèles telles que la souris, la drosophile et le zèbre-poison peuvent fournir des informations précieuses sur les régions régulatrices de l'ADN et leur fonction [125]. Cependant, l'analyse des données ChIP-seq peut être compliquée en raison de la diversité des annotations. En intégrant ces espèces au catalogue, ReMap simplifierait donc l'analyse de ces données en fournissant une référence commune pour les chercheurs travaillant sur ces espèces.

En outre, l'ajout de catalogues de régions régulatrices de plusieurs espèces modèles permettrait également de mieux comprendre l'évolution des régions régulatrices à travers les espèces et de fournir des informations précieuses sur la manière dont les processus biologiques ont évolué au fil du temps [289].

En comparant avec l'atlas de l'homme, 412 régulateurs transcriptionnels sont communs avec les catalogues de souris, ce qui pourrait permettre l'exploration de la conservation évolutive des modules *cis*-régulateurs. Cependant dans l'état actuel du catalogue nous ne pouvons pas prétendre à effectuer une analyse comparative complète. C'est pour cela que je propose d'étudier d'autres espèces modèles : *Rattus norvegicus* (205 datasets dans GEO), *Caenorhabditis elegans* (594 datasets dans ENCODE, et 588 dans GEO) et Zebrafish (180 datasets dans GEO).

L'ajout de ces données permettrait d'analyser les mécanismes de régulation conservés à travers l'évolution, les différences entre les espèces qui peuvent être liées à des adaptations spécifiques [123]. L'utilisation d'espèces modèles présentant des maladies génétiques similaires à celles de l'homme peut aider à identifier les mécanismes sous-jacents à ces maladies, tel que les maladies immunitaires [319].

5.1.4. Analyse gène cible

ReMap pourrait intégrer dans ces fonctionnalités la possibilité d'identifier des gènes qui sont ciblés par les TF des régions régulatrices identifiées par ReMap. Cela peut être fait en utilisant différentes approches, telles que :

- L'analyse de co-localisation : il s'agit d'identifier les gènes qui se trouvent proches des régions régulatrices identifiées par ReMap avec des outils tel que RegulatorTrail et TIP [149, 48].

- L'analyse de la corrélation génétique : il s'agit de mesurer la corrélation entre l'expression génique et les régions régulatrices identifiées par ReMap [251, 82].

- L'analyse de l'expression différentielle : il s'agit de mesurer les différences d'expression génique entre différents groupes cellulaires ou tissulaires [174].

5.2. Implication des TE dans la régulation

5.2.1. Espèces analysées

Dans nos travaux récents, nous nous sommes concentrés sur l'homme et la souris en tant que modèles pour étudier les interactions entre les éléments transposables et les facteurs de transcription. Cependant, nous n'avons pas eu le temps de faire une comparaison complète entre ces deux espèces, comme l'ont fait d'autres travaux [35, 162]. Par ailleurs, nous nous sommes concentrés sur l'homme et la souris mais il est tout à fait possible de réaliser ces mêmes analyses sur les données ReMap disponibles de la Drosophile et *Arabidopsis Thaliana*. Dans cette partie nous discuterons des travaux interespèce réalisés dans le domaine afin de mettre en valeur l'utilité d'une telle analyse sur nos données.

Des travaux récents ont déjà examiné l'implication des TE dans la régulation chez la Drosophile. Par exemple, dans un article de A.Ullastres et al. [295], les auteurs ont utilisé une approche de génétique de population pour identifier les insertions d'éléments transposables susceptibles d'augmenter la tolérance de *Drosophila melanogaster* à l'infection bactérienne en affectant l'expression des gènes liés à l'immunité. L'analyse a identifié 12 insertions associées à des changements d'expression génique spécifiques à l'allèle dans les gènes liés à l'immunité. Ces résultats ont été validés expérimentalement pour trois de ces insertions, y compris une qui agit probablement en tant que régulateur négatif, une en tant que régulateur positif et une avec un rôle double en tant que régulateur positif et promoteur. La direction du changement d'expression génique associé à la présence de plusieurs de ces insertions est cohérente avec une augmentation de la survie à l'infection.

En nous inspirant de cet article, nous pourrions étendre la recherche d'un lien entre TE et TFBS sur un jeu de données plus vaste (550 TF disponible avec ReMap) qui mettrait en valeur surement plus de 12 associations significatives. Appliquée à la plante, cette analyse sera une des premières à être réalisée sur *A. thaliana*. A partir de ces résultats nous pourrions faire une comparaison interespèces des résultats de l'analyse de l'implication de TE dans l'insertion des TFBS dans le génome. Cette analyse permettrait de mettre en évidence l'évolution des réseaux de régulation et permettrait de déterminer en quoi l'insertion de nouveaux sites TFBS par les TE a pu influencer cette évolution. Certains TE étant spécifique aux espèces nous pourrions comparer les associations TE/TF chez différentes espèces, afin de découvrir des mécanismes spécifiques à chaque espèce qui sont peut-être associés à leur adaptation au cours de l'évolution.

Par exemple, Bourque et al. [35] ont utilisé l'âge des familles de TE pour montrer qu'ils sont associés à une expansion significative de la régulation au cours de l'évolution des mammifères. Les auteurs révèlent des exemples de TE associés à la régulation, tels que ESR1 sur MIR, qui a été identifié comme étant antérieur à la radiation des mammifères, et CTCF sur B2, qui est associé à des TE actuellement actifs chez les rongeurs. Ces résultats suggèrent que les réseaux de régulation transcriptionnelle de l'homme et de la souris peuvent être similaires dans des processus biologiques importants tels que le cancer et les cellules souches.

Les contributions des TE aux paysages régulateurs sont particulièrement marquées lorsque l'on considère les enhancers spécifiques à une espèce, y compris à travers des vagues d'expansions de TE [35] (Figure 5.3).

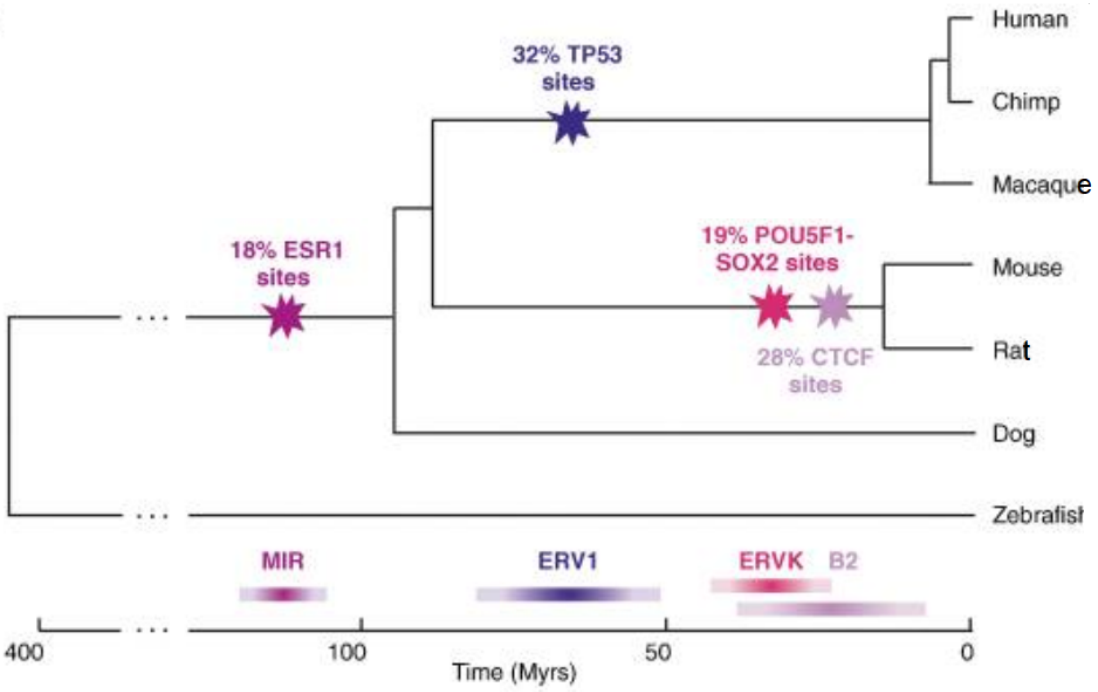


FIGURE 5.3. – ***Évolution du répertoire des TFBS chez les mammaliens via les TE.*** *Superposition de l'âge des répétitions sur l'arbre des espèces permet de déterminer l'âge des RABS (équivalent aux TE des paires de TE/TF associés lors de notre analyse)* [1].

3. Bourque, Guillaume, et al. "Evolution of the mammalian transcription factor binding repertoire via transposable elements." Genome research 18.11 (2008)

Par exemple, la majorité des enhancers spécifiques aux singes et humains du foie chevauchent des transposons [291]. Les séquences des transposons contribuent donc à l'évolution des enhancers entre les espèces proches et conduisent probablement à des différences phénotypiques [284]. Dans la plupart des tissus humains et murins, les TE enrichis associés aux enhancers sont généralement plus anciens que les TE génomiques non régulateurs [228, 266]. Cela suggère que les mutations qui se sont produites au cours de l'évolution ont permis aux TE d'acquérir des mutations favorisant l'activité enhancer. En effet, les enhancers utérins et hépatiques ont souvent évolué à partir de séquences de TE anciennes [185, 301].

Edward B. Chuong et al. [50] en superposant l'apparition de TE (associé a des TF) à l'arbre des espèces, ont déterminé que les sites identifiés pour ESR1 sont probablement ancestraux aux mammifères, ceux pour TP53 sont spécifiques aux primates, ceux pour POU5F1-SOX2 sont spécifiques aux rongeurs, et ceux pour CTCF sont soit spécifiques aux rongeurs ou à la souris en particulier. Cependant, il reste encore beaucoup de questions sans réponses quant à la fonction biologique de ces TFBS dérivés des TE. Bien que les événements de liaison rapportés indiquent une activité biochimique de l'association TF-ADN, il reste à déterminer si ces TFBS dérivés des TE peuvent influencer l'expression des gènes.

5.2.2. Tissu spécificité

Notre analyse a été effectuée sur plus de 700 biotypes différents, ce qui la rend intéressante comparée aux analyses similaires publiées précédemment. Celles-ci se sont concentrées sur des petits jeux de données et sur des types cellulaires bien spécifiques. Par exemple, Kunarso et al. [162] décrivent l'insertion de nouveaux TFBS dans le génome par l'intégration de TE dans des cellules souches embryonnaires humaines. Les travaux de Jiang et al. [133] ont également identifié le rôle des L1PA2 dans la régulation dans des cellules cancéreuses du sein.

En suivant ces exemples, l'analyse de l'impact des TE sur les réseaux de régulation génétique nous permettrait de déterminer la spécificité cellulaire de certaines associations entre TE et TF. En comparant nos résultats avec des analyses similaires sur des biotypes spécifiques, nous pouvons distinguer les associations qui sont ubiquitaires et celles qui sont spécifiques à la lignée. A la fin de ma thèse j'ai lancé les analyses sur trois tissus principaux : le sang (Blood), le sein (Breast) et le foie (Liver). Cependant je n'ai pas eu les temps d'analyser les résultats avant la fin de ma thèse.

En conclusion, notre analyse offre une vue d'ensemble plus globale de l'impact des TE sur les réseaux de régulation génétique et de la spécificité cellulaire de ces associations. Il est important de préciser que les analyses tissus-spécifiques ont déjà été lancées à la fin de ma thèse mais les résultats n'ont pas encore été interprétés.

5.2.3. Fonction biologique des gènes régulé par les TF associés

Il est important d'étudier la régulation des gènes cibles des TF insérés par les TE pour comprendre comment ces derniers ont influencé l'évolution des réseaux de régulation transcriptionnelle en insérant des TFBS proches de certains gènes. Plusieurs des travaux cités précédemment ont notamment effectué une analyse complémentaire portant sur l'expression des gènes cibles de ces TFBS.

L'article de Kunarso et al. [162] a permis de déterminer l'importance fonctionnelle de RABS en déplétant les cellules souches embryonnaires humaines de POU5F1 par RNAi (ARN à interférence) et en examinant l'expression différentielle des gènes par analyse microarray. Les résultats ont montré un enrichissement de TFBS de POU5F1 et de NANOG, en particulier autour de 137 gènes downregulés.

Les travaux de Bourque et al. [35] concernent l'analyse de la relation entre les TE associées aux TFBS (RABS) et la régulation génétique. L'hypothèse était que si les RABS sont impliqués dans la régulation génétique, ils devraient se trouver à proximité de gènes régulés par les TF associés. L'analyse de données expérimentales a montré que, pour les TF ESR1 et POU5F1-SOX2, les RABS se trouvaient plus près de gènes régulés que de gènes choisis au hasard. Les résultats suggèrent que les RABS sont fonctionnels au niveau de la fixation des TF et sont susceptibles de réguler de nombreux gènes associés.

En 2007, T. Wang et al. [304] ont collecté des données publiées pour 392 gènes régulés par p53. Cette liste a été comparée à un ensemble de 440 gènes les plus proches qui ne se trouvent pas à plus de 1 Mb de chacun des 497 LTR10 et MER61 LTRs avec un site p53. Trente et un des 392 cibles connues de p53 étaient parmi les 440 gènes associés à un site dérivé d'un LTR p53. En s'inspirant des articles précédents, nous prévoyons de réaliser une analyse target genes sur nos TE associés à des TF. Il serait également intéressant de coupler cette analyse à des données d'expression afin de déterminer si les gènes ciblés sont réellement régulés par des TF insérés par des TE. De plus, nous aimerions également combiner cette analyse avec l'âge des TE afin de déterminer s'il existe une potentielle corrélation entre l'âge d'insertion des TE et la fonctionnalité de TFBS insérés.

Des recherches sont en cours afin de terminer nos analyses sur l'impact des TE sur la régulation. Nous aimerions identifier des processus biologiques communs dans les gènes situés à proximité de nos séquences de TE enrichies en facteurs de transcription. Pour cela, je souhaiterais effectuer une analyse en utilisant la méthode GSEA (Gene Set Enrichment Analysis) sur les gènes proches des séquences de TE associées à des TFBS.

Dans le cas des gènes proches des séquences de TE (Transposable Elements) associées à des TFBS (site de fixation de facteur de transcription), une analyse GSEA peut aider à comprendre si ces gènes sont impliqués dans des voies ou des processus biologiques pertinents. Les TE sont des séquences d'ADN répétitives qui peuvent se déplacer dans le génome et influencer la régulation en se fixant à des sites de reconnaissance spécifiques des TF. Les TFBS associés aux TE peuvent donc jouer un rôle important dans la régulation de l'expression génique.

Il serait particulièrement utile de réaliser une analyse GSEA séparément pour les TE anciens et les TE récents. Cela pourrait nous permettre de déterminer les différences dans les fonctions biologiques associées à ces deux types de TE. Par exemple, on peut s'attendre à ce que les TE anciens soient associés à des gènes avec des fonctions biologiques anciennes tandis que les TE récents soient associés à des gènes avec des fonctions biologiques plus récentes et spécifiques.

5.2.4. Analyse sur les STR

Les répétitions en tandem courtes (STR), également appelées microsatellites ou répétitions de séquences simples, sont des séquences d'ADN répétées en tandem qui impliquent une unité répétitive de 1 à 6 paires de bases [286], formant des séries d'une longueur allant jusqu'à 100 nucléotides. Les STRs sont largement présents dans les procaryotes et les eucaryotes, y compris les humains. Ils apparaissent plus ou moins régulièrement tout au long du génome humain, représentant environ 3% de l'ensemble du génome. La plupart des STRs se trouvent dans les régions non codantes, tandis qu'environ 8% se situent dans les régions codantes [77]. Bien que les STRs existent largement dans les organismes, la plupart d'entre eux ont été considérés comme ayant aucune utilisation biologique et considérés comme de l'ADN "inutile". Cependant, certains STRs peuvent participer à la régulation de la transcription. Par exemple, des recherches ont montré que certains STRs peuvent réguler la transcription du gène de facteur de croissance épidermique [101]. Les STRs $(TGYCC)_n$ sont également fixés par p53 dans le promoteur du gène PIG3 [58], ce microsattelite serait necessaire à l'activation transcriptionel du gène PIG3.

Bien que notre méthodologie et celle appliquée dans ces articles soit différente, il serait tout de même intéressant d'essayer d'adapter nos analyses aux spécificités des STR et ainsi comprendre leur contribution dans le paysage des éléments régulateurs.

5.3. Identification des régions régulatrices autour du gène ALDH1A1

Les premiers résultats de la validation expérimententale n'ont pas abouti. En effet, leChIP qPCR nous ont permis d'dentifier des TF potentiels qui devrons être validés par une autre technique comme le CRISPR cas-9. Celle de luciferasse assay est complement raté et n'a permis de valider aucune région. La manipe de siRNA devra être encore optimisée afin de tester l'effet du silencing des TF d'intérêt sur l'expression de l'ALDH1A1. Ce plan expérimental aussi complet et complexe n'a pas permis d'avoir des résultats concluants et définitifs malgré un travail qui a durée pratiquement 2 ans. Chacune des expériences a été optimisée et refaite plusieurs fois afin de valider les résultats.

Nous avons décidé avec nos collaborateurs d'élargir la région étudiée jusqu'au gène ANXA1 en aval du gène ALDH1A1. En effet, dans la première version créée en 2020 nous avions uniquement identifié les régions présentes entre les TAD délimitées par l'expérience Hi-C. Dans cette nouvelle version, nous avons également mis à jour le trackhub pour y intégrer les nouvelles données de ReMap 2022, ce qui nous a également permis de mettre en valeur des régions qui ne semblaient pas pertinentes en 2020. La Figure 5.4 représente le trackhub 2022 avec toutes les régions potentiellement régulatrices identifiées autour du gène ALDH1A1. Nous avons identifié 24 régions régulatrices autour du gène ALDH1A1. Deux de ces régions sont considérées comme des régions promoteurs en raison de leur position et de la présence de signaux ChIP-seq d'histone H3K4me3.

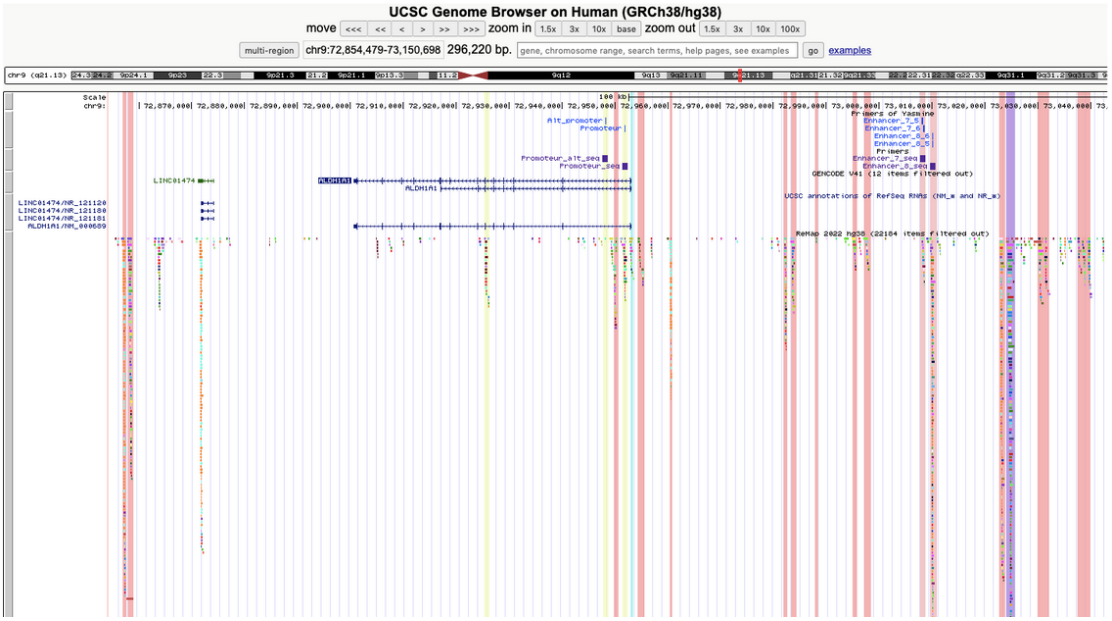


FIGURE 5.4. – *Amélioration du trackhub de l'ALDH1A1. Trackhub mis à jour en 2022 pour les régions potentiellement régulatrices de l'ALDH dans les lignées cellulaires AML.*

Afin de simplifier les travaux de nos collaborateurs, j'ai créé un fichier (Figure 5.5) qui regroupe la liste des 24 potentielles régions régulatrices identifiées, numérotées de gauche à droite. Ce fichier comporte également les TF présents dans chaque région, ainsi que le nombre de pics ChIP-seq associés à ces TF provenants de sources différentes. Les coordonnées génomiques de chaque région sont également incluses. Le but de ce fichier est de faciliter la sélection des nouveaux TF et régions à tester lors des validations expérimentales, en s'éloignant des TF identifiés dans la littérature ayant échoué lors des premières validations, pour se concentrer sur les régions contenant le plus de données expérimentales et ainsi que les TF identifiés par plusieurs sources différentes.

Regions	Coordinates	ADNP	AFF1	AR	ARID1B	ARID2	ARID3A	ARNT	ATF1	ATF2	ATF3	ATF4	ATF7	BACH1
Enhancer1	chr9:72856679-7285760	0	0	0	0	0	1	0	0	0	0	0	0	0
Enhancer2	chr9:72857699-7285899	0	0	0	1	0	0	0	1	0	0	0	1	1
Enhancer3	chr9:72925374-7292626	0	0	1	0	0	0	0	0	0	0	0	0	0
Enhancer4(Promoteur_alt	chr9:72947724-7294872	0	0	0	0	0	0	0	0	0	0	0	0	0
Enhancer5	chr9:72949660-7295045	0	0	0	1	0	0	0	0	0	2	1	0	0
Enhancer6(Promoteur)	chr9:72951419-7295241	0	1	0	0	0	0	0	0	0	0	0	0	0
Enhancer7	chr9:72952788-7295344	0	0	0	0	0	1	0	0	0	0	0	0	0
Enhancer8	chr9:72954270-7295557	0	0	0	0	0	0	0	0	0	0	1	0	0
Enhancer9	chr9:72960293-7296091	0	0	0	0	0	0	0	0	0	0	0	0	0
Enhancer10	chr9:72981993-7298264	0	0	1	1	0	0	0	0	0	2	1	0	0
Enhancer11	chr9:72983181-7298419	0	0	0	1	0	0	0	0	1	2	0	0	0
Enhancer12	chr9:72987900-7298830	0	0	0	0	0	0	0	0	0	0	0	0	0
Enhancer13	chr9:72994861-7299570	0	0	0	0	0	0	0	0	0	0	0	0	0
Enhancer14	chr9:72996970-7299811	0	1	0	0	0	0	0	1	0	0	0	0	0
Enhancer15(Enhancer7)	chr9:73007715-7300871	0	0	0	1	0	0	0	0	0	0	0	0	0
Enhancer16(Enhancer8/F	chr9:73009594-7301059	0	1	0	1	0	1	1	1	0	1	0	1	0
Enhancer17	chr9:73022730-7302389	1	0	1	0	0	1	0	2	1	2	0	1	0
Enhancer18(Fr)	chr9:73023932-7302561	1	2	0	1	0	1	3	1	1	2	0	1	0
Enhancer19	chr9:73029244-7303178	0	2	0	1	0	0	0	1	0	0	0	0	0
Enhancer20	chr9:73037400-7304032	1	0	0	0	0	0	0	0	0	0	0	0	0
Enhancer21	chr9:73081834-7308264	0	0	0	0	1	1	1	0	0	0	0	0	0
Enhancer22	chr9:73107250-7310792	0	0	0	1	0	0	0	0	0	0	0	0	0
Enhancer23(Fr)	chr9:73131466-7313346	0	2	1	1	0	0	2	2	1	2	1	1	1
Enhancer24	chr9:73143282-7314498	0	2	1	1	0	1	1	0	1	3	1	1	1
Total	-	3	11	5	11	1	7	8	9	5	16	5	6	3
Total_uniq_exp	-	3	7	5	11	1	7	5	7	5	8	5	6	3

FIGURE 5.5. – ***Liste des régions régulatrices identifié autour de l'ALDH1A1.*** *Capture d'écran d'une partie du fichier contenant la liste des régions régulatrices identifiées, leurs coordonnées ainsi que les TF se fixant potentiellement sur ces régions. Pour chaque TF nous avons compté le nombre d'expériences différentes ayant identifié le TF sur la région donnée.*

Après une analyse, une short liste de TF a été sorti (Figure 5.6).

Regions Coordinates	Total -	Total_uniq_exp -
HDAC1	24	16
BRD4	33	14
IKZF1	12	12
SMARCA4	16	12
ZEB2	18	12
ARID1B	11	11
ZBTB40	13	11
CEBPB	22	10
GATA2	16	7

FIGURE 5.6. – *Liste des TF à tester.*

Un nouveau plan expérimental été mis en place (par ABD) afin de tester l'implication de ces régions et facteurs de transcription dans la régulation de l'ALDH1A1. Ces techniques déjà mises en place seront complétées par du CRISPR cas 9 pour valider les TF et les régions les plus significatives.

Bibliographie

[1] European Bioinformatics Institute : Birney Ewan 3 Goldman Nick 3 Kasprzyk Arkadiusz 3 Mongin Emmanuel 3 Rust Alistair G. 3 Slater Guy 3 Stabenau Arne 3 Ureta-Vidal Abel 3 Whelan Simon 3, Research Group in BIOMEDICAL INFORMATICS ABRIL JOSEP F. 5 GUIGÓ RODERIC 5 PARRA GENIÉS 5, Bioinformatics Agarwal Pankaj 6 et al. « Initial sequencing and comparative analysis of the mouse genome ». In : *Nature* 420.6915 (2002), p. 520-562 (cf. p. 86, 162).

[2] US DOE Joint Genome Institute : Hawkins Trevor 4 Branscomb Elbert 4 Predki Paul 4 Richardson Paul 4 Wenning Sarah 4 Slezak Tom 4 Doggett Norman 4 Cheng Jan-Fang 4 Olsen Anne 4 Lucas Susan 4 Elkin Christopher 4 Uberbacher Edward 4 Frazier Marvin 4, RIKEN Genomic Sciences Center : Sakaki Yoshiyuki 9 Fujiyama Asao 9 Hattori Masahira 9 Yada Tetsushi 9 Toyoda Atsushi 9 Itoh Takehiko 9 Kawagoe Chiharu 9 Watanabe Hidemi 9 Totoki Yasushi 9 Taylor Todd 9, GENOSCOPE et al. « Initial sequencing and analysis of the human genome ». In : *nature* 409.6822 (2001), p. 860-921 (cf. p. 84, 86, 162, 175).

[3] Jeff D AALFS et Robert E KINGSTON. « What does 'chromatin remodeling'mean ? » In : *Trends in biochemical sciences* 25.11 (2000), p. 548-555 (cf. p. 42, 44).

[4] Mohamed Ibrahim ABOUELHODA, Stefan KURTZ et Enno OHLEBUSCH. « Replacing suffix trees with enhanced suffix arrays ». In : *Journal of discrete algorithms* 2.1 (2004), p. 53-86 (cf. p. 72).

[5] Imad ABUGESSAISA, Jordan A RAMILOWSKI, Marina LIZIO et al. « FANTOM enters 20th year : expansion of transcriptomic atlases and functional annotation of non-coding RNAs ». In : *Nucleic acids research* 49.D1 (2021), p. D892-D898 (cf. p. 82, 155).

[6] David ADAMS, Lucia ALTUCCI, Stylianos E ANTONARAKIS et al. « BLUEPRINT to decode the epigenetic signature written in blood ». In : *Nature biotechnology* 30.3 (2012), p. 224-226 (cf. p. 82).

[7] Andrew ADEY, Hilary G MORRISON, Xu XUN et al. « Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition ». In : *Genome biology* 11 (2010), p. 1-17 (cf. p. 67).

[8] A ALMER et W HÖRZ. « Nuclease hypersensitive regions with adjacent positioned nucleosomes mark the gene boundaries of the PHO5/PHO3 locus in yeast. » In : *The EMBO journal* 5.10 (1986), p. 2681-2687 (cf. p. 41).

[9] Joanna S AMBERGER, Carol A BOCCHINI, Alan F SCOTT et al. « OMIM. org : leveraging knowledge across phenotype–gene relationships ». In : *Nucleic acids research* 47.D1 (2019), p. D1038-D1043 (cf. p. 83).

[10] Haley M AMEMIYA, Anshul KUNDAJE et Alan P BOYLE. « The ENCODE blacklist : identification of problematic regions of the genome ». In : *Scientific reports* 9.1 (2019), p. 9354 (cf. p. 145).

[11] Simon ANDREWS et al. *FastQC : a quality control tool for high throughput sequence data*. 2010 (cf. p. 71).

[12] Cosmas D ARNOLD, Daniel GERLACH, Christoph STELZER et al. « Genome-wide quantitative enhancer activity maps identified by STARR-seq ». In : *Science* 339.6123 (2013), p. 1074-1077 (cf. p. 45).

[13] Raymond K AUERBACH, Bin CHEN et Atul J BUTTE. « Relating genes to function : identifying enriched transcription factors using the ENCODE ChIP-Seq significance tool ». In : *Bioinformatics* 29.15 (2013), p. 1922-1924 (cf. p. 155).

[14] Timothy L BAILEY, Mikael BODEN, Fabian A BUSKE et al. « MEME SUITE : tools for motif discovery and searching ». In : *Nucleic acids research* 37.suppl_2 (2009), W202-W208 (cf. p. 154).

[15] Timothy L BAILEY, James JOHNSON, Charles E GRANT et al. « The MEME suite ». In : *Nucleic acids research* 43.W1 (2015), W39-W49 (cf. p. 158).

[16] Amos BAIROCH. « The cellosaurus, a cell-line knowledge resource ». In : *Journal of biomolecular techniques : JBT* 29.2 (2018), p. 25 (cf. p. 119).

[17] Julian BANERJI, Laura OLSON et Walter SCHAFFNER. « A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes ». In : *Cell* 33.3 (1983), p. 729-740 (cf. p. 39).

[18] Julian BANERJI, Sandro RUSCONI et Walter SCHAFFNER. « Expression of a β-globin gene is enhanced by remote SV40 DNA sequences ». In : *Cell* 27.2 (1981), p. 299-308 (cf. p. 39).

[19] Gráinne BARKESS et Adam G WEST. « Chromatin insulator elements : establishing barriers to set heterochromatin boundaries ». In : *Epigenomics* 4.1 (2012), p. 67-80 (cf. p. 46).

[20] Artem BARSKI, Suresh CUDDAPAH, Kairong CUI et al. « High-resolution profiling of histone methylations in the human genome ». In : *Cell* 129.4 (2007), p. 823-837 (cf. p. 34, 60, 103, 188, 219).

[21] Sofia BATTAGLIA, Kevin DONG, Jingyi WU et al. « Long-range phasing of dynamic, tissue-specific and allele-specific regulatory elements ». In : *Nature Genetics* 54.10 (2022), p. 1504-1513 (cf. p. 142).

[22] Adam C BELL, Adam G WEST et Gary FELSENFELD. « Insulators and boundaries : versatile regulatory elements in the eukaryotic genome ». In : *Science* 291.5503 (2001), p. 447-450 (cf. p. 46).

[23] Paolo BENATTI, Diletta DOLFINI, Alessandra VIGANO et al. « Specific inhibition of NF-Y subunits triggers different cell proliferation defects ». In : *Nucleic acids research* 39.13 (2011), p. 5356-5368 (cf. p. 173).

[24] John M BENNETT, Daniel CATOVSKY, Marie-Theregse DANIEL et al. « Proposals for the classification of the acute leukaemias French-American-British (FAB) co-operative group ». In : *British journal of haematology* 33.4 (1976), p. 451-458 (cf. p. 96).

[25] Christophe BENOIST et Pierre CHAMBON. « In vivo sequence requirements of the SV40 early promoter region ». In : *Nature* 290.5804 (1981), p. 304-310 (cf. p. 39).

[26] Dennis A BENSON, Mark CAVANAUGH, Karen CLARK et al. « GenBank ». In : *Nucleic acids research* 46.Database issue (2018), p. D41 (cf. p. 83).

[27] David R BENTLEY, Shankar BALASUBRAMANIAN, Harold P SWERDLOW et al. « Accurate whole human genome sequencing using reversible terminator chemistry ». In : *nature* 456.7218 (2008), p. 53-59 (cf. p. 58).

[28] Charlotte A BERKES, Donald A BERGSTROM, Bennett H PENN et al. « Pbx marks genes for activation by MyoD indicating a role for a homeodomain protein in establishing myogenic potential ». In : *Molecular cell* 14.4 (2004), p. 465-477 (cf. p. 53).

[29] William J BLACK, Dimitrios STAGOS, Satori A MARCHITTI et al. « Human aldehyde dehydrogenase genes : alternatively spliced transcriptional variants and their suggested nomenclature ». In : *Pharmacogenetics and genomics* 19.11 (2009), p. 893-902 (cf. p. 100).

[30] Yuval BLAT et Nancy KLECKNER. « Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region ». In : *Cell* 98.2 (1999), p. 249-259 (cf. p. 60).

[31] Carles A BOIX, Benjamin T JAMES, Yongjin P PARK et al. « Regulatory genomic circuitry of human disease loci by integrative epigenomics ». In : *Nature* 590.7845 (2021), p. 300-307 (cf. p. 217).

[32] Dan BOLSER, Daniel M STAINES, Emily PRITCHARD et al. « Ensembl plants : integrating tools for visualizing, mining, and analyzing plant genomics data ». In : *Plant bioinformatics : Methods and protocols* (2016), p. 115-140 (cf. p. 119).

[33] Beatrice BORSARI, Pablo VILLEGAS-MIRÓN, Siélvia PÉREZ-LLUCH et al. « Enhancers with tissue-specific activity are enriched in intronic regions ». In : *Genome research* 31.8 (2021), p. 1325-1336 (cf. p. 141, 144).

[34] Guillaume BOURQUE, Kathleen H BURNS, Mary GEHRING et al. « Ten things you should know about transposable elements ». In : *Genome biology* 19 (2018), p. 1-12 (cf. p. 86).

[35] Guillaume BOURQUE, Bernard LEONG, Vinsensius B VEGA et al. « Evolution of the mammalian transcription factor binding repertoire via transposable elements ». In : *Genome research* 18.11 (2008), p. 1752-1762 (cf. p. 92, 148, 162, 166, 167, 169, 171, 178-180, 183, 224, 225, 227).

[36] Alan P BOYLE, Sean DAVIS, Hennady P SHULHA et al. « High-resolution mapping and characterization of open chromatin across the genome ». In : *Cell* 132.2 (2008), p. 311-322 (cf. p. 65, 66, 189, 219).

[37] Roy J BRITTEN et Eric H DAVIDSON. « Gene Regulation for Higher Cells : A Theory : New facts regarding the organization of the genome provide clues to the nature of gene regulation. » In : *Science* 165.3891 (1969), p. 349-357 (cf. p. 25).

[38] Roy J BRITTEN et Eric H DAVIDSON. « Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty ». In : *The Quarterly review of biology* 46.2 (1971), p. 111-138 (cf. p. 183).

[39] Nigel P BROWN, Christophe LEROY et Chris SANDER. « MView : a web-compatible database search or multiple alignment viewer. » In : *Bioinformatics (Oxford, England)* 14.4 (1998), p. 380-381 (cf. p. 161).

[40] Jason D BUENROSTRO, Paul G GIRESI, Lisa C ZABA et al. « Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position ». In : *Nature methods* 10.12 (2013), p. 1213-1218 (cf. p. 67, 189, 219).

[41] Jason D BUENROSTRO, Beijing WU, Ulrike M LITZENBURGER et al. « Single-cell chromatin accessibility reveals principles of regulatory variation ». In : *Nature* 523.7561 (2015), p. 486-490 (cf. p. 67, 221).

[42] Michael BULGER et Mark GROUDINE. « Functional and mechanistic diversity of distal transcription enhancers ». In : *Cell* 144.3 (2011), p. 327-339 (cf. p. 39).

[43] Stephen K BURLEY, Helen M BERMAN, Gerard J KLEYWEGT et al. « Protein Data Bank (PDB) : the single global macromolecular structure archive ». In : *Protein crystallography : methods and protocols* (2017), p. 627-641 (cf. p. 83).

[44] Michael BURROWS. « A block-sorting lossless data compression algorithm ». In : *SRC Research Report, 124* (1994) (cf. p. 72).

[45] Diego CALDERON, Michelle LT NGUYEN, Anja MEZGER et al. « Landscape of stimulation-responsive chromatin across diverse human immune cells ». In : *Nature genetics* 51.10 (2019), p. 1494-1505 (cf. p. 221).

[46] Jaime A CASTRO-MONDRAGON, Rafael RIUDAVETS-PUIG, Ieva RAULUSEVICIUTE et al. « JASPAR 2022 : the 9th release of the open-access database of transcription factor binding profiles ». In : *Nucleic acids research* 50.D1 (2022), p. D165-D173 (cf. p. 77, 145, 154, 261).

[47] Ethan CERAMI, Jianjiong GAO, Ugur DOGRUSOZ et al. « The cBio cancer genomics portal : an open platform for exploring multidimensional cancer genomics data ». In : *Cancer discovery* 2.5 (2012), p. 401-404 (cf. p. 101).

[48] Chao CHENG, Renqiang MIN et Mark GERSTEIN. « TIP : a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles ». In : *Bioinformatics* 27.23 (2011), p. 3221-3227 (cf. p. 223).

[49] Adithya CHENNAMADHAVUNI, Varun LYENGAR et Alex SHIMANOVSKY. « Continuing Education Activity ». In : () (cf. p. 95).

[50] Edward B CHUONG, Nels C ELDE et Cédric FESCHOTTE. « Regulatory evolution of innate immunity through co-option of endogenous retroviruses ». In : *Science* 351.6277 (2016), p. 1083-1087 (cf. p. 92, 180, 226).

[51] Emily CLOUGH et Tanya BARRETT. « The gene expression omnibus database ». In : *Statistical Genomics : Methods and Protocols* (2016), p. 93-110 (cf. p. 83).

[52] Andrew B CONLEY, Jittima PIRIYAPONGSA et I King JORDAN. « Retroviral promoters in the human genome ». In : *Bioinformatics* 24.14 (2008), p. 1563-1567 (cf. p. 92).

[53] ENCODE Project CONSORTIUM et al. « An integrated encyclopedia of DNA elements in the human genome ». In : *Nature* 489.7414 (2012), p. 57 (cf. p. 52, 65, 77).

[54] ENCODE Project CONSORTIUM et al. « Expanded encyclopaedias of DNA elements in the human and mouse genomes ». In : *Nature* 583.7818 (2020), p. 699-710 (cf. p. 217).

[55] FlyBase CONSORTIUM. « The FlyBase database of the Drosophila genome projects and community literature ». In : *Nucleic acids research* 31.1 (2003), p. 172-175 (cf. p. 119).

[56] Gene Ontology CONSORTIUM. « The gene ontology resource : 20 years and still GOing strong ». In : *Nucleic acids research* 47.D1 (2019), p. D330-D338 (cf. p. 83).

[57] UniProt CONSORTIUM. « UniProt : a worldwide hub of protein knowledge ». In : *Nucleic acids research* 47.D1 (2019), p. D506-D515 (cf. p. 83).

[58] Ana CONTENTE, Alexandra DITTMER, Manuela C KOCH et al. « A polymorphic microsatellite that mediates induction of PIG3 by p53 ». In : *Nature genetics* 30.3 (2002), p. 315-320 (cf. p. 228).

[59] Françoise COUSTRY, Qianghua HU, Benoit DE CROMBRUGGHE et al. « CBF/NF-Y Functions Both in Nucleosomal Disruption and Transcription Activation of the Chromatin-assembled Topoisomerase IIα Promoter : TRANSCRIPTION ACTIVATION BY CBF/NF-Y IN CHROMATIN IS DEPENDENT ON THE PROMOTER STRUCTURE ». In : *Journal of Biological Chemistry* 276.44 (2001), p. 40621-40630 (cf. p. 173).

[60] Francis H CRICK. « On protein synthesis ». In : *Symp Soc Exp Biol.* T. 12. 138-63. 1958, p. 8 (cf. p. 28).

[61] Fiona CUNNINGHAM, James E ALLEN, Jamie ALLEN et al. « Ensembl 2022 ». In : *Nucleic acids research* 50.D1 (2022), p. D988-D995 (cf. p. 82, 83).

[62] Darren A CUSANOVICH, Riza DAZA, Andrew ADEY et al. « Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing ». In : *Science* 348.6237 (2015), p. 910-914 (cf. p. 221).

[63] Darren A CUSANOVICH, Bryan PAVLOVIC, Jonathan K PRITCHARD et al. « The functional consequences of variation in transcription factor binding ». In : *PLoS genetics* 10.3 (2014), e1004226 (cf. p. 52).

[64] Garrett M DANCIK, Ioannis F VOUTSAS et Spiros VLAHOPOULOS. « Aldehyde Dehydrogenase Enzyme Functions in Acute Leukemia Stem Cells ». In : *Frontiers in Bioscience-Scholar* 14.1 (2022), p. 8 (cf. p. 186).

[65] Lan TM DAO, Ariel O GALINDO-ALBARRÁN, Jaime A CASTRO-MONDRAGON et al. « Genome-wide characterization of mammalian promoters with distal enhancer functions ». In : *Nature genetics* 49.7 (2017), p. 1073-1081 (cf. p. 36).

[66] Gretchen J DARLINGTON, Sarah E ROSS et Ormond A MACDOUGALD. « The role of C/EBP genes in adipocyte differentiation ». In : *Journal of Biological Chemistry* 273.46 (1998), p. 30057-30060 (cf. p. 175).

[67] Robert L DAVIS, Harold WEINTRAUB et Andrew B LASSAR. « Expression of a single transfected cDNA converts fibroblasts to myoblasts ». In : *Cell* 51.6 (1987), p. 987-1000 (cf. p. 50).

[68] Job DEKKER, Karsten RIPPE, Martijn DEKKER et al. « Capturing chromosome conformation ». In : *science* 295.5558 (2002), p. 1306-1311 (cf. p. 45, 69).

[69] Paolo DI TOMMASO, Maria CHATZOU, Evan W FLODEN et al. « Nextflow enables reproducible computational workflows ». In : *Nature biotechnology* 35.4 (2017), p. 316-319 (cf. p. 78).

[70] Jesse R DIXON, David U GORKIN et Bing REN. « Chromatin domains : the unit of chromosome organization ». In : *Molecular cell* 62.5 (2016), p. 668-680 (cf. p. 55).

[71] Hartmut DÖHNER, Elihu H ESTEY, Sergio AMADORI et al. « Diagnosis and management of acute myeloid leukemia in adults : recommendations from an international expert panel, on behalf of the European LeukemiaNet ». In : *Blood, The Journal of the American Society of Hematology* 115.3 (2010), p. 453-474 (cf. p. 98).

[72] Hartmut DÖHNER, Andrew H WEI, Frederick R APPELBAUM et al. « Diagnosis and management of AML in adults : 2022 recommendations from an international expert panel on behalf of the ELN ». In : *Blood, The Journal of the American Society of Hematology* 140.12 (2022), p. 1345-1377 (cf. p. 98).

[73] Hartmut DÖHNER, Daniel J WEISDORF et Clara D BLOOMFIELD. « Acute myeloid leukemia ». In : *New England Journal of Medicine* 373.12 (2015), p. 1136-1152 (cf. p. 96, 186).

[74] Maureen J DONLIN. « Using the generic genome browser (GBrowse) ». In : *Current protocols in bioinformatics* 28.1 (2009), p. 9-9 (cf. p. 78).

[75] Dale DORSETT et Matthias MERKENSCHLAGER. « Cohesin at active genes : a unifying theme for cohesin and gene expression from model organisms to humans ». In : *Current opinion in cell biology* 25.3 (2013), p. 327-333 (cf. p. 44, 45).

[76] Robert C EDGAR. « MUSCLE : multiple sequence alignment with high accuracy and high throughput ». In : *Nucleic acids research* 32.5 (2004), p. 1792-1797 (cf. p. 152, 161).

[77] Hans ELLEGREN. « Heterogeneous mutation processes in human microsatellite DNA sequences ». In : *Nature genetics* 24.4 (2000), p. 400-402 (cf. p. 85, 228).

[78] Janan T EPPIG. « Mouse genome informatics (MGI) resource : genetic, genomic, and biological knowledgebase for the laboratory mouse ». In : *ILAR journal* 58.1 (2017), p. 17-41 (cf. p. 119).

[79] Keith ERICKSON. « The jukes-cantor model of molecular evolution ». In : *Primus* 20.5 (2010), p. 438-445 (cf. p. 162).

[80] Jason ERNST et Manolis KELLIS. « ChromHMM : automating chromatin-state discovery and characterization ». In : *Nature methods* 9.3 (2012), p. 215-216 (cf. p. 142).

[81] Jason ERNST, Pouya KHERADPOUR, Tarjei S MIKKELSEN et al. « Mapping and analysis of chromatin state dynamics in nine human cell types ». In : *Nature* 473.7345 (2011), p. 43-49 (cf. p. 42, 144).

[82] Ahmed ESSAGHIR, Federica TOFFALINI, Laurent KNOOPS et al. « Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data ». In : *Nucleic acids research* 38.11 (2010), e120-e120 (cf. p. 223).

[83] Brent EWING, LaDeana HILLIER, Michael C WENDL et al. « Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment ». In : *Genome research* 8.3 (1998), p. 175-185 (cf. p. 71).

[84] Hao FAN et Jia-You CHU. « A brief review of short tandem repeat mutation ». In : *Genomics, proteomics & bioinformatics* 5.1 (2007), p. 7-14 (cf. p. 85).

[85] EA FEINGOLD et L PACHTER. « The ENCODE (ENCyclopedia of DNA elements) project ». In : *Science* 306.5696 (2004), p. 636-640 (cf. p. 104, 105).

[86] Anthony P FEJES, Gordon ROBERTSON, Mikhail BILENKY et al. « FindPeaks 3.1 : a tool for identifying areas of enrichment from massively parallel short-read sequencing technology ». In : *Bioinformatics* 24.15 (2008), p. 1729-1730 (cf. p. 72).

[87] Xosé M FERNÁNDEZ et Ewan BIRNEY. « Ensembl genome browser ». In : *Vogel and Motulsky's Human Genetics* (2010), p. 923-939 (cf. p. 78).

[88] Paolo FERRAGINA et Giovanni MANZINI. « Opportunistic data structures with applications ». In : *Proceedings 41st annual symposium on foundations of computer science*. IEEE. 2000, p. 390-398 (cf. p. 72).

[89] Cédric FESCHOTTE. « Transposable elements and the evolution of regulatory networks ». In : *Nature Reviews Genetics* 9.5 (2008), p. 397-405 (cf. p. 92, 148, 183).

[90] Cédric FESCHOTTE, Ning JIANG et Susan R WESSLER. « Plant transposable elements : where genetics meets genomics ». In : *Nature Reviews Genetics* 3.5 (2002), p. 329-341 (cf. p. 91).

[91] Cédric FESCHOTTE et Ellen J PRITHAM. « DNA transposons and the evolution of eukaryotic genomes ». In : *Annu. Rev. Genet.* 41 (2007), p. 331-368 (cf. p. 86).

[92] Stephanie Feupe FOTSING, Jonathan MARGOLIASH, Catherine WANG et al. « The impact of short tandem repeat variation on gene expression ». In : *Nature genetics* 51.11 (2019), p. 1652-1659 (cf. p. 153).

[93] Adam FRANKISH, Mark DIEKHANS, Irwin JUNGREIS et al. « GENCODE 2021 ». In : *Nucleic acids research* 49.D1 (2021), p. D916-D923 (cf. p. 82).

[94] Geoffrey FUDENBERG, Maxim IMAKAEV, Carolyn LU et al. « Formation of chromosomal domains by loop extrusion ». In : *Cell reports* 15.9 (2016), p. 2038-2049 (cf. p. 44, 47, 55, 219).

[95] Anthony V FURANO. « The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons ». In : (2000) (cf. p. 175, 181).

[96] George GAMOW. *On information transfer from nucleic acids to proteins*. I kommission hos E. Munksgaard, 1955 (cf. p. 26).

[97] Sunil GANGADHARAN, Loris MULARONI, Jennifer FAIN-THORNTON et al. « DNA transposon Hermes inserts into DNA in nucleosome-free regions in vivo ». In : *Proceedings of the National Academy of Sciences* 107.51 (2010), p. 21966-21972 (cf. p. 67).

[98] Jianjiong GAO, Bülent Arman AKSOY, Ugur DOGRUSOZ et al. « Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal ». In : *Science signaling* 6.269 (2013), pl1-pl1 (cf. p. 101).

[99] Margarita GARCIA-HERNANDEZ, Tanya BERARDINI, Guanghong CHEN et al. « TAIR : a resource for integrated Arabidopsis data ». In : *Functional & integrative genomics* 2 (2002), p. 239-253 (cf. p. 119).

[100] Annie GAREL et Richard AXEL. « Selective digestion of transcriptionally active ovalbumin genes from oviduct nuclei. » In : *Proceedings of the National Academy of Sciences* 73.11 (1976), p. 3966-3970 (cf. p. 65).

[101] Frank GEBHARDT, Kurt S ZÄNKER et Burkhard BRANDT. « Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1 ». In : *Journal of Biological Chemistry* 274.19 (1999), p. 13176-13180 (cf. p. 228).

[102] C GENOME. « Genomic and epigenomic landscapes of adult de novo acute myeloid 310 leukemia ». In : *N Engl J Med* 368.2059-2074 (2013), p. 311 (cf. p. 97).

[103] Marc GILLESPIE, Bijay JASSAL, Ralf STEPHAN et al. « The reactome pathway knowledgebase 2022 ». In : *Nucleic acids research* 50.D1 (2022), p. D687-D692 (cf. p. 83).

[104] GB GOLDING. « Simple sequence is abundant in eukaryotic proteins ». In : *Protein science* 8.6 (1999), p. 1358-1361 (cf. p. 153).

[105] Manfred G GRABHERR, Brian J HAAS, Moran YASSOUR et al. « Full-length transcriptome assembly from RNA-Seq data without a reference genome ». In : *Nature biotechnology* 29.7 (2011), p. 644-652 (cf. p. 59).

[106] Charles E GRANT, Timothy L BAILEY et William Stafford NOBLE. « FIMO : scanning for occurrences of a given motif ». In : *Bioinformatics* 27.7 (2011), p. 1017-1018 (cf. p. 77, 152, 158).

[107] Philipp A GREIF, Nikola P KONSTANDIN, Klaus H METZELER et al. « RUNX1 mutations in cytogenetically normal acute myeloid leukemia are associated with a poor prognosis and up-regulation of lymphoid genes ». In : *Haematologica* 97.12 (2012), p. 1909 (cf. p. 215).

[108] Aurélien GRIFFON, Quentin BARBIER, Jordi DALINO et al. « Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape ». In : *Nucleic acids research* 43.4 (2015), e27-e27 (cf. p. 105).

[109] David GRIMWADE, Robert K HILLS, Anthony V MOORMAN et al. « Refinement of cytogenetic classification in acute myeloid leukemia : determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the United Kingdom Medical Research Council trials ». In : *Blood, The Journal of the American Society of Hematology* 116.3 (2010), p. 354-365 (cf. p. 96).

[110] Fabian GRUBERT, Rohith SRIVAS, Damek V SPACEK et al. « Landscape of cohesin-mediated chromatin loops in the human genome ». In : *Nature* 583.7818 (2020), p. 737-743 (cf. p. 218).

[111] Peter GRUSS, Ravi DHAR et George KHOURY. « Simian virus 40 tandem repeated sequences as an element of the early promoter. » In : *Proceedings of the National Academy of Sciences* 78.2 (1981), p. 943-947 (cf. p. 39).

[112] Ravi GUPTA, Priyankara WIKRAMASINGHE, Anirban BHATTACHARYYA et al. « Annotation of gene promoters by integrative data-mining of ChIP-seq Pol-II enrichment data ». In : *BMC bioinformatics* 11.1 (2010), p. 1-9 (cf. p. 219).

[113] Vanja HABERLE et Alexander STARK. « Eukaryotic core promoters and the functional basis of transcription initiation ». In : *Nature reviews Molecular cell biology* 19.10 (2018), p. 621-637 (cf. p. 34, 188).

[114] Wilfried HAERTY et G Brian GOLDING. « Low-complexity sequences and single amino acid repeats : not just “junk” peptide sequences ». In : *Genome* 53.10 (2010), p. 753-762 (cf. p. 153).

[115] Marc S HALFON, Ana CARMENA, Stephen GISSELBRECHT et al. « Ras pathway specificity is determined by the integration of multiple signal-activated and tissue-restricted transcription factors ». In : *Cell* 103.1 (2000), p. 63-74 (cf. p. 40).

[116] Fayrouz HAMMAL, Pierre de LANGEN, Aurélie BERGON et al. « ReMap 2022 : a database of human, mouse, drosophila and arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments ». In : *Nucleic Acids Research* 50.D1 (2022), p. D316-D325 (cf. p. 106).

[117] Xiaoping HAN, Ziming ZHOU, Lijiang FEI et al. « Construction of a human cell landscape at single-cell level ». In : *Nature* 581.7808 (2020), p. 303-309 (cf. p. 59).

[118] Robert J HASELBECK, Ines HOFFMANN et Gregg DUESTER. « Distinct functions for Aldh1 and Raldh2 in the control of ligand production for embryonic retinoid signaling pathways ». In : *Developmental genetics* 25.4 (1999), p. 353-364 (cf. p. 100).

[119] Qiye HE, Jeff JOHNSTON et Julia ZEITLINGER. « ChIP-nexus enables improved detection of in vivo transcription factor binding footprints ». In : *Nature biotechnology* 33.4 (2015), p. 395-401 (cf. p. 63).

[120] Nathaniel D HEINTZMAN, Rhona K STUART, Gary HON et al. « Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome ». In : *Nature genetics* 39.3 (2007), p. 311-318 (cf. p. 44).

[121] Clara HERMANT et Maria-Elena TORRES-PADILLA. « TFs for TEs : the transcription factor repertoire of mammalian transposable elements ». In : *Genes & Development* 35.1-2 (2021), p. 22-39 (cf. p. 150).

[122] Jay R HESSELBERTH, Xiaoyu CHEN, Zhihong ZHANG et al. « Global mapping of protein-DNA interactions in vivo by digital genomic footprinting ». In : *Nature methods* 6.4 (2009), p. 283-289 (cf. p. 65, 66, 189, 219).

[123] Hopi E HOEKSTRA et Jerry A COYNE. « The locus of evolution : evo devo and the genetics of adaptation ». In : *Evolution* 61.5 (2007), p. 995-1016 (cf. p. 222).

[124] Connor A HORTON, Amr M ALEXANDARI, Michael GB HAYES et al. « Short tandem repeats bind transcription factors to tune eukaryotic gene expression ». In : *bioRxiv* (2022), p. 2022-05 (cf. p. 153).

[125] H-T HUANG et LI ZON. « Regulation of stem cells in the zebra fish hematopoietic system ». In : *Cold Spring Harbor Symposia on Quantitative Biology*. T. 73. Cold Spring Harbor Laboratory Press. 2008, p. 111-118 (cf. p. 222).

[126] Maxwell A HUME, Luis A BARRERA, Stephen S GISSELBRECHT et al. « UniPROBE, update 2015 : new tools and content for the online database of protein-binding microarray data on protein–DNA interactions ». In : *Nucleic acids research* 43.D1 (2015), p. D117-D122 (cf. p. 77).

[127] Tony HUNTER et Michael KARIN. « The regulation of transcription by phosphorylation ». In : *Cell* 70.3 (1992), p. 375-387 (cf. p. 28).

[128] Alexandra IOURANOVA, Delphine GRUN, Tamara ROSSY et al. « KRAB zinc finger protein ZNF676 controls the transcriptional influence of LTR12-related endogenous retrovirus sequences ». In : *Mobile Dna* 13.1 (2022), p. 1-17 (cf. p. 173, 183).

[129] François JACOB et Jacques MONOD. « On the regulation of gene activity ». In : *Cold Spring Harbor symposia on quantitative biology*. T. 26. Cold Spring Harbor Laboratory Press. 1961, p. 193-211 (cf. p. 29).

[130] Pierre-Étienne JACQUES, Justin JEYAKANI et Guillaume BOURQUE. « The majority of primate-specific regulatory sequences are derived from transposable elements ». In : *PLoS genetics* 9.5 (2013), e1003504 (cf. p. 92, 149, 151, 181, 183, 184).

[131] Hongkai JI, Hui JIANG, Wenxiu MA et al. « An integrated software system for analyzing ChIP-chip and ChIP-seq data ». In : *Nature biotechnology* 26.11 (2008), p. 1293-1300 (cf. p. 72).

[132] Hongshan JIANG, Rong LEI, Shou-Wei DING et al. « Skewer : a fast and accurate adapter trimmer for next-generation sequencing paired-end reads ». In : *BMC bioinformatics* 15 (2014), p. 1-12 (cf. p. 71).

[133] Jiayue-Clara JIANG, Joseph A ROTHNAGEL et Kyle R UPTON. « Integrated transcription factor profiling with transcriptome analysis identifies L1PA2 transposons as global regulatory modulators in a breast cancer model ». In : *Scientific Reports* 11.1 (2021), p. 8083 (cf. p. 170, 226).

[134] Jiayue-Clara JIANG, Joseph A ROTHNAGEL et Kyle R UPTON. « Widespread exaptation of L1 transposons for transcription factor binding in breast cancer ». In : *International Journal of Molecular Sciences* 22.11 (2021), p. 5625 (cf. p. 150, 169).

[135] Chunyuan JIN, Chongzhi ZANG, Gang WEI et al. « H3. 3/H2A. Z double variant–containing nucleosomes mark'nucleosome-free regions' of active promoters and other regulatory regions ». In : *Nature genetics* 41.8 (2009), p. 941-945 (cf. p. 34).

[136] Wenfei JIN, Qingsong TANG, Mimi WAN et al. « Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples ». In : *Nature* 528.7580 (2015), p. 142-146 (cf. p. 221).

[137] David S JOHNSON, Ali MORTAZAVI, Richard M MYERS et al. « Genome-wide mapping of in vivo protein-DNA interactions ». In : *Science* 316.5830 (2007), p. 1497-1502 (cf. p. 72, 103, 219).

[138] Peter F JOHNSON. « Molecular stop signs : regulation of cell-cycle arrest by C/EBP transcription factors ». In : *Journal of cell science* 118.12 (2005), p. 2545-2555 (cf. p. 175).

[139] Arttu JOLMA, Teemu KIVIOJA, Jarkko TOIVONEN et al. « Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities ». In : *Genome research* 20.6 (2010), p. 861-873 (cf. p. 77).

[140] I King JORDAN, Igor B ROGOZIN, Galina V GLAZKO et al. « Origin of a substantial fraction of human regulatory sequences from transposable elements ». In : *TRENDS in Genetics* 19.2 (2003), p. 68-72 (cf. p. 92, 171).

[141] Raja JOTHI, Suresh CUDDAPAH, Artem BARSKI et al. « Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data ». In : *Nucleic acids research* 36.16 (2008), p. 5221-5231 (cf. p. 72).

[142] Avak KAHVEJIAN, John QUACKENBUSH et John F THOMPSON. « What would you do if you could sequence everything? » In : *Nature biotechnology* 26.10 (2008), p. 1125-1133 (cf. p. 81).

[143] Gizem KALAY et Patricia J WITTKOPP. « Nomadic enhancers : tissue-specific cis-regulatory elements of yellow have divergent genomic positions among Drosophila species ». In : *PLoS Genetics* 6.11 (2010), e1001222 (cf. p. 221).

[144] Minoru KANEHISA. « The KEGG database ». In : *'In Silico'Simulation of Biological Processes : Novartis Foundation Symposium 247*. T. 247. Wiley Online Library. 2002, p. 91-103 (cf. p. 83).

[145] Minoru KANEHISA, Miho FURUMICHI, Mao TANABE et al. « KEGG : new perspectives on genomes, pathways, diseases and drugs ». In : *Nucleic acids research* 45.D1 (2017), p. D353-D361 (cf. p. 83).

[146] Carola KANZ, Philippe ALDEBERT, Nicola ALTHORPE et al. « The EMBL nucleotide sequence database ». In : *Nucleic acids research* 33.suppl_1 (2005), p. D29-D33 (cf. p. 83).

[147] Donna KAROLCHIK, Robert BAERTSCH, Mark DIEKHANS et al. « The UCSC genome browser database ». In : *Nucleic acids research* 31.1 (2003), p. 51-54 (cf. p. 78).

[148] Hiroko KAWATA, Kazuya YAMADA, Zhangfei SHOU et al. « Zinc-fingers and homeoboxes (ZHX) 2, a novel member of the ZHX family, functions as a transcriptional repressor ». In : *Biochemical Journal* 373.3 (2003), p. 747-757 (cf. p. 173).

[149] Tim KEHL, Lara SCHNEIDER, Florian SCHMIDT et al. « RegulatorTrail : a web service for the identification of key transcriptional regulators ». In : *Nucleic Acids Research* 45.W1 (2017), W146-W153 (cf. p. 223).

[150] J KENT. *KentUtils*. 2016 (cf. p. 189).

[151] Peter V KHARCHENKO, Michael Y TOLSTORUKOV et Peter J PARK. « Design and analysis of ChIP-seq experiments for DNA-binding proteins ». In : *Nature biotechnology* 26.12 (2008), p. 1351-1359 (cf. p. 72).

[152] Purvesh KHATRI, Marina SIROTA et Atul J BUTTE. « Ten years of pathway analysis : current approaches and outstanding challenges ». In : *PLoS computational biology* 8.2 (2012), e1002375 (cf. p. 155).

[153] Tae-Kyung KIM, Martin HEMBERG, Jesse M GRAY et al. « Widespread transcription at neuronal activity-regulated enhancers ». In : *Nature* 465.7295 (2010), p. 182-187 (cf. p. 37).

[154] Martin KIRCHER, Patricia HEYN et Janet KELSO. « Addressing challenges in the production and analysis of illumina sequencing data ». In : *BMC genomics* 12 (2011), p. 1-14 (cf. p. 58).

[155] Nikolay KOLESNIKOV, Emma HASTINGS, Maria KEAYS et al. « ArrayExpress update—simplifying data submissions ». In : *Nucleic acids research* 43.D1 (2015), p. D1113-D1116 (cf. p. 83).

[156] Semyon KOLMYKOV, Ivan YEVSHIN, Mikhail KULYASHOV et al. « GTRD : an integrated view of transcription regulation ». In : *Nucleic acids research* 49.D1 (2021), p. D104-D111 (cf. p. 117).

[157] Yong KONG. « Btrim : a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies ». In : *Genomics* 98.2 (2011), p. 152-153 (cf. p. 71).

[158] Johannes KÖSTER et Sven RAHMANN. « Snakemake—a scalable bioinformatics workflow engine ». In : *Bioinformatics* 28.19 (2012), p. 2520-2522 (cf. p. 78).

[159] Sonja K KRÖNUNG, Ulrike BEYER, Maria Luisa CHIARAMONTE et al. « LTR12 promoter activation in a broad range of human tumor cells by HDAC inhibition ». In : *Oncotarget* 7.23 (2016), p. 33484 (cf. p. 173).

[160] Felix KRUEGER. « Trim Galore : a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries ». In : *URL http ://www. bioinformatics. babraham. ac. uk/projects/trim_galore/.(Date of access : 28/04/2016)* (2012) (cf. p. 71).

[161] Michelle M KUDRON, Alec VICTORSEN, Louis GEVIRTZMAN et al. « The ModERN resource : genome-wide binding profiles for hundreds of Drosophila and Caenorhabditis elegans transcription factors ». In : *Genetics* 208.3 (2018), p. 937-949 (cf. p. 124).

[162] Galih KUNARSO, Na-Yu CHIA, Justin JEYAKANI et al. « Transposable elements have rewired the core regulatory network of human embryonic stem cells ». In : *Nature genetics* 42.7 (2010), p. 631-634 (cf. p. 92, 149, 169, 178, 183, 224, 226, 227).

[163] Roadmap Epigenomics Consortium Integrative analysis coordination KUNDAJE ANSHUL 1 2 3 MEULEMAN WOUTER 1 2 ERNST JASON 1 2 4 BILENKY MISHA 5, Scientific program management CHADWICK LISA H. 53 et Principal investigators BERNSTEIN BRADLEY E. 2 26 42 COSTELLO JOSEPH F. 14 ECKER JOSEPH R. 9 HIRST MARTIN 5 18 MEISSNER ALEXANDER 2 6 MILOSAVLJEVIC ALEKSANDAR 7 REN BING 8 13 STAMATOYANNOPOULOS JOHN A. 10 WANG TING 21 KELLIS MANOLIS 1 2. « Integrative analysis of 111 reference human epigenomes ». In : *Nature* 518.7539 (2015), p. 317-330 (cf. p. 65, 217).

[164] ML KUNITZ. « Crystalline desoxyribonuclease : I. Isolation and general properties spectrophotometric method for the measurement of desoxyribonuclease activity ». In : *The Journal of general physiology* 33.4 (1950), p. 349-362 (cf. p. 65).

[165] Bryan R LAJOIE, Job DEKKER et Noam KAPLAN. « The Hitchhiker's guide to Hi-C analysis : practical guidelines ». In : *Methods* 72 (2015), p. 65-75 (cf. p. 69).

[166] Stephen G LANDT, Georgi K MARINOV, Anshul KUNDAJE et al. « ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia ». In : *Genome research* 22.9 (2012), p. 1813-1831 (cf. p. 77, 104, 105).

[167] Ben LANGMEAD et Steven L SALZBERG. « Fast gapped-read alignment with Bowtie 2 ». In : *Nature methods* 9.4 (2012), p. 357-359 (cf. p. 72).

[168] Brian T LEE, Galt P BARBER, Anna BENET-PAGÈS et al. « The UCSC genome browser database : 2022 update ». In : *Nucleic Acids Research* 50.D1 (2022), p. D1115-D1122 (cf. p. 153).

[169] Wes LEE, Alois HASLINGER, Michael KARIN et al. « Activation of transcription by two factors that bind promoter and enhancer sequences of the human metallothionein gene and SV40 ». In : *Nature* 325.6102 (1987), p. 368-372 (cf. p. 40).

[170] Boris LENHARD, Albin SANDELIN et Piero CARNINCI. « Metazoan promoters : emerging characteristics and insights into transcriptional regulation ». In : *Nature Reviews Genetics* 13.4 (2012), p. 233-245 (cf. p. 34).

[171] Heng LI et Richard DURBIN. « Fast and accurate short read alignment with Burrows–Wheeler transform ». In : *bioinformatics* 25.14 (2009), p. 1754-1760 (cf. p. 72).

[172] Heng LI, Bob HANDSAKER, Alec WYSOKER et al. « The sequence alignment/map format and SAMtools ». In : *bioinformatics* 25.16 (2009), p. 2078-2079 (cf. p. 72).

[173] Kun LIANG et Sündüz KELEŞ. « Normalization of ChIP-seq data with control ». In : *BMC bioinformatics* 13 (2012), p. 1-10 (cf. p. 62).

[174] Peng LIANG et Arthur B PARDEE. « Analysing differential gene expression in cancer ». In : *Nature Reviews Cancer* 3.11 (2003), p. 869-876 (cf. p. 223).

[175] Erez LIEBERMAN-AIDEN, Nynke L VAN BERKUM, Louise WILLIAMS et al. « Comprehensive mapping of long-range interactions reveals folding principles of the human genome ». In : *science* 326.5950 (2009), p. 289-293 (cf. p. 45, 55, 219).

[176] Charles Y LIN, Serap ERKEK, Yiai TONG et al. « Active medulloblastoma enhancers reveal subgroup-specific cellular origins ». In : *Nature* 530.7588 (2016), p. 57-62 (cf. p. 45).

[177] Ryan LISTER, Ronan C O'MALLEY, Julian TONTI-FILIPPINI et al. « Highly integrated single-base resolution maps of the epigenome in Arabidopsis ». In : *Cell* 133.3 (2008), p. 523-536 (cf. p. 59).

[178] Qian LIU, Yao TONG et Kai WANG. « Genome-wide detection of short tandem repeat expansions by long-read sequencing ». In : *BMC bioinformatics* 21 (2020), p. 1-15 (cf. p. 153).

[179] Tao LIU, Jorge A ORTIZ, Len TAING et al. « Cistrome : an integrative platform for transcriptional regulation studies ». In : *Genome biology* 12.8 (2011), p. 1-10 (cf. p. 105).

[180] John LONSDALE, Jeffrey THOMAS, Mike SALVATORE et al. « The genotype-tissue expression (GTEx) project ». In : *Nature genetics* 45.6 (2013), p. 580-585 (cf. p. 138).

[181] Jakob LOVÉN, Heather A HOKE, Charles Y LIN et al. « Selective inhibition of tumor oncogenes by disruption of super-enhancers ». In : *Cell* 153.2 (2013), p. 320-334 (cf. p. 45).

[182] Bob LOWENBERG, James R DOWNING et Alan BURNETT. « Acute myeloid leukemia ». In : *New England Journal of Medicine* 341.14 (1999), p. 1051-1062 (cf. p. 96).

[183] Nicholas M LUSCOMBE, Susan E AUSTIN, Helen M BERMAN et al. « An overview of the structures of protein-DNA complexes ». In : *Genome biology* 1 (2000), p. 1-37 (cf. p. 51).

[184] Vincent J LYNCH, Robert D LECLERC, Gemma MAY et al. « Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals ». In : *Nature genetics* 43.11 (2011), p. 1154-1159 (cf. p. 171).

[185] Vincent J LYNCH, Mauris C NNAMANI, Aurélie KAPUSTA et al. « Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy ». In : *Cell reports* 10.4 (2015), p. 551-561 (cf. p. 226).

[186] Donna MAGLOTT, Jim OSTELL, Kim D PRUITT et al. « Entrez Gene : gene-centered information at NCBI ». In : *Nucleic acids research* 33.suppl_1 (2005), p. D54-D58 (cf. p. 83).

[187] Shaun MAHONY et B Franklin PUGH. « Protein–DNA binding in high-resolution ». In : *Critical reviews in biochemistry and molecular biology* 50.4 (2015), p. 269-283 (cf. p. 73).

[188] Pascal MAIRE, Jérôme WUARIN et Ueli SCHIBLER. « The role of cis-acting promoter elements in tissue-specific albumin gene expression ». In : *Science* 244.4902 (1989), p. 343-346 (cf. p. 221).

[189] Marcel MARTIN. « Cutadapt removes adapter sequences from high-throughput sequencing reads ». In : *EMBnet. journal* 17.1 (2011), p. 10-12 (cf. p. 71).

[190] Vivien MARX. « The big challenges of big data ». In : *Nature* 498.7453 (2013), p. 255-260 (cf. p. 80).

[191] Kert MÄTLIK, Kaja REDIK et Mart SPEEK. « L1 antisense promoter drives tissue-specific transcription of human genes ». In : *Journal of biomedicine and biotechnology* 2006 (2006) (cf. p. 177).

[192] Vea MATYS, Ellen FRICKE, Robert GEFFERS et al. « TRANSFAC® : transcriptional regulation, from patterns to profiles ». In : *Nucleic acids research* 31.1 (2003), p. 374-378 (cf. p. 77).

[193] Barbara MCCLINTOCK et al. « Mutable loci in maize. » In : *Mutable loci in maize.* (1947) (cf. p. 86).

[194] Barbara MCCLINTOCK. « Controlling elements and the gene ». In : *Cold Spring Harbor symposia on quantitative biology*. T. 21. Cold Spring Harbor Laboratory Press. 1956, p. 197-216 (cf. p. 183).

[195] Barbara MCCLINTOCK. « The significance of responses of the genome to challenge ». In : *Science* 226.4676 (1984), p. 792-801 (cf. p. 84).

[196] Iain W MCKINNELL, Jeff ISHIBASHI, Fabien LE GRAND et al. « Pax7 activates myogenic genes by recruitment of a histone methyltransferase complex ». In : *Nature cell biology* 10.1 (2008), p. 77-84 (cf. p. 53).

[197] Cory Y MCLEAN, Dave BRISTOR, Michael HILLER et al. « GREAT improves functional interpretation of cis-regulatory regions ». In : *Nature biotechnology* 28.5 (2010), p. 495-501 (cf. p. 155).

[198] Carlos A MELO, Jarno DROST, Patrick J WIJCHERS et al. « eRNAs are required for p53-dependent enhancer activity and gene transcription ». In : *Molecular cell* 49.3 (2013), p. 524-535 (cf. p. 40).

[199] Eric M MENDENHALL, Kaylyn E WILLIAMSON, Deepak REYON et al. « Locus-specific editing of histone modifications at endogenous enhancers ». In : *Nature biotechnology* 31.12 (2013), p. 1133-1136 (cf. p. 39).

[200] Wouter MEULEMAN, Alexander MURATOV, Eric RYNES et al. « Index and biological spectrum of human DNase I hypersensitive sites ». In : *Nature* 584.7820 (2020), p. 244-251 (cf. p. 154, 217).

[201] Matthias MEYER et Martin KIRCHER. « Illumina sequencing library preparation for highly multiplexed target capture and sequencing ». In : *Cold Spring Harbor Protocols* 2010.6 (2010), pdb-prot5448 (cf. p. 58).

[202] Tarjei S MIKKELSEN, Manching KU, David B JAFFE et al. « Genome-wide maps of chromatin state in pluripotent and lineage-committed cells ». In : *Nature* 448.7153 (2007), p. 553-560 (cf. p. 42, 103, 219).

[203] Jaina MISTRY, Sara CHUGURANSKY, Lowri WILLIAMS et al. « Pfam : The protein families database in 2021 ». In : *Nucleic acids research* 49.D1 (2021), p. D412-D419 (cf. p. 83).

[204] Celine MOORMAN, Ling V SUN, Junbai WANG et al. « Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster ». In : *Proceedings of the National Academy of Sciences* 103.32 (2006), p. 12027-12032 (cf. p. 52).

[205] Mary MUERS. « The modENCODE guide to the genome ». In : *Nature Reviews Genetics* 12.2 (2011), p. 80-80 (cf. p. 124).

[206] Martıén MUÑOZ-LÓPEZ et José L GARCIÉA-PÉREZ. « DNA transposons : nature and applications in genomics ». In : *Current genomics* 11.2 (2010), p. 115-128 (cf. p. 67, 86, 91).

[207] Takashi NAGANO, Yaniv LUBLING, Tim J STEVENS et al. « Single-cell Hi-C reveals cell-to-cell variability in chromosome structure ». In : *Nature* 502.7469 (2013), p. 59-64 (cf. p. 55).

[208] Jairo NAVARRO GONZALEZ, Ann S ZWEIG, Matthew L SPEIR et al. « The UCSC genome browser database : 2021 update ». In : *Nucleic acids research* 49.D1 (2021), p. D1046-D1057 (cf. p. 83).

[209] Christoph NEUMAYR, Vanja HABERLE, Leonid SEREBRENI et al. « Differential cofactor dependencies define distinct types of human enhancers ». In : *Nature* 606.7913 (2022), p. 406-413 (cf. p. 173).

[210] Thomas A NGUYEN, Richard D JONES, Andrew R SNAVELY et al. « High-throughput functional comparison of promoter and enhancer activities ». In : *Genome research* 26.8 (2016), p. 1023-1033 (cf. p. 36).

[211] Marshall W NIRENBERG et J Heinrich MATTHAEI. « The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides ». In : *Proceedings of the National Academy of Sciences* 47.10 (1961), p. 1588-1602 (cf. p. 26).

[212] Hidenori NISHIHARA. « Retrotransposons spread potential cis-regulatory elements during mammary gland evolution ». In : *Nucleic Acids Research* 47.22 (2019), p. 11551-11562 (cf. p. 163, 260).

[213] Darryl NISHIMURA. « RepeatMasker ». In : *Biotech Software & Internet Report* 1.1-2 (2000), p. 36-39 (cf. p. 153, 173).

[214] Kazuhiro R NITTA, Arttu JOLMA, Yimeng YIN et al. « Conservation of transcription factor binding specificities across 600 million years of bilateria evolution ». In : *elife* 4 (2015), e04837 (cf. p. 48).

[215] Cédric NOTREDAME, Desmond G HIGGINS et Jaap HERINGA. « T-Coffee : A novel method for fast and accurate multiple sequence alignment ». In : *Journal of molecular biology* 302.1 (2000), p. 205-217 (cf. p. 161).

[216] Sergey NURK, Sergey KOREN, Arang RHIE et al. « The complete sequence of a human genome ». In : *Science* 376.6588 (2022), p. 44-53 (cf. p. 81, 84).

[217] Yishai OFRAN, Martin S TALLMAN et Jacob M ROWE. « How I treat acute myeloid leukemia presenting with preexisting comorbidities ». In : *Blood, The Journal of the American Society of Hematology* 128.4 (2016), p. 488-496 (cf. p. 96).

[218] Steven OGBOURNE et Toni M ANTALIS. « Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes ». In : *Biochemical Journal* 331.1 (1998), p. 1-14 (cf. p. 46).

[219] Andrew J OLDFIELD, Pengyi YANG, Amanda E CONWAY et al. « Histone-fold domain protein NF-Y promotes chromatin accessibility for cell type-specific master transcription factors ». In : *Molecular cell* 55.5 (2014), p. 708-722 (cf. p. 173).

[220] John K PACE et Cédric FESCHOTTE. « The evolutionary history of human DNA transposons : evidence for intense activity in the primate lineage ». In : *Genome research* 17.4 (2007), p. 422-432 (cf. p. 162, 180, 182).

[221] Vera PANCALDI, Enrique CARRILLO-DE-SANTA-PAU, Biola Maria JAVIERRE et al. « Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity ». In : *Genome biology* 17.1 (2016), p. 1-19 (cf. p. 36).

[222] Shen PANG, Jens DANNULL, Randhir KABOO et al. « Identification of a positive regulatory element responsible for tissue-specific expression of prostate-specific antigen ». In : *Cancer Research* 57.3 (1997), p. 495-499 (cf. p. 142).

[223] Daniel PANNE. « The enhanceosome ». In : *Current opinion in structural biology* 18.2 (2008), p. 236-242 (cf. p. 40).

[224] Mathilde PARIS, Tommy KAPLAN, Xiao Yong LI et al. « Extensive divergence of transcription factor binding in Drosophila embryos with highly conserved gene expression ». In : *PLoS genetics* 9.9 (2013), e1003748 (cf. p. 52).

[225] Daechan PARK, Yaelim LEE, Gurvani BHUPINDERSINGH et al. « Widespread misinterpretable ChIP-seq bias in yeast ». In : *PloS one* 8.12 (2013), e83506 (cf. p. 62).

[226] Peter J PARK. « ChIP–seq : advantages and challenges of a maturing technology ». In : *Nature reviews genetics* 10.10 (2009), p. 669-680 (cf. p. 60, 72).

[227] Typhaine PAYSAN-LAFOSSE, Matthias BLUM, Sara CHUGURANSKY et al. « InterPro in 2022 ». In : *Nucleic Acids Research* 51.D1 (2023), p. D418-D427 (cf. p. 83).

[228] Erica C PEHRSSON, Mayank NK CHOUDHARY, Vasavi SUNDARAM et al. « The epigenomic landscape of transposable elements across normal human development and anatomy ». In : *Nature communications* 10.1 (2019), p. 5640 (cf. p. 226).

[229] Len A PENNACCHIO, Wendy BICKMORE, Ann DEAN et al. « Enhancers : five essential questions ». In : *Nature Reviews Genetics* 14.4 (2013), p. 288-295 (cf. p. 37).

[230] Rebecca PETRI, Per Ludvik BRATTÅS, Yogita SHARMA et al. « LINE-2 transposable elements are a source of functional human microRNAs and target sites ». In : *PLoS genetics* 15.3 (2019), e1008036 (cf. p. 181).

[231] Paz POLAK et Eytan DOMANY. « Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes ». In : *BMC genomics* 7 (2006), p. 1-15 (cf. p. 148, 169).

[232] Sue POVEY, Ruth LOVERING, Elspeth BRUFORD et al. « The HUGO gene nomenclature committee (HGNC) ». In : *Human genetics* 109 (2001), p. 678-680 (cf. p. 119).

[233] Alkes L PRICE, Eleazar ESKIN et Pavel A PEVZNER. « Whole-genome analysis of Alu repeat elements reveals complex evolutionary history ». In : *Genome research* 14.11 (2004), p. 2245-2252 (cf. p. 181).

[234] Kim D PRUITT, Tatiana TATUSOVA, Garth R BROWN et al. « NCBI Reference Sequences (RefSeq) : current status, new features and genome annotation policy ». In : *Nucleic acids research* 40.D1 (2012), p. D130-D135 (cf. p. 83).

[235] Aaron R QUINLAN et Ira M HALL. « BEDTools : a flexible suite of utilities for comparing genomic features ». In : *Bioinformatics* 26.6 (2010), p. 841-842 (cf. p. 161).

[236] Sanjida H RANGWALA, Anatoliy KUZNETSOV, Victor ANANIEV et al. « Accessing NCBI data using the NCBI sequence viewer and genome data viewer (GDV) ». In : *Genome research* 31.1 (2021), p. 159-169 (cf. p. 78).

[237] Suhas SP RAO, Miriam H HUNTLEY, Neva C DURAND et al. « A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping ». In : *Cell* 159.7 (2014), p. 1665-1680 (cf. p. 70, 219).

[238] Aviv REGEV, Sarah A TEICHMANN, Eric S LANDER et al. « The human cell atlas ». In : *elife* 6 (2017), e27041 (cf. p. 221).

[239] Knut REINERT, Ben LANGMEAD, David WEESE et al. « Alignment of next-generation sequencing reads ». In : *Annual review of genomics and human genetics* 16 (2015), p. 133-151 (cf. p. 71, 72).

[240] Zachary RENFRO, Bryan E WHITE et Kimberly E STEPHENS. « CCAAT enhancer binding protein gamma (C/EBP-γ) : An understudied transcription factor ». In : *Advances in Biological Regulation* (2022), p. 100861 (cf. p. 175).

[241] Ho Sung RHEE et B Franklin PUGH. « ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy ». In : *Current protocols in molecular biology* 100.1 (2012), p. 21-24 (cf. p. 63, 104).

[242] David B ROBERTS. « Drosophila melanogaster : the model organism ». In : *Entomologia experimentalis et applicata* 121.2 (2006), p. 93-103 (cf. p. 124).

[243] Gordon ROBERTSON, Martin HIRST, Matthew BAINBRIDGE et al. « Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing ». In : *Nature methods* 4.8 (2007), p. 651-657 (cf. p. 72, 103).

[244] Matthew J ROSSI, William KM LAI et B Franklin PUGH. « Simplified ChIP-exo assays ». In : *Nature communications* 9.1 (2018), p. 2842 (cf. p. 63).

[245] Assaf ROTEM, Oren RAM, Noam SHORESH et al. « Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state ». In : *Nature biotechnology* 33.11 (2015), p. 1165-1172 (cf. p. 221).

[246] Joel ROZOWSKY, Ghia EUSKIRCHEN, Raymond K AUERBACH et al. « PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls ». In : *Nature biotechnology* 27.1 (2009), p. 66 (cf. p. 62, 72).

[247] Marilyn SAFRAN, Irina DALAH, Justin ALEXANDER et al. « GeneCards Version 3 : the human gene integrator ». In : *Database* 2010 (2010) (cf. p. 83).

[248] Manuel SÁNCHEZ-CASTILLO, David RUAU, Adam C WILKINSON et al. « CODEX : a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities ». In : *Nucleic acids research* 43.D1 (2015), p. D1117-D1123 (cf. p. 156).

[249] Fred SANGER et Alan R COULSON. « A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase ». In : *Journal of molecular biology* 94.3 (1975), p. 441-448 (cf. p. 57).

[250] David SANTIAGO-ALGARRA, Charbel SOUAID, Himanshu SINGH et al. « Epromoters function as a hub to recruit key transcription factors required for the inflammatory response ». In : *Nature communications* 12.1 (2021), p. 6660 (cf. p. 36, 46).

[251] Theresa SCHACHT, Marcus OSWALD, Roland EILS et al. « Estimating the activity of transcription factors by the effect on their target genes ». In : *Bioinformatics* 30.17 (2014), p. i401-i407 (cf. p. 223).

[252] Sabine SCHIRM, Josef JIRICNY et Walter SCHAFFNER. « The SV40 enhancer can be dissected into multiple segments, each with a different cell type specificity. » In : *Genes & development* 1.1 (1987), p. 65-74 (cf. p. 39).

[253] Dominic SCHMIDT, Petra C SCHWALIE, Michael D WILSON et al. « Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages ». In : *Cell* 148.1-2 (2012), p. 335-348 (cf. p. 92, 183).

[254] Dominic SCHMIDT, Michael D WILSON, Benoit BALLESTER et al. « Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding ». In : *Science* 328.5981 (2010), p. 1036-1040 (cf. p. 52).

[255] Anthony D SCHMITT, Ming HU, Inkyung JUNG et al. « A compendium of chromatin contact maps reveals spatially active regions in the human genome ». In : *Cell reports* 17.8 (2016), p. 2042-2059 (cf. p. 55).

[256] Patrick S SCHNABLE, Doreen WARE, Robert S FULTON et al. « The B73 maize genome : complexity, diversity, and dynamics ». In : *science* 326.5956 (2009), p. 1112-1115 (cf. p. 86).

[257] Valerie A SCHNEIDER, Tina GRAVES-LINDSAY, Kerstin HOWE et al. « Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly ». In : *Genome research* 27.5 (2017), p. 849-864 (cf. p. 81).

[258] Ida SCHOMBURG, Antje CHANG, Oliver HOFMANN et al. « BRENDA : a resource for enzyme data and metabolic information ». In : *Trends in biochemical sciences* 27.1 (2002), p. 54-56 (cf. p. 119).

[259] Mikkel SCHUBERT, Stinus LINDGREEN et Ludovic ORLANDO. « AdapterRemoval v2 : rapid adapter trimming, identification, and read merging ». In : *BMC research notes* 9.1 (2016), p. 1-7 (cf. p. 71).

[260] Walter A SCOTT et Dianne J WIGMORE. « Sites in simian virus 40 chromatin which are preferentially cleaved by endonucleases ». In : *Cell* 15.4 (1978), p. 1511-1518 (cf. p. 65).

[261] Madhobi SEN, Xin WANG, Feda H HAMDAN et al. « ARID1A facilitates KRAS signaling-regulated enhancer activity in an AP1-dependent manner in colorectal cancer cells ». In : *Clinical epigenetics* 11.1 (2019), p. 1-16 (cf. p. 145).

[262] Aurelien A SERANDOUR, Gordon D BROWN, Joshua D COHEN et al. « Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties ». In : *Genome biology* 14 (2013), p. 1-9 (cf. p. 63).

[263] Nathan C SHEFFIELD et Christoph BOCK. « LOLA : enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor ». In : *Bioinformatics* 32.4 (2016), p. 587-589 (cf. p. 155).

[264] Nathan C SHEFFIELD, Robert E THURMAN, Lingyun SONG et al. « Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions ». In : *Genome research* 23.5 (2013), p. 777-788 (cf. p. 156).

[265] Fabian SIEVERS et Desmond G HIGGINS. « Clustal omega ». In : *Current protocols in bioinformatics* 48.1 (2014), p. 3-13 (cf. p. 161).

[266] Corinne N SIMONTI, Mihaela PAVLIČEV et John A CAPRA. « Transposable element exaptation into regulatory regions is rare, influenced by evolutionary age, and subject to pleiotropic constraints ». In : *Molecular Biology and Evolution* 34.11 (2017), p. 2856-2869 (cf. p. 226).

[267] Surendra SINGH, Chad BROCKER, Vindhya KOPPAKA et al. « Aldehyde dehydrogenases in cellular responses to oxidative/electrophilicstress ». In : *Free radical biology and medicine* 56 (2013), p. 89-101 (cf. p. 100).

[268] Stephen T SMALE et James T KADONAGA. « The RNA polymerase II core promoter ». In : *Annual review of biochemistry* 72.1 (2003), p. 449-479 (cf. p. 34).

[269] Andrew D SMITH, Pavel SUMAZIN et Michael Q ZHANG. « Tissue-specific regulatory elements in mammalian promoters ». In : *Molecular systems biology* 3.1 (2007), p. 73 (cf. p. 142).

[270] Catherine C SMITH. « The growing landscape of FLT3 inhibition in AML ». In : *Hematology 2014, the American Society of Hematology Education Program Book* 2019.1 (2019), p. 539-547 (cf. p. 97).

[271] Robert M SMITH. « BLOOD The Journal of The American Society of Hematology ». In : (2011) (cf. p. 45).

[272] Mark J SOLOMON, Pamela L LARSEN et Alexander VARSHAVSKY. « Mapping proteinDNA interactions in vivo with formaldehyde : Evidence that histone H4 is retained on a highly transcribed gene ». In : *Cell* 53.6 (1988), p. 937-947 (cf. p. 103).

[273] Lingyun SONG et Gregory E CRAWFORD. « DNase-seq : a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells ». In : *Cold Spring Harbor Protocols* 2010.2 (2010), pdb-prot5384 (cf. p. 66, 67).

[274] Natalie de SOUZA. « The ENCODE project ». In : *Nature methods* 9.11 (2012), p. 1046-1046 (cf. p. 82, 155).

[275] Demetrios A SPANDIDOS et Neil M WILKIE. « Host-specificities of papillomavirus, Moloney murine sarcoma virus and simian virus 40 enhancer sequences. » In : *The EMBO journal* 2.7 (1983), p. 1193-1199 (cf. p. 39).

[276] François SPITZ et Eileen EM FURLONG. « Transcription factors : from enhancer binding to developmental control ». In : *Nature reviews genetics* 13.9 (2012), p. 613-626 (cf. p. 40, 52).

[277] Klara STEFFLOVA, David THYBERT, Michael D WILSON et al. « Cooperativity and rapid evolution of cobound transcription factors in closely related mammals ». In : *Cell* 154.3 (2013), p. 530-540 (cf. p. 52).

[278] Zachary D STEPHENS, Skylar Y LEE, Faraz FAGHRI et al. « Big data : astronomical or genomical? » In : *PLoS biology* 13.7 (2015), e1002195 (cf. p. 81).

[279] Jennitte L STEVENS, Greg T CANTIN, Gang WANG et al. « Transcription control by E1A and MAP kinase pathway via Sur2 mediator subunit ». In : *Science* 296.5568 (2002), p. 755-758 (cf. p. 44).

[280] Gary D STORMO. « Modeling the specificity of protein-DNA interactions ». In : *Quantitative biology* 1.2 (2013), p. 115-130 (cf. p. 75, 77).

[281] Gary D STORMO, Thomas D SCHNEIDER, Larry GOLD et al. « Use of the 'Perceptron'algorithm to distinguish translational initiation sites in E. coli ». In : *Nucleic acids research* 10.9 (1982), p. 2997-3011 (cf. p. 75).

[282] Hao SUN, Jiejun WU, Priyankara WICKRAMASINGHE et al. « Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq ». In : *Nucleic acids research* 39.1 (2011), p. 190-201 (cf. p. 219).

[283] Vasavi SUNDARAM, Yong CHENG, Zhihai MA et al. « Widespread contribution of transposable elements to the innovation of gene regulatory networks ». In : *Genome research* 24.12 (2014), p. 1963-1976 (cf. p. 92, 149, 169-171).

[284] Vasavi SUNDARAM et Joanna WYSOCKA. « Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes ». In : *Philosophical Transactions of the Royal Society B* 375.1795 (2020), p. 20190347 (cf. p. 149, 177, 183, 226).

[285] Kazutoshi TAKAHASHI et Shinya YAMANAKA. « Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors ». In : *cell* 126.4 (2006), p. 663-676 (cf. p. 48, 50).

[286] Diethard TAUTZ. « Notes on the definition and nomenclature of tandemly repetitive DNA sequences ». In : *DNA fingerprinting : State of the science* (1993), p. 21-28 (cf. p. 228).

[287] Paul D THOMAS, Dustin EBERT, Anushya MURUGANUJAN et al. « PANTHER : Making genome-scale phylogenetics accessible to all ». In : *Protein Science* 31.1 (2022), p. 8-22 (cf. p. 83).

[288] Morgane THOMAS-CHOLLIER, Olivier SAND, Jean-Valéry TURATSINZE et al. « RSAT : regulatory sequence analysis tools ». In : *Nucleic acids research* 36.suppl_-2 (2008), W119-W127 (cf. p. 77).

[289] Dawn THOMPSON, Aviv REGEV et Sushmita ROY. « Comparative analysis of gene regulatory networks : from network reconstruction to evolution ». In : *Annual review of cell and developmental biology* 31 (2015), p. 399-428 (cf. p. 222).

[290] Aurore TOUZART, Etienne LENGLINÉ, Mehdi LATIRI et al. « Epigenetic Silencing Affects l-Asparaginase Sensitivity and Predicts Outcome in T-ALLASNS Promoter Methylation Predicts Outcome in T-ALL ». In : *Clinical cancer research* 25.8 (2019), p. 2483-2493 (cf. p. 46).

[291] Marco TRIZZINO, YoSon PARK, Marcia HOLSBACH-BELTRAME et al. « Transposable elements are the primary source of novelty in primate gene regulation ». In : *Genome research* 27.10 (2017), p. 1623-1633 (cf. p. 226).

[292] Alice P TSANG, Jane E VISVADER, C Alexander TURNER et al. « FOG, a multitype zinc finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and megakaryocytic differentiation ». In : *Cell* 90.1 (1997), p. 109-119 (cf. p. 53).

[293] Susan TWEEDIE, Bryony BRASCHI, Kristian GRAY et al. « Genenames. org : the HGNC and VGNC resources in 2021 ». In : *Nucleic acids research* 49.D1 (2021), p. D939-D946 (cf. p. 83).

[294] Mathias UHLÉN, Linn FAGERBERG, Björn M HALLSTRÖM et al. « Tissue-based map of the human proteome ». In : *Science* 347.6220 (2015), p. 1260419 (cf. p. 48).

[295] Anna ULLASTRES, Miriam MERENCIANO et Josefa GONZÁLEZ. « Regulatory regions in natural transposable element insertions drive interindividual differences in response to immune challenges in Drosophila ». In : *Genome Biology* 22.1 (2021), p. 1-30 (cf. p. 224).

[296] Anton VALOUEV, David S JOHNSON, Andreas SUNDQUIST et al. « Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data ». In : *Nature methods* 5.9 (2008), p. 829-834 (cf. p. 52, 72).

[297] Flora E VAN LEEUWEN. « 4 Risk of acute myelogenous leukaemia and myelodysplasia following cancer treatment ». In : *Baillière's clinical haematology* 9.1 (1996), p. 57-85 (cf. p. 97).

[298] James W VARDIMAN, Nancy Lee HARRIS et Richard D BRUNNING. « The World Health Organization (WHO) classification of the myeloid neoplasms ». In : *Blood, The Journal of the American Society of Hematology* 100.7 (2002), p. 2292-2302 (cf. p. 97).

[299] Vasilis VASILIOU, Aglaia PAPPA et Tia ESTEY. « Role of human aldehyde dehydrogenases in endobiotic and xenobiotic metabolism ». In : *Drug metabolism reviews* 36.2 (2004), p. 279-299 (cf. p. 100).

[300] G VENTON, M PÉREZ-ALEA, C BAIER et al. « Aldehyde dehydrogenases inhibition eradicates leukemia stem cells while sparing normal progenitors ». In : *Blood cancer journal* 6.9 (2016), e469-e469 (cf. p. 101, 186).

[301] Diego VILLAR, Camille BERTHELOT, Sarah ALDRIDGE et al. « Enhancer evolution across 20 mammalian species ». In : *Cell* 160.3 (2015), p. 554-566 (cf. p. 52, 226).

[302] Diego VILLAR, Paul FLICEK et Duncan T ODOM. « Evolution of transcription factor binding in metazoans—mechanisms and functional implications ». In : *Nature Reviews Genetics* 15.4 (2014), p. 221-233 (cf. p. 52).

[303] Jianrong WANG, Nathan J BOWEN, Leonardo MARIÑO-RAMIÉREZ et al. « A c-Myc regulatory subnetwork from human transposable element sequences ». In : *Molecular BioSystems* 5.12 (2009), p. 1831-1839 (cf. p. 92, 171).

[304] Ting WANG, Jue ZENG, Craig B LOWE et al. « Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53 ». In : *Proceedings of the National Academy of Sciences* 104.47 (2007), p. 18613-18618 (cf. p. 92, 171, 183, 227).

[305] Wachiraporn WANICHNOPPARAT, Kulachanya SUWANWONGSE, Piyapat PIN-ON et al. « Genes associated with the cis-regulatory functions of intragenic LINE-1 elements ». In : *BMC genomics* 14.1 (2013), p. 1-9 (cf. p. 177).

[306] James D WATSON et Francis HC CRICK. « Molecular structure of nucleic acids : a structure for deoxyribose nucleic acid ». In : *Nature* 171.4356 (1953), p. 737-738 (cf. p. 26, 28).

[307] Amy S WEINMANN, Stephanie M BARTLEY, Theresa ZHANG et al. « Use of chromatin immunoprecipitation to clone novel E2F target promoters ». In : *Molecular and cellular biology* 21.20 (2001), p. 6820-6832 (cf. p. 103).

[308] Harold WEINTRAUB et Mark GROUDINE. « Chromosomal Subunits in Active Genes Have an Altered Conformation : Globin genes are digested by deoxyribonuclease I in red blood cell nuclei but not in fibroblast nuclei. » In : *Science* 193.4256 (1976), p. 848-856 (cf. p. 65).

[309] Jacob WHITE. « PubMed 2.0 ». In : *Medical reference services quarterly* 39.4 (2020), p. 382-387 (cf. p. 83).

[310] Warren A WHYTE, David A ORLANDO, Denes HNISZ et al. « Master transcription factors and mediator establish super-enhancers at key cell identity genes ». In : *Cell* 153.2 (2013), p. 307-319 (cf. p. 45).

[311] Thomas WICKER, François SABOT, Aurélie HUA-VAN et al. « A unified classification system for eukaryotic transposable elements ». In : *Nature reviews genetics* 8.12 (2007), p. 973-982 (cf. p. 86, 88, 90).

[312] Carl WU. « The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I ». In : *Nature* 286.5776 (1980), p. 854-860 (cf. p. 65).

[313] Carl WU, Paul M BINGHAM, Kenneth J LIVAK et al. « The chromatin structure of specific genes : I. Evidence for higher order domains of defined DNA sequence ». In : *Cell* 16.4 (1979), p. 797-806 (cf. p. 65).

[314] Ray WU et Ellen TAYLOR. « Nucleotide sequence analysis of DNA : II. Complete nucleotide sequence of the cohesive ends of bacteriophage λ DNA ». In : *Journal of molecular biology* 57.3 (1971), p. 491-511 (cf. p. 57).

[315] Shwu-Yuan WU et Cheng-Ming CHIANG. « The double bromodomain-containing chromatin adaptor Brd4 and transcriptional regulation ». In : *Journal of Biological Chemistry* 282.18 (2007), p. 13141-13145 (cf. p. 45).

[316] Dan XIE, Chieh-Chun CHEN, Leon M PTASZEK et al. « Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species ». In : *Genome research* 20.6 (2010), p. 804-815 (cf. p. 171).

[317] Haibin XU, Xiang LUO, Jun QIAN et al. « FastUniq : a fast de novo duplicates removal tool for paired short reads ». In : *PloS one* 7.12 (2012), e52249 (cf. p. 72).

[318] Xiangchou YANG, Rongxin YAO et Hong WANG. « Update of ALDH as a potential biomarker and therapeutic target for AML ». In : *BioMed Research International* 2018 (2018) (cf. p. 101).

[319] Jeffrey A YODER, Michael E NIELSEN, Chris T AMEMIYA et al. « Zebrafish as an immunological model system ». In : *Microbes and Infection* 4.14 (2002), p. 1469-1478 (cf. p. 222).

[320] Xueping YU, Jimmy LIN, Donald J ZACK et al. « Identification of tissue-specific cis-regulatory modules based on interactions between transcription factors ». In : *BMC bioinformatics* 8 (2007), p. 1-13 (cf. p. 221).

[321] Muhammad A ZABIDI, Cosmas D ARNOLD, Katharina SCHERNHUBER et al. « Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation ». In : *Nature* 518.7540 (2015), p. 556-559 (cf. p. 36).

[322] Yong ZHANG, Tao LIU, Clifford A MEYER et al. « Model-based analysis of ChIP-Seq (MACS) ». In : *Genome biology* 9.9 (2008), p. 1-9 (cf. p. 62, 72).

[323] Shanrong ZHAO. « Assessment of the impact of using a reference transcriptome in mapping short RNA-Seq reads ». In : *PLoS One* 9.7 (2014), e101374 (cf. p. 59).

[324] Rongbin ZHENG, Changxin WAN, Shenglin MEI et al. « Cistrome Data Browser : expanded datasets and new tools for gene regulatory analysis ». In : *Nucleic acids research* 47.D1 (2019), p. D729-D735 (cf. p. 117).

[325] Maggie ZHOU, Archna BHASIN et William S REZNIKOFF. « Molecular genetic analysis of transposase-end DNA sequence recognition : cooperativity of three adjacent base-pairs in specific interaction with a mutant Tn5 transposase ». In : *Journal of molecular biology* 276.5 (1998), p. 913-925 (cf. p. 67).

[326] Zhaonan ZOU, Tazro OHTA, Fumihito MIURA et al. « ChIP-Atlas 2021 update : a data-mining suite for exploring epigenomic landscapes by fully integrating chip-seq, ATAC-seq and Bisulfite-seq data ». In : *Nucleic acids research* 50.W1 (2022), W175-W182 (cf. p. 117, 222).

ANNEXES

A. Proportion en pb de TE sur le génome

Nous avons également représenté la proportion en pb de TE sur le génome (Figure 1). Sur cette figure nous pouvons voir que, comme dans la Figure H.Nishihara [212], le pourcentage de TE sur le génome est d'environ 50%.

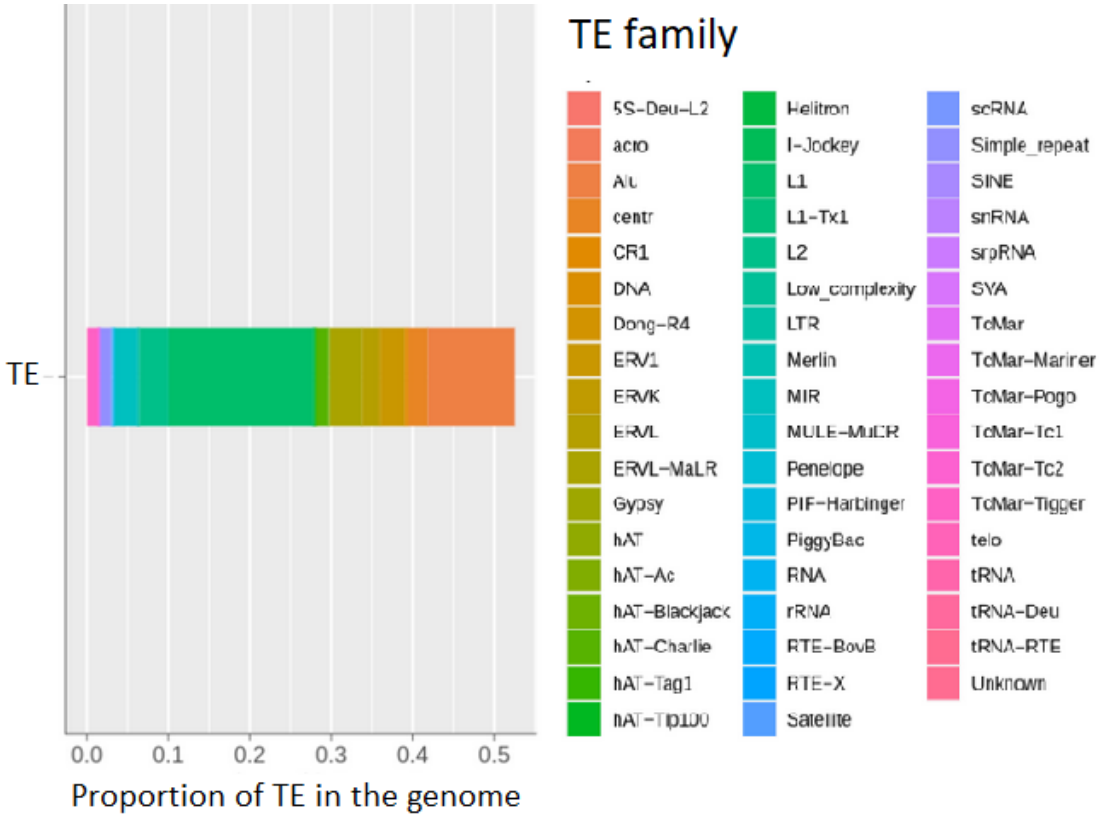


FIGURE 1. – ***Barplot représentant le pourcentage en pb de chaque famille de TE en fonction de la longueur totale du génome.*** *La proportion de TE sur le génome est d'environ 50% comme le montre la figure de H.Nishira [212]. Ce calcul a été fait sur à partir de la longeur en pb.*

B. Article JASPAR

Jaime A CASTRO-MONDRAGON, Rafael RIUDAVETS-PUIG, Ieva RAULUSEVICIUTE et al. « JASPAR 2022 : the 9th release of the open-access database of transcription factor binding profiles ». In : *Nucleic acids research* 50.D1 (2022), p. D165-D173

Published online 30 November 2021 *Nucleic Acids Research, 2022, Vol. 50, Database issue* ***D165–D173***
https://doi.org/10.1093/nar/gkab1113

JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles

Jaime A. Castro-Mondragon [1,†], Rafael Riudavets-Puig [1,†], Ieva Rauluseviciute [1,†], Roza Berhanu Lemma [1], Laura Turchi [2], Romain Blanc-Mathieu [2], Jeremy Lucas [2], Paul Boddie[1], Aziz Khan [3], Nicolás Manosalva Pérez [4,5], Oriol Fornes [6], Tiffany Y. Leung [6], Alejandro Aguirre [6], Fayrouz Hammal [7], Daniel Schmelter[8], Damir Baranasic[9,10], Benoit Ballester [7], Albin Sandelin [11,*], Boris Lenhard [9,10,*], Klaas Vandepoele [4,5,12], Wyeth W. Wasserman [6,*], François Parcy [2,*] and Anthony Mathelier [1,13,*]

[1]Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway, [2]Laboratoire Physiologie Cellulaire et Végétale, Univ. Grenoble Alpes, CNRS, CEA, INRAE, IRIG-DBSCI-LPCV, 17 avenue des martyrsF-38054, Grenoble, France, [3]Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA94305, USA, [4]Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 71, 9052 Ghent, Belgium, [5]VIB Center for Plant Systems Biology, Technologiepark 71, 9052 Ghent, Belgium, [6]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, [7]Aix Marseille Univ, INSERM, TAGC, Marseille, France, [8]UCSC Genome Browser, University of California Santa Cruz, Santa Cruz, CA95060, USA, [9]MRC London Institute of Medical Sciences, Du Cane Road, London, W12 0NN, UK, [10]Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, Hammersmith Hospital Campus, Du Cane Road, London W12 0NN, UK, [11]The Bioinformatics Centre, Department of Biology & Biotech Research and Innovation Centre, University of Copenhagen, Ole Maaloes Vej 5, DK2200 Copenhagen N, Denmark, [12]Bioinformatics Institute Ghent, Ghent University, Technologiepark 71, 9052 Ghent, Belgium and [13]Department of Medical Genetics, Institute of Clinical Medicine, University of Oslo and Oslo University Hospital, Oslo, Norway

Received September 15, 2021; Revised October 20, 2021; Editorial Decision October 20, 2021; Accepted October 22, 2021

ABSTRACT

JASPAR (http://jaspar.genereg.net/) is an open-access database containing manually curated, non-redundant transcription factor (TF) binding profiles for TFs across six taxonomic groups. In this 9th release, we expanded the CORE collection with 341 new profiles (148 for plants, 101 for vertebrates, 85 for urochordates, and 7 for insects), which corresponds to a 19% expansion over the previous release. We added 298 new profiles to the Unvalidated collection when no orthogonal evidence was found in the literature. All the profiles were clustered to provide familial binding profiles for each taxonomic group. Moreover, we revised the structural classification of DNA binding domains to consider plant-specific TFs. This release introduces word clouds to represent the scientific knowledge associated with each TF. We updated the genome tracks of TFBSs predicted with JASPAR profiles in eight organisms; the human and mouse TFBS predictions can be visualized as native tracks in the UCSC Genome Browser. Finally, we provide a new tool to perform JASPAR TFBS enrichment analysis in user-provided genomic regions. All the data is accessible through the JASPAR website, its associated RESTful API, the R/Bioconductor data package, and a new Python package, pyJASPAR, that facilitates serverless access to the data.

*To whom correspondence should be addressed. Email: anthony.mathelier@ncmm.uio.no
Correspondence may also be addressed to François Parcy. Email: francois.parcy@cea.fr
Correspondence may also be addressed to Wyeth W. Wasserman. Email: wyeth@cmmt.ubc.ca
Correspondence may also be addressed to Boris Lenhard. Email: b.lenhard@imperial.ac.uk
Correspondence may also be addressed to Albin Sandelin. Email: albin@binf.ku.dk
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

D166 Nucleic Acids Research, 2022, Vol. 50, Database issue

INTRODUCTION

Transcription factors are proteins that interact with the DNA in a sequence-specific manner through recognition of their TF binding sites (TFBSs) located at cis-regulatory regions (promoters, enhancers) to regulate transcription (1). TF binding to these regions occurs through direct interactions between the DNA-binding domains (DBDs) of TFs and the DNA. DBDs are classified into structural classes and families, and TFs with related DBDs typically have similar DNA binding preferences (2). The binding of TFs to cis-regulatory regions promotes or inhibits the assembly of the transcription machinery, thereby controlling gene expression regulation (1,3–5).

Sequence-specific TF-DNA interactions at TFBSs can be experimentally determined either *in vitro* or *in vivo*. High-throughput *in vitro* methods include systematic evolution of ligands by exponential enrichment (SELEX) (6) and protein binding-microarrays (PBM) (7) where TFs are exposed to synthesized DNA sequences. High-throughput *in vivo* assays include chromatin immunoprecipitation-based methods such as ChIP-seq (8), ChIP-exo (9) and ChIP-nexus (10), and cleavage-based methods such as cleavage under targets and tagmentation (11) or cleavage under targets and release using nuclease (12). These high-throughput assays (reviewed in (1)) provide unprecedented means to characterize the binding properties of individual TFs. Nevertheless, a challenge lies in our understanding of how TFs interact cooperatively at regulatory elements, for instance by forming dimers (13). Recently, CAP-SELEX revealed that TF pairs can bind in a DNA-dependent manner and that the combined binding of TFs can alter their individual binding specificities (14).

Despite the establishment of a wide variety of experimental techniques that delineate TF-DNA binding interactions and TF binding specificities, experimentally identifying all TFBSs for all TFs in various systems and biological conditions is intractable. To address this challenge, researchers rely on computational modeling to predict and investigate TF-DNA interactions. Such methods are helpful for investigating results of experimental methods with low resolution. For instance, ChIP-seq peaks are typically an order of magnitude larger than the actual binding sites of a targeted TF, and therefore computational methods can be used to pinpoint the binding sites within the peaks (15,16).

Given the importance of understanding TF-DNA interactions in studying gene expression regulation, various computational methods have been devised to model and predict TFBSs. The methods utilize experimentally identified TFBSs to build models and computationally predict TFBSs in a given genomic sequence (5). These computational methods range from basic representations such as sequence consensus-based models and position frequency matrices (PFMs) to more complex representations such as Markov and deep learning-based models (reviewed in (13,17–18)). PFMs, which summarize occurrences of each nucleotide at each position in a set of observed TF-DNA interactions, are largely and most commonly used to capture TF binding specificities. Unlike the simple consensus-based models, PFMs can be transformed to probabilistic or energy-based models to obtain position weight matrices (PWMs) (or position-specific scoring matrices (PSSMs)) that can be used to scan any DNA sequence and predict TFBSs with sum weights above a defined threshold (reviewed in (17)). Hence, TF binding preferences can be represented as PFMs, which can be interpreted as TF binding profiles or motifs. In this manuscript, we will use the term PFM, motif, and TF binding profile interchangeably.

JASPAR is a popular and regularly maintained open-access and manually curated database storing TF binding preferences as PFMs. The JASPAR CORE collection provides non-redundant binding preferences for TFs (one versioned profile per TF per taxon, except when a TF has multiple DNA-binding preferences) across 6 taxa: urochordata, vertebrates, plants, insects, nematodes, and fungi. Inclusion of new profiles requires orthogonal evidence for the binding preferences of the TFs, which is rigorously evaluated by our expert curators. To complement the CORE collection, we previously introduced the Unvalidated collection to store high-quality TF-binding profiles that are lacking orthogonal supporting evidence in the literature (19). Beyond the high-quality TF binding profiles and metadata stored in JASPAR, the popularity of the database originates from its simplicity, the tools embedded in its web-interface, and the multitude of popular resources and tools directly integrating JASPAR profiles. Some of these tools include: (i) the MEME suite, allowing various motif enrichment and discovery analysis (20), (ii) TFBSshape allowing investigation of DNA shape features for TFBSs to provide insight on the mechanism of protein–DNA interaction (21,22), (iii) CiiiDER (23) for TFBS prediction and analysis such as enrichment assessment in DNA sequences, (iv) RSAT, allowing motif discovery, TFBS motif analyses (24) and (v) i-*cis*Target, which allows the prediction of *cis*-regulatory modules and regulatory features (25,26).

In this paper, we present the 9th release of the JASPAR database, which provides a substantial update and expansion of TF binding profiles in the six taxonomic groups. The update includes not only binding profiles (as PFMs) but also revisited metadata. Additionally, we added word clouds to display enriched terms associated with TFs in the scientific literature. Furthermore, a rigorous structural classification of plant TF DBDs is provided to adequately consider the numerous plant-specific TFs. Finally, the update comes with a range of new or updated functionalities and resources such as a TFBS enrichment tool, the pyJASPAR package, new familial binding profiles, and native UCSC human and mouse genome tracks with TFBSs predicted from JASPAR TF binding profiles.

RESULTS

Expansion and update of the JASPAR database

TF binding profiles. In the 9th release of JASPAR, we discarded unused collections introduced in early releases of the database (27–29) that either did not correspond to TF-specific binding profiles or were data-type specific; we maintained the CORE and Unvalidated collections. We computed and compiled TF binding profiles obtained from CAP-SELEX (14), NCAP-SELEX (30), SELEX-seq (31), PBMs (32), ChIP-seq (33–36) and DAP-seq experiments from ReMap 2022 (36) and GEO (37), and ChIP-exo (38)

Nucleic Acids Research, 2022, Vol. 50, Database issue **D167**

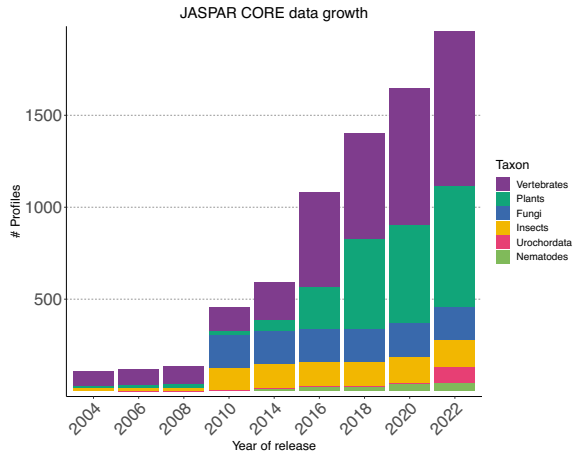


Figure 1. JASPAR CORE collection growth. The number of non-redundant profiles in each taxon (see legend) and overall through all JASPAR releases.

data (Supplementary Data 1 - Text for detailed list of datasets and method details). After manual curation of these profiles to confirm orthogonal supports in the literature, we augmented the CORE collection with 341 new binding profiles for TFs in four taxa (Table 1; Figure 1): 148 profiles in plants (a 24% expansion for this taxon), 101 profiles in vertebrates (a 13% expansion), 85 profiles in urochordates (only one motif was present since the second release of JASPAR in 2006 (27)), and seven profiles in insects (a 5% expansion). Out of these added profiles, 52 were upgraded from the Unvalidated to the CORE collection (27 and 25 for plants and vertebrates, respectively). Moreover, out of the newly introduced PFMs, 31 are associated with TF dimers. The literature that provides orthogonal evidence for the newly introduced TF binding profiles is provided in the metadata. Additionally, we updated 160 TF binding profiles across the six taxa with new PFMs (Table 1).

High-quality PFMs lacking orthogonal support were included in the Unvalidated collection (298 new profiles; Supplementary Data 1—Supplementary Figure S1, Supplementary Data 2—Supplementary Table S1). Specifically, 115 TF binding profiles are associated with zinc-finger TFs and 95 associated with TFs binding DNA as dimers. We provide the Unvalidated collection of TF binding profiles to the community to use with due caution since they are not yet supported with orthogonal evidence. We extend our invitation to the user community to be involved in the motif curation process by providing either new unvalidated profiles to consider or support to existing profiles in the collection.

We exhaustively revised the metadata to update information about the TF names, the structural class and family of the TF DBDs (following TFClass (39)), and links to external databases such as UniProt (40), ReMap (36), UniBind (15,16) and DNA Readout Viewer (41), whenever possible. Finally, we removed 32 profiles from the CORE collection (22 plant, 6 vertebrate and 4 fungi profiles) as they corresponded to synonyms of already present TF profiles, had low information content, or were derived from consensus strings (Table 1). In addition, we removed 85 profiles from the Unvalidated collection (44 vertebrate, 40 plant and 1 fungi profiles) because: (i) the corresponding profile or a new profile for the same TF was added to the CORE collection; (ii) the profile was of insufficient quality or (iii) the profile was misannotated (Supplementary Data 2—Supplementary Table S1; detailed list of all removed profiles at https://jaspar.genereg.net/changelog/).

The JASPAR 2022 CORE collection now stores 1955 non-redundant PFMs (841 for vertebrates, 656 for plants, 179 for fungi, 150 for insects, 43 for nematodes, and 86 for urochordates) (Table 1; Figure 1). Additionally, we maintained the associated collection of transcription factor flexible models (TFFMs; hidden Markov-based models capturing dinucleotide dependencies in TF–DNA interactions (42)) that were initialized using JASPAR CORE PFMs and trained on ChIP-seq data (Supplementary Data 1—Text). This process resulted in 303 new TFFMs (207 for vertebrates and 96 for plants).

Improved structural classification of plant TF DNA-binding domains. In JASPAR, TFs are classified based on TFClass (39), which provides a hierarchical structural classification (including superclass, class, and family) originally designed for human TFs and later extended to mammals. Since plant genomes contain many classes of TFs absent from TFClass, we expanded the TF structural classification using TFClass guidelines (39) and published structural evidence (Supplementary Data 2—Supplementary Table S2). In some rare cases (e.g. GARP and NF-Y TFs), we slightly diverged from TFClass so that the TF common name expected by users is provided in the structural class or family name. We arbitrarily decided to classify plant specific RAV TFs that contain two types of DBD (B3 and AP2) in the B3 Class. WRKY TFs that have a Zinc finger and a DBD derived from a GCM fold have been classified under the GCM domain factors class and WRKY family, and not in the Zinc-coordinating DNA-binding domains superclass. This homogenised classification introduced 27 novel entries in the TF DBD structural classification (Supplementary Data 2—Supplementary Table S2) and led to numerous corrections in the class and family fields compared to previous JASPAR releases.

Word clouds of terms associated with TFs in the scientific literature

Biological information about TFs, or genes in general, is scattered across many different resources, with PubMed possibly being the most extensive one. In an attempt to provide rich annotations for the TFs in JASPAR, we mined the corpus of article abstracts available in the PubMed database (43). We compiled sets of abstracts associated with each TF and weighted each word present by its relative importance when compared to all abstracts associated with other TFs in the same taxon (Supplementary Data 1—Text for method details). For each TF, the 200 highest weighted words were used to create a word cloud summarizing the annotations associated with that TF. As an example, Figure 2 illustrates the word cloud of terms associated with the PAX6

D168 *Nucleic Acids Research, 2022, Vol. 50, Database issue*

Table 1. Growth overview of the CORE collection of JASPAR 2022 compared to the previous release

Taxonomic Group	Non-redundant PFMs in JASPAR 2020	New non-redundant PFMs in JASPAR 2022	Removed profiles	Upgraded profiles (from Unvalidated to CORE)	Updated PFMs in JASPAR 2022	Total PFMs (non-redundant) in JASPAR 2022
Plants	530	121	22	27	44	656
Vertebrates	746	76	6	25	102	841
Urochordata	1	85	-	-	-	86
Insects	143	7	-	-	-	150
Nematodes	43	-	-	-	-	43
Fungi	183	-	4	-	14	179
CORE total	**1646**	**289**	**32**	**52**	**160**	**1955**

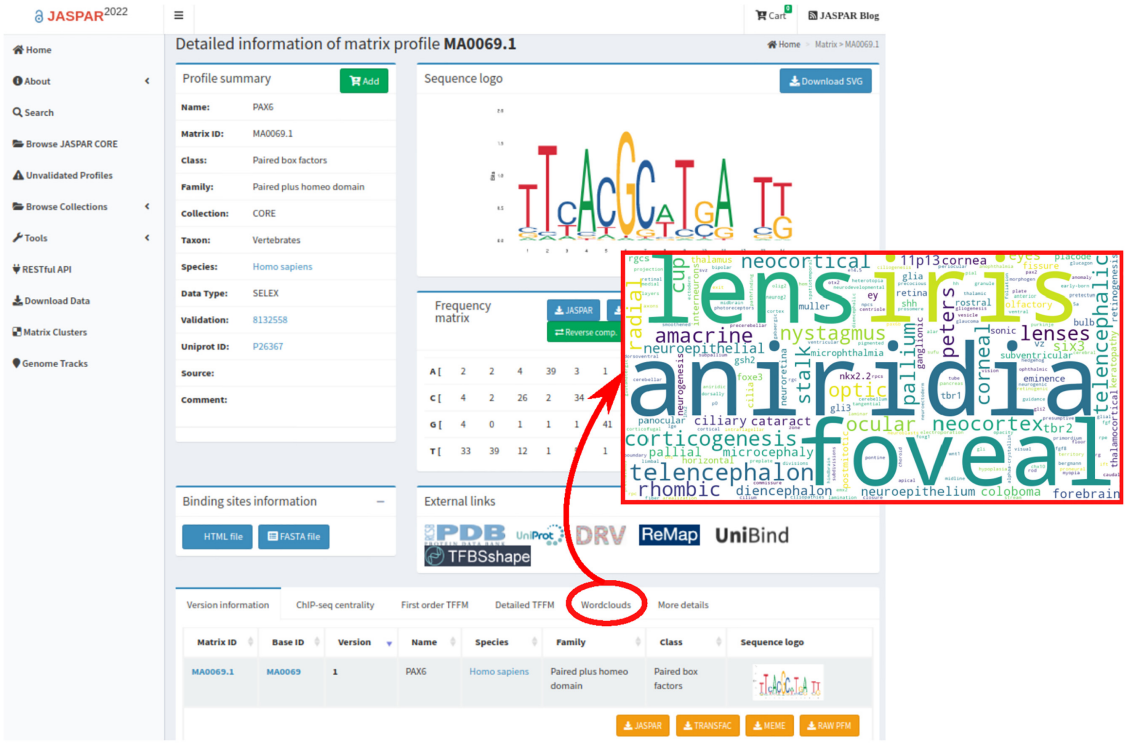


Figure 2. JASPAR TF word clouds. Webpage providing information about the binding profile associated with PAX6. The word cloud of terms obtained for PAX6 is highlighted in red, which supports the role of this TF in eye development and its implication in causing the genetic disorder aniridia.

TF in the scientific literature. Among the most significant terms, we find 'lens', 'iris', and 'foveal' that are representative of the importance of PAX6 in the development of the eye, while the term 'aniridia' reflects the link between some PAX6 mutations and the genetic disorder aniridia (44,45).

TF binding profile clusters, familial binding profiles, and genomic tracks

We updated the hierarchical clustering of the JASPAR TF binding profiles for each taxon with the RSAT matrix-clustering tool (46). Users can explore the CORE and Unvalidated collections through radial trees, which highlight the TF DBD structural classes, and directly access the underlying profiles by clicking on the TF name (https://jaspar.genereg.net/matrix-clusters).

The hierarchical clustering of JASPAR PFMs was used to generate a collection of familial binding profiles (5,47), following previously published methodologies (16,48). Such familial motifs are useful in applications where motif redundancy (many TFs have similar binding preferences) is not desired. In brief, we defined clusters based on the DBD structural classes along the hierarchical clustering of PFMs. Next, we computed a familial binding profile for each cluster, summarizing the profiles within the clusters following (48) (Supplementary Data 1—Text for method details; Supplementary Data 1—Supplementary Figure S2). The familial binding profiles, also referred to as archetypes in (48), can be explored and downloaded at https://jaspar.

Nucleic Acids Research, 2022, Vol. 50, Database issue **D169**

genereg.net/matrix-clusters and https://jaspar.genereg.net/downloads/, respectively.

One of the primary uses of PFMs is to predict binding sites. To facilitate this, we created ready-made prediction tracks for genome visualization and interpretation. Specifically, we scanned the genomes of eight organisms (*Arabidopsis thaliana, Caenorhabditis elegans, Ciona intestinalis, Danio rerio, Drosophila melanogaster, Homo sapiens, Mus musculus*, and *Saccharomyces cerevisiae*) with the JASPAR CORE PFMs associated with the same taxon to predict TFBSs and update the JASPAR TFBS genomic tracks. Moreover, we created a collection of familial TFBSs by merging overlapping TFBSs that were predicted from PFMs associated with the same familial binding profile (Supplementary Data 1—Text for method details). The TFBS predictions associated with all PFMs are available at http://expdata.cmmt.ubc.ca/JASPAR/downloads/UCSC_tracks/2022/. The familial binding TFBSs are available at https://jaspar.genereg.net/downloads/. Finally, we provide JASPAR TFBS predictions as genomic tracks, which can be visualized in genome browsers. Notably, the UCSC Genome Browser (49) now presents predicted human (for the hg19 and hg38 genome assemblies) and mouse (for the mm10 and mm39 genome assemblies) JASPAR TFBS data as a native tracks for the human and mouse genomes with information such as TF names, TFBS prediction scores, and PFM logos (Supplementary Data 1 - Supplementary Figure S3).

A command-line tool to evaluate JASPAR TFBS enrichment in genomic regions

A common challenge in the field of transcriptional regulation is to predict the TF(s) most likely to control a set of cis-regulatory regions. This challenge is classically addressed by evaluating the enrichment for potential TFBSs associated with candidate TFs in the genomic regions of interest compared to background regions (16,26,50–53). We previously introduced an enrichment tool that evaluates the enrichment for sets of direct TF–DNA interactions from UniBind in user-provided DNA regions compared to background regions (16). Following the same strategy, we introduce a TFBS enrichment tool to predict TFs with an enrichment of JASPAR TFBSs using the Locus Overlap Analysis (LOLA) tool (54). The enrichment tool is available as a command-line tool (https://jaspar.genereg.net/enrichment/, https://bitbucket.org/CBGR/jaspar_enrichment/).

As a use case, we studied the differential enrichment of predicted TFBSs at DNase-seq peaks observed in A549 cells before and after 2 h treatment with 100 nM dexamethasone. DNase-seq is an assay capturing open chromatin regions (55). Dexamethasone is a known agonist of the glucocorticoid receptor (NR3C1), a nuclear receptor that binds the DNA upon ligand-based activation. Figure 3 provides a visual representation of the differential TFBS enrichment analysis results when considering DNase-seq peaks in treated versus untreated cells. As expected, NR3C1 (a member of the Steroid hormone receptors (NR3) family) was the top enriched TF ($-\log_{10}(P) = 58.77$). Among other TFs showing a high enrichment of TFBSs, we observed many members of the Three-zinc finger Kruppel-related family (e.g. KLF factors, SP3, and SP9) (Supplementary Data 2—Supplementary Table S3). In another example, we observed the enrichment of TFBSs for the TFs FOXA1 and GATA3 in regions surrounding CpGs that are hypomethylated in estrogen receptor positive (ER+) breast cancers (56) (Supplementary Data 1—Supplementary Figure S4, Supplementary Data 2—Supplementary Supplementary Table S4). These TFs are well established drivers of ER+ breast cancers binding to hypomethylated enhancers in ER+ breast cancers (56).

pyJASPAR—serverless pythonic interface to JASPAR data

All data is accessible through the JASPAR website (https://jaspar.genereg.net/), its associated RESTful API (https://jaspar.genereg.net/api/) (57), and the JASPAR2022 R/Bioconductor data package (source code at https://github.com/da-bar/JASPAR2022). The JASPAR database can also be accessed using Biopython (58) but it requires a local MySQL server to query the underlying database, which limits its access and use. To make access to JASPAR data easier, we introduce a new Python package, pyJASPAR (59), which allows users to query and access all JASPAR data without setting up the underlying MySQL database.

pyJASPAR is implemented in Python 3 using the Biopython *motifs* module and SQLite3 to provide a serverless Pythonic interface to the JASPAR database. The package allows users to query and access TF binding profiles across various releases of JASPAR. The releases currently available are: JASPAR2014, JASPAR2016, JASPAR2018, JASPAR2020, and JASPAR2022. The pyJASPAR package will be updated when future JASPAR releases become available. TF binding profiles can be retrieved using JASPAR matrix IDs, TF names, or other metadata information (Supplementary Data 1—Text for more details).

pyJASPAR is open source and the code is available at https://github.com/asntech/pyjaspar/ under the GPL-3.0 License. The module can easily be installed with Conda from the bioconda channel (https://anaconda.org/bioconda/pyjaspar) (60) or from the Python Package Index with the *pip* command. Detailed documentation with usage examples is available at https://pyjaspar.rtfd.io/.

CONCLUSIONS AND PERSPECTIVES

For the 9th release of the JASPAR database, we substantially expanded the JASPAR CORE collection by 19% (341 added motifs). The newly introduced TF binding profiles were obtained after manual curation of PFMs predicted de novo from >3500 ChIP-seq/-exo datasets (from ReMap 2022 (36) and GEO (61)) or retrieved from publically available repositories. While we continued our commitment to provide non-redundant, high-quality TF binding profiles for TFs across six taxa, this release comes with an important increase in the number of profiles for urochordata, with 86 PFMs available when JASPAR has contained a single one since 2006 (27). We now also provide TFBS predictions in *Ciona intestinalis* using the 86 JASPAR binding profiles. This increase exemplifies how the investigation of transcriptional regulation is expanding across more model organisms.

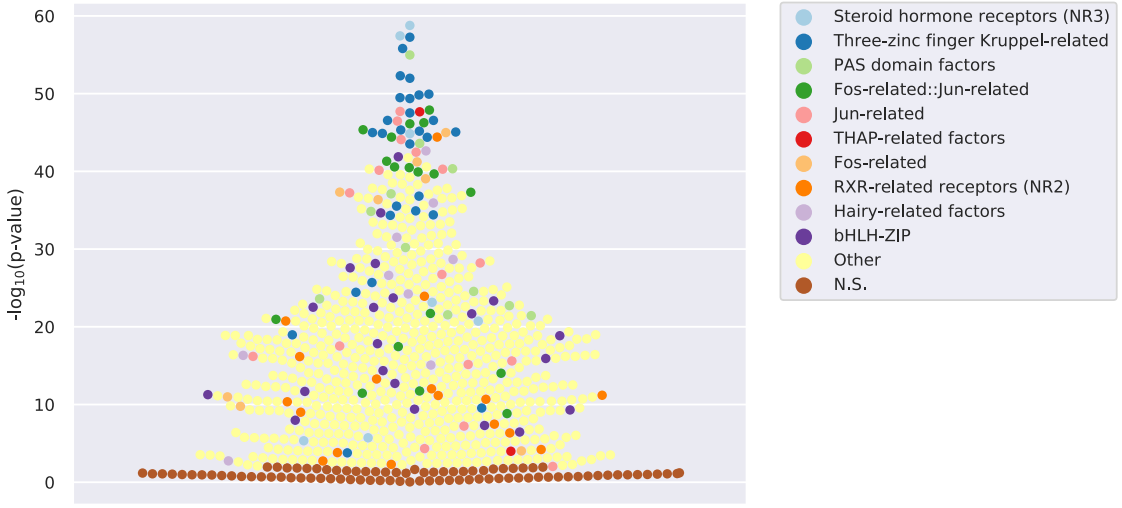


Figure 3. TFBS differential enrichment analysis on DNase-seq data for A549 cells before and after 2 h of dexamethasone treatment. Enrichment significance for each JASPAR profile from the vertebrate CORE collection is shown in the y-axis as -$\log_{10}(P)$ in this beeswarm plot. Each point depicts the Fisher exact test *P*-value (*P*) corresponding to a TF. The Points are colored based on the TF DBD structural family annotation, with a distinct color for each of the top 10 enriched families (see legend). Light yellow represents TF families outside of the top 10 enriched and with -$\log_{10}(P) > 3$ (Other) and brown represents TF families for which -$\log_{10}(P) \leq 3$ (non significant, N.S.).

An important question is what fraction of TFs have a binding profile in JASPAR. For humans, the JASPAR vertebrates CORE collection contains a binding profile for 43% of the 1639 human TFs (1), 56% when including the Unvalidated collection. If we consider the 1717 reported TFs for *A. thaliana* (62), 21% of these TFs have a profile in the JASPAR plants CORE collection, 22% when including the Unvalidated collection.

From the previous version of the Unvalidated collection (19), we found literature support for 81 profiles. Unfortunately, our team of curators did not succeed in identifying orthogonal validation in the literature for several high-quality motifs found enriched at ChIP-seq/-exo peak summits. As a result, 298 of such profiles were added to the previously introduced Unvalidated collection (19). The lack of experimental support for these profiles indicates an opportunity for the research field to explore these understudied TFs (63). Notably, 61% of the profiles in the vertebrates Unvalidated collection is associated with C2H2 zinc finger factors. A potential contributing challenge to obtaining orthogonal evidence may be the fact that many zinc-fingers, which represent the largest class of TFs, have been reported to regulate a limited number or even a single gene (e.g. Zfp568 (64), ZNF558 (65), ZNF410 (66) and ZFP64 (67)).

This JASPAR update comes with a new tool to compute TFBS enrichment given user-provided input and background sequences, mimicking a similar tool available with the UniBind database (16). The tool relies on the genome-wide TFBSs predicted using PFMs from the JASPAR CORE collection. Even though JASPAR predicted TFBSs will contain a high number of false positives, the enrichment tool could be useful to suggest roles for TFs for which no direct TF-DNA interactions are available in UniBind (16).

Consistent with Weidemüller *et al.* (63), we noticed that limited scientific literature (i.e. at most a single manuscript in PubMed) exists for many TFs, which clearly impacts the utility of the JASPAR word clouds. This constraint varies between taxa. For example, while the average number of PubMed manuscripts per vertebrate TF was ∼500, urochordata TFs were associated with an average of only four manuscripts. Furthermore, a large number of TFs associated with individual PubMed manuscripts was observed. The average number of vertebrate TFs associated with PubMed IDs was ∼19 with some associated with hundreds of TFs. An example is PubMed ID 21873635 that describes methods development of the Gene Ontology database (822 TFs), PubMed ID 12477932 that describes the Mammalian Gene Collection (MGC) Program (805 TFs), and PubMed ID 15618518 that analyzes the expression of TFs in the mouse brain (722 TFs). These manuscripts include general information about TFs. Therefore, we see opportunities to further improve the literature annotation engine, by decreasing the influence of outlier manuscripts and incorporating emerging natural language processing methods.

PFMs are still the most widely used models to represent TF binding preferences to DNA, despite their well-established caveats such as fixed-length and the failure to account for nucleotide interdependencies. A novel generation of computational models based on machine learning approaches such as deep learning are arising (68,69). Nevertheless, how to best share these models in a unified manner is still unclear despite some recent efforts (70) and will require discussion in the community. As the field moves to-

Nucleic Acids Research, 2022, Vol. 50, Database issue **D171**

wards a unified framework to share such models, we expect their inclusion in future JASPAR releases.

AUTHORS' NOTE

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint first authors. The order of co-first authors provided here was decided through a mushroom picking competition around the Sognsvann lake, Oslo, Norway. Co-first authors can prioritise their names when adding this paper's reference to their résumés.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

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 Shaun Mahony and Franklin Pugh for providing early access to ChIP-exo data from (38) and Emma Farley for pointers to urochordata data sets. We thank Vipin Kumar for contribution in the early curation sessions. We thank Walter Santana for his technical assistance to generate the motif radial trees, the UCSC Genome Browser project team for their assistance with the genome tracks, Harold Gutch and the NCMM IT team for their IT support, 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 Mathelier group); Norwegian Research Council [288404 to R.R.P., J.A.C.M., Mathelier group]; Norwegian Cancer Society [197884 to R.B.L., Mathelier group]; GRAL program [ANR-10-LABX-49-01] with the frame of the CBH-EUR-GS [ANR-17-EURE-0003 to Parcy group]; PhD fellowship from CNRS Prime80 (to L.T.); NHGRI [5U41HG002371-20 to D.S.]; BOF grant from Ghent University [BOF24Y2019001901 to N.M.P.]; PhD Fellowship from the Provence-Alpes-Côte d'Azur Regional Council (Région SUD); Institut National de la Santé et de la Recherche Médicale (INSERM) (to F.H.); Novo Nordisk Foundation [NNF20OC0059951, NNF19OC0058262]; Danish Cancer Foundation [R204-A12359]; Danish Independent Research Fund [6110-00207B, 7014-00120B]; Carlsberg Foundation; ERC the European Union's Horizon 2020 research and innovation programme (MSCA ITN pHioniC) (to Sandelin group and collaborators); Canadian Institutes of Health Research [PJT-162120]; Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant [RGPIN-2017-06824]; BC Children's Hospital Foundation and Research Institute (to Wasserman group). 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. 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. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., 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.
3. Reiter,F., Wienerroither,S. and Stark,A. (2017) Combinatorial function of transcription factors and cofactors. *Curr. Opin. Genet. Dev.*, **43**, 73–81.
4. Venters,B.J. and Pugh,B.F. (2009) How eukaryotic genes are transcribed. *Crit. Rev. Biochem. Mol. Biol.*, **44**, 117–141.
5. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
6. 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.
7. 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.
8. 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.
9. Franklin Pugh,B. (2012) Ultra-high resolution mapping of protein-genome interactions using ChIP-exo. *BMC Proc.*, **6**, O27.
10. 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.
11. Kaya-Okur,H.S., Wu,S.J., Codomo,C.A., Pledger,E.S., Bryson,T.D., Henikoff,J.G., Ahmad,K. and Henikoff,S. (2019) CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.*, **10**, 1930.
12. Skene,P.J. and Henikoff,S. (2017) An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife*, **6**, e21856.
13. 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.
14. Jolma,A., Yin,Y., Nitta,K.R., Dave,K., Popov,A., Taipale,M., Enge,M., Kivioja,T., Morgunova,E. and Taipale,J. (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, **527**, 384–388.
15. 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.
16. Puig,R.R., Boddie,P., Khan,A., Castro-Mondragon,J.A. and Mathelier,A. (2021) UniBind: maps of high-confidence direct TF-DNA interactions across nine species. *BMC Genomics*, **22**, 482.
17. Stormo,G.D. (2013) Modeling the specificity of protein-DNA interactions. *Quant Biol*, **1**, 115–130.
18. Koo,P.K. and Ploenzke,M. (2020) Deep learning for inferring transcription factor binding sites. *Curr Opin Syst Biol.*, **19**, 16–23.
19. Fornes,O., Castro-Mondragon,J.A., Khan,A., van der Lee,R., Zhang,X., Richmond,P.A., Modi,B.P., Correard,S., Gheorghe,M., Baranašić,D. *et al.* (2020) JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **48**, D87–D92.
20. Bailey,T.L., Johnson,J., Grant,C.E. and Noble,W.S. (2015) The MEME Suite. *Nucleic Acids Res.*, **43**, W39–W49.
21. 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.
22. Chiu,T.-P., Xin,B., Markarian,N., Wang,Y. and Rohs,R. (2020) TFBSshape: an expanded motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res.*, **48**, D246–D255.

23. Gearing,L.J., Cumming,H.E., Chapman,R., Finkel,A.M., Woodhouse,I.B., Luu,K., Gould,J.A., Forster,S.C. and Hertzog,P.J. (2019) CiiiDER: a tool for predicting and analysing transcription factor binding sites. *PLoS One*, **14**, e0215495.
24. 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.
25. Herrmann,C., Van de Sande,B., Potier,D. and Aerts,S. (2012) i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. *Nucleic Acids Res.*, **40**, e114.
26. 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.
27. 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.
28. 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.
29. 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–D110.
30. Zhu,F., Farnung,L., Kaasinen,E., Sahu,B., Yin,Y., Wei,B., Dodonova,S.O., Nitta,K.R., Morgunova,E., Taipale,M. *et al.* (2018) The interaction landscape between transcription factors and the nucleosome. *Nature*, **562**, 76–81.
31. Brozovic,M., Dantec,C., Dardaillon,J., Dauga,D., Faure,E., Gineste,M., Louis,A., Naville,M., Nitta,K.R., Piette,J. *et al.* (2018) ANISEED 2017: extending the integrated ascidian database to the exploration and evolutionary comparison of genome-scale datasets. *Nucleic Acids Res.*, **46**, D718–D725.
32. 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.
33. Ricardi,M.M., González,R.M., Zhong,S., Domínguez,P.G., Duffy,T., Turjanski,P.G., Salgado Salter,J.D., Alleva,K., Carrari,F., Giovannoni,J.J. *et al.* (2014) Genome-wide data (ChIP-seq) enabled identification of cell wall-related and aquaporin genes as targets of tomato ASR1, a drought stress-responsive transcription factor. *BMC Plant Biol.*, **14**, 29.
34. Du,M., Zhao,J., Tzeng,D.T.W., Liu,Y., Deng,L., Yang,T., Zhai,Q., Wu,F., Huang,Z., Zhou,M. *et al.* (2017) MYC2 orchestrates a hierarchical transcriptional cascade that regulates jasmonate-mediated plant immunity in tomato. *Plant Cell*, **29**, 1883–1906.
35. Liu,Y., Shi,Y., Zhu,N., Zhong,S., Bouzayen,M. and Li,Z. (2020) SlGRAS4 mediates a novel regulatory pathway promoting chilling tolerance in tomato. *Plant Biotechnol. J.*, **18**, 1620–1633.
36. Hammal,F., de Langen,P., Bergon,A., Lopez,F. and Ballester,B. (2021) ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res.*, https://doi.org/10.1093/nar/gkab996.
37. 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.
38. Rossi,M.J., Kuntala,P.K., Lai,W.K.M., Yamada,N., Badjatia,N., Mittal,C., Kuzu,G., Bocklund,K., Farrell,N.P., Blanda,T.R. *et al.* (2021) A high-resolution protein architecture of the budding yeast genome. *Nature*, **592**, 309–314.
39. 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.
40. UniProt,Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
41. Adam,K., Gyorgypal,Z. and Hegedus,Z. (2020) DNA Readout Viewer (DRV): visualization of specificity determining patterns of protein-binding DNA segments. *Bioinformatics*, **36**, 2286–2287.
42. Mathelier,A. and Wasserman,W.W. (2013) The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, **9**, e1003214.
43. Sayers,E.W., Beck,J., Bolton,E.E., Bourexis,D., Brister,J.R., Canese,K., Comeau,D.C., Funk,K., Kim,S., Klimke,W. *et al.* (2021) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **49**, D10–D17.
44. Jordan,T., Hanson,I., Zaletayev,D., Hodgson,S., Prosser,J., Seawright,A., Hastie,N. and van Heyningen,V. (1992) The human PAX6 gene is mutated in two patients with aniridia. *Nat. Genet.*, **1**, 328–332.
45. Gehring,W.J. and Ikeo,K. (1999) Pax 6: mastering eye morphogenesis and eye evolution. *Trends Genet.*, **15**, 371–377.
46. 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.
47. Mahony,S., Auron,P.E. and Benos,P.V. (2007) DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput. Biol.*, **3**, e61.
48. Vierstra,J., Lazar,J., Sandstrom,R., Halow,J., Lee,K., Bates,D., Diegel,M., Dunn,D., Neri,F., Haugen,E. *et al.* (2020) Global reference mapping of human transcription factor footprints. *Nature*, **583**, 729–736.
49. Navarro Gonzalez,J., Zweig,A.S., Speir,M.L., Schmelter,D., Rosenbloom,K.R., Raney,B.J., Powell,C.C., Nassar,L.R., Maulding,N.D., Lee,C.M. *et al.* (2021) The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.*, **49**, D1046–D1057.
50. 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.
51. Puente-Santamaria,L., Wasserman,W.W. and Del Peso,L. (2019) TFEA.ChIP: a tool kit for transcription factor binding site enrichment analysis capitalizing on ChIP-seq datasets. *Bioinformatics*, **35**, 5339–5340.
52. Roopra,A. (2020) MAGIC: A tool for predicting transcription factors and cofactors driving gene sets using ENCODE data. *PLoS Comput. Biol.*, **16**, e1007800.
53. Arenillas,D.J., Forrest,A.R.R., Kawaji,H., Lassmann,T., Wasserman,W.W., Mathelier,A and Consortium FANTOM Consortium (2016) CAGEd-oPOSSUM: motif enrichment analysis from CAGE-derived TSSs. *Bioinformatics*, **32**, 2858–2860.
54. Sheffield,N.C. and Bock,C. (2015) LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics*, **32**, 587–589.
55. Song,L. and Crawford,G.E. (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.*, **2010**, db.prot5384.
56. Fleischer,T., Tekpli,X., Mathelier,A., Wang,S., Nebdal,D., Dhakal,H.P., Sahlberg,K.K., Schlichting,E., Børresen-Dale,A.-L., Oslo Breast Cancer Research Consortium (OSBREAC) *et al.* (2017) DNA methylation at enhancers identifies distinct breast cancer lineages. *Nat. Commun.*, **8**, 1379.
57. Khan,A. and Mathelier,A. (2017) JASPAR RESTful API: accessing JASPAR data from any programming language. *Bioinformatics*, **34**, 1612–1614.
58. Cock,P.J.A., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
59. Khan,A. (2021) *pyJASPAR: a Pythonic interface to JASPAR transcription factor motifs*. https://doi.org/10.5281/zenodo.5062370.
60. Grüning,B., Dale,R., Sjödin,A., Chapman,B.A., Rowe,J., Tomkins-Tinch,C.H., Valieris,R. and Köster,J. (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods*, **15**, 475–476.

61. Edgar,R. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
62. Jin,J., Tian,F., Yang,D.-C., Meng,Y.-Q., Kong,L., Luo,J. and Gao,G. (2017) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic. Acids. Res.*, **45**, D1040–D1045.
63. Weidemüller,P., Kholmatov,M., Petsalaki,E. and Zaugg,J.B. (2021) Transcription factors: Bridge between cell signaling and gene regulation. *Proteomics*, e2000034.
64. Yang,P., Wang,Y., Hoang,D., Tinkham,M., Patel,A., Sun,M.-A., Wolf,G., Baker,M., Chien,H.-C., Lai,K.-Y.N. *et al.* (2017) A placental growth factor is silenced in mouse embryos by the zinc finger protein ZFP568. *Science*, **356**, 757–759.
65. Johansson,P.A., Brattås,P.L., Douse,C.H., Hsieh,P., Pontis,J., Grassi,D., Garza,R., Jönsson,M.E., Atacho,D.A.M., Pircs,K. *et al.* (2020) A human-specific structural variation at the ZNF558 locus controls a gene regulatory network during forebrain development. bioRxiv doi: https://www.biorxiv.org/content/10.1101/2020.08.18.255562, 18 August 2020, preprint: not peer reviewed.
66. Lan,X., Ren,R., Feng,R., Ly,L.C., Lan,Y., Zhang,Z., Aboreden,N., Qin,K., Horton,J.R., Grevet,J.D. *et al.* (2021) ZNF410 uniquely activates the NuRD component CHD4 to silence fetal hemoglobin expression. *Mol. Cell*, **81**, 239–254.
67. Lu,B., Klingbeil,O., Tarumoto,Y., Somerville,T.D.D., Huang,Y.-H., Wei,Y., Wai,D.C., Low,J.K.K., Milazzo,J.P., Wu,X.S. *et al.* (2018) A transcription factor addiction in leukemia imposed by the MLL promoter sequence. *Cancer Cell*, **34**, 970–981.
68. Avsec,Ž., Weilert,M., Shrikumar,A., Krueger,S., Alexandari,A., Dalal,K., Fropf,R., McAnany,C., Gagneur,J., Kundaje,A. *et al.* (2021) Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.*, **53**, 354–366.
69. Minnoye,L., Taskiran,I.I., Mauduit,D., Fazio,M., Van Aerschot,L., Hulselmans,G., Christiaens,V., Makhzami,S., Seltenhammer,M., Karras,P. *et al.* (2020) Cross-species analysis of enhancer logic using deep learning. *Genome Res.*, **30**, 1815–1834.
70. Avsec,Ž., Kreuzhuber,R., Israeli,J., Xu,N., Cheng,J., Shrikumar,A., Banerjee,A., Kim,D.S., Beier,T., Urban,L. *et al.* (2019) The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.*, **37**, 592–600.